ADAPT OR FORGET: PROVABLE TRADEOFFS BETWEEN ADAM AND SGD IN NONSTATIONARY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We provide a theoretical analysis of Adam under non-stationary stochastic objectives, separating two regimes: Euclidean tracking under adaptive strong monotonicity of the Adam-preconditioned mean-gradient operator, and high-probability projected stationarity guarantees under general L-smooth objectives. In the tracking regime, we derive finite-time expected and high-probability bounds that decompose sharply into four components: initialization, objective drift, a first-moment tracking error governed by β_1, and a preconditioner perturbation governed by β_2. We characterize the burn-in time required for the transient terms to decay to the asymptotic tracking bound under constant and step-decay schedules. We also prove a high-probability bound on the average projected stationarity gap for Adam under distribution shift. Across both analyses, our bounds reveal a noise–drift tradeoff: in noise-dominated regimes, first-moment averaging and adaptive preconditioning can yield favorable upper guarantees, whereas in drift-dominated regimes, stale first-moment information and preconditioner perturbations can enlarge Adam's tracking guarantee, potentially allowing vanilla SGD to attain a smaller tracking error. Our explicit $(\beta_1, \beta_2, \epsilon)$-dependent bounds identify mechanisms through which adaptive step-sizing can help or hurt under nonstationarity and provide theoretical explanations consistent with Adam's empirical instability and stabilization under distribution shift.

1 INTRODUCTION

Consider the optimization problem on a compact convex set $\Theta \subset \mathbb{R}^d$ posed by a smooth objective function $G_t : \mathbb{R}^d \to \mathbb{R}$ defined as:

$$\boldsymbol{\theta}_t^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} G_t(\boldsymbol{\theta}), \quad G_t(\boldsymbol{\theta}) = \mathbb{E}_{X_t\sim\Pi_t}[g(\boldsymbol{\theta}, X_t)]. \tag{Opt}$$

Here $g(\boldsymbol{\theta}, X_t)$ is a noise perturbed measurement of $G_t(\boldsymbol{\theta})$ and X_t is a random variable sampled from a time-varying distribution Π_t. This setting connects naturally to classical stochastic tracking in signal processing (Kushner & Yin, 1997; Sayed, 2003), concept drift in online learning (Hazan, 2016; Cutler et al., 2021), and minimum divergence-type statistical objectives (Roy et al., 2026), where the goal is not convergence to a fixed solution but faithful tracking of an unknown, time-varying minimizer.

Adaptive gradient methods such as Adam (Kingma & Ba, 2015) have become the default optimizer for training large-scale deep learning models (Brown et al., 2020; Dosovitskiy et al., 2020), combining coordinate-wise adaptive step sizes derived from second-moment estimates with first-moment gradient averaging. In stationary settings, Adam is known to enjoy favorable convergence properties (Hong & Lin, 2024; Jin et al., 2025), and its adaptivity is widely credited with reducing sensitivity to hyperparameter tuning and accelerating progress on ill-conditioned objectives. Recent work analyzing stochastic gradient descent (SGD) and its momentum-based variants (SGDM), including Polyak's Heavy-Ball (Polyak, 1963) and Nesterov acceleration (Nesterov, 1983), in non-stationary settings Sahu et al. (2026) has shown that gradient averaging induces tracking lag under distribution shift. Since Adam incorporates momentum in both its first and second-moment estimates, it is natural to ask if similar degradation occurs, and whether the adaptive preconditioning compounds or

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.

alleviates it. These failure modes have indeed been observed empirically in continual learning and non-stationary optimization (Dohare et al., 2023; Lyle et al., 2023; Ellis et al., 2024; Dong & Wu, 2026), yet a precise theoretical account of when and why Adam degrades under distribution shift has remained unclear.

These observations motivate the following central question in the non-stationary tracking setting:

When does Adam's adaptive preconditioning help under distribution shift, and when can its moment estimates provably hurt relative to vanilla SGD, for L-smooth time-varying functions $G_t(\boldsymbol{\theta})$?

Our contributions. We answer this question under predictable distribution shift and make the following advances.

1. **Finite-time tracking bounds under adaptive strong monotonicity.** We prove finite-time tracking bounds for Adam that separate four distinct error sources: initialization decay, drift-induced tracking lag, a first-moment tracking error governed by β_1, and a preconditioner perturbation governed by β_2. To obtain a tracking recursion, we impose strong monotonicity on the mean-gradient map after Adam-style preconditioning. Since the realized Adam preconditioner is computed from the fresh sample X_{t+1} at the same step, we formulate this condition using a predictable proxy rather than the sample-dependent preconditioner itself. This yields a population-level contraction condition, while the difference between the realized preconditioner and its predictable proxy is controlled explicitly as an Adam-specific perturbation term.
2. **A noise–drift tradeoff separating Adam from SGD.** Our explicit $(\beta_1, \beta_2, \epsilon)$-dependent bounds reveal an optimizer-specific noise–drift tradeoff that is absent from the corresponding vanilla SGD guarantee and takes a different form from that of SGDM. In noise-dominated regimes, first-moment averaging and adaptive preconditioning can yield favorable high-probability tracking guarantees, and our bounds identify parameter regimes in which Adam's upper bound is possibly smaller than the corresponding SGD upper bound, consistent with our numerical simulations. In drift-dominated regimes, however, stale first-moment information together with preconditioner perturbations can enlarge Adam's asymptotic tracking bound, potentially allowing vanilla SGD to attain a smaller tracking error. The bounds also provide a theoretical mechanism consistent with the empirical observation of Lyle et al. (2023); Lee et al. (2024); Bai et al. (2025); Gogianu et al. (2026) that increasing ϵ stabilizes Adam under task changes: larger ϵ dampens the effect of adaptive second-moment variability, while also weakening the effective adaptive step size and thereby potentially slowing adaptation to a changing objective.
3. **Projected stationarity bounds under general Adam preconditioners.** Without the adaptive monotonicity condition, we prove a high-probability bound on the average projected stationarity gap for Adam under distribution shift. The same Adam-specific error structure persists: objective drift, first-moment bias-variance tradeoffs, and second-moment preconditioner perturbation all contribute explicitly, showing that the qualitative picture identified in the tracking regime extends to general Adam preconditioners.
4. **Experiments corroborating the noise–drift tradeoff.** We validate the theory on strongly convex least squares and three non-convex online problems: teacher–student MLP regression, phase retrieval, and matrix factorization under controlled drift and time-varying noise. Across these settings, the results support the predicted tradeoff: SGD performs better when drift dominates because it adapts more quickly, whereas Adam is more effective when stochastic noise dominates due to momentum and adaptive scaling.

1.1 Related Work

Adam (Kingma & Ba, 2015) has been the subject of extensive theoretical study in stationary settings. Indeed, in a variety of nonconvex settings, Adam provably converges under realistic conditions and adapts to the geometry of the problem in ways that SGD cannot (Li et al., 2023; Hong & Lin, 2024; Taniguchi et al., 2024; Jin et al., 2026). The advantage of Adam over SGD has been attributed to several sources, including heavy-tailed gradient distributions in language modeling (Kunstner et al., 2024), Hessian heterogeneity across layers in Transformer architectures (Zhang et al., 2024), and

the implicit sign-descent behavior of its updates (Kunstner et al., 2023). However, Adam does not universally outperform simpler methods: it has been shown to generalize worse than SGD in certain settings (Wilson et al., 2017), fail to converge without modification in convex problems (Reddi et al., 2018), and exhibit instability and loss of plasticity under distribution shift (Lyle et al., 2023; Dohare et al., 2023; Ellis et al., 2024). Most existing convergence theory for Adam assumes a fixed stationary objective.

Under non-stationarity, past gradients become stale and temporal averaging can be actively harmful. Building on the foundational dynamic regret framework of Zinkevich (2003), non-asymptotic tracking and regret guarantees have been established for SGD under time drift and dependent data (Cao et al., 2019; Cutler et al., 2023; Zhao et al., 2024; Shen et al., 2026). Most directly related to our work, Sahu et al. (2026) provide finite-time tracking bounds for SGD and its momentum variants under strongly convex smooth objectives, showing that momentum amplifies drift-induced tracking lag. Nonconvex tracking guarantees under distribution shift remain scarce even for SGD, and no prior work addresses Adam in this setting. The failure of Adam under distribution shift has been documented empirically in continual learning and reinforcement learning (Lyle et al., 2023; Dohare et al., 2023; Ellis et al., 2024), yet no prior work provides a theoretical framework characterizing when and why Adam degrades, how its $(\beta_1, \beta_2, \epsilon)$ hyperparameters interact with the drift-noise balance, or how it compares to SGD in a unified tracking setting. We survey the broader literature on adaptive methods and non-stationary optimization more extensively in Appendix A.

2 PRELIMINARIES

We first summarize the notation used throughout the paper. Scalars, vectors, and matrices are denoted by lowercase, bold lowercase, and bold uppercase letters, respectively; calligraphic letters denote sets, operators, or σ-algebras. For $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|$ denotes the ℓ_2 norm, $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ the inner product, and $\boldsymbol{x}^{\odot 2} := (x_1^2, \ldots, x_d^2)^\top$ the coordinatewise square. For a positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, we write $\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$. For $x \in \mathbb{R}$, we write $(x)_+ := \max\{x, 0\}$. We write $a_m = \mathcal{O}(b_m)$ if $a_m \le C b_m$ for some $C > 0$, $a_m = \Theta(b_m)$ if a_m and b_m are of the same order, and $a_m = o(b_m)$ if $a_m/b_m \to 0$ as $m \to \infty$. The notation $a_m \lesssim b_m$ (resp. $\gtrsim, \asymp$) indicates inequality up to constants independent of m and problem parameters. For $f : \mathbb{R}^d \to \mathbb{R}$, we write ∇f for the gradient and $f^\star = \min_{x\in\mathbb{R}^d} f(x)$. We use $\mathbb{E}[\cdot]$ for expectation and $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$ for the natural filtration. We write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We define $\log_+(x) := \log(e \vee x)$ for $x \ge 0$. We write $\mathcal{P}_\Theta$ to denote the projection operator onto the set Θ.

We next introduce conditional Ψ_α–Orlicz norms, which we will use throughout our analysis to control random quantities in a filtration-adapted (i.e., history-dependent) manner. Fix $\alpha \ge 1$. For a real-valued random variable X, recall the (unconditional) Ψ_α–Orlicz norm

$$\|X\|_{\Psi_\alpha} := \inf\Big\{u > 0 : \ \mathbb{E}\exp\big((|X|/u)^\alpha\big) \le 2\Big\}.$$

Given a σ-algebra $\mathcal{F}$, we write $\|X\|_{\Psi_\alpha|\mathcal{F}} \le K_{\mathcal{F}}$ for some $\mathcal{F}$-measurable $K_{\mathcal{F}} > 0$ if $\mathbb{E}\big[\exp\big(\big(|X|/K_{\mathcal{F}}\big)^\alpha\big)\big|\mathcal{F}\big] \le 2$. Equivalently, one may define for u

$$\|X\|_{\Psi_\alpha|\mathcal{F}} := \inf\Big\{u > 0 : \ \mathbb{E}\big[\exp\big((|X|/u)^\alpha\big) \mid \mathcal{F}\big] \le 2\Big\}$$

provided u is $\mathcal{F}$ measurable. Vector and matrix conditional Orlicz norms are defined analogously to their unconditional counterparts by taking suprema over one-dimensional projections (see Sahu et al. (2026) for details).

2.1 PROBLEM SETUP

Recall the optimization problem (Opt). Let $(\mathcal{F}_t)_{t\ge1}$ be the natural filtration $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$. Throughout our analysis, we adopt the following assumption from Sahu et al. (2026), which models stochasticity in the non-stationary setting:

Assumption 2.1 (Stochastic predictability framework)**.** *There exists a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge0}, \mathbb{P})$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $(X_t)_{t\ge0}$ be an $\mathbb{F}$–adapted process, i.e., X_t is $\mathcal{F}_t$–measurable for all t. For each $t \ge 0$, let Π_{t+1} denote the regular conditional law of X_{t+1} given $\mathcal{F}_t$, i.e., $\Pi_{t+1}(A) = \mathbb{P}(X_{t+1} \in A \mid \mathcal{F}_t)$ a.s. for every measurable set A, and note Π_{t+1} is $\mathcal{F}_t$–measurable. Define the conditional risk*

$$G_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t] = \mathbb{E}_{X\sim\Pi_{t+1}}[g(\boldsymbol{\theta}, X)]\,,$$

and let $\boldsymbol{\theta}^\star_{t+1} \in \arg\min_{\boldsymbol{\theta}\in\Theta} G_{t+1}(\boldsymbol{\theta})$ denote a (measurable) minimizer contained within the interior of Θ where $\Theta \subset \mathbb{R}^d$ is compact and convex. Assume the following hold for all $t \geq 0$:

1. ***(Predictable minimizer)** $\boldsymbol{\theta}^\star_{t+1}$ is $\mathcal{F}_t$–measurable.*[1]
2. ***(Algorithm adaptedness)** The iterate $\boldsymbol{\theta}_t$ is $\mathcal{F}_t$–measurable.*
3. ***(Martingale difference noise)** Define the conditional mean gradient $\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t]$, the conditional second moment $\boldsymbol{s}_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}))^{\odot 2} \mid \mathcal{F}_t]$, and the associated noise terms $\boldsymbol{\xi}_{t+1} := \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)$ and $\boldsymbol{\chi}_{t+1} := (\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2} - \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t)$. Then $\boldsymbol{\xi}_{t+1}$ and $\boldsymbol{\chi}_{t+1}$ are $\mathcal{F}_{t+1}$-measurable and satisfy $\mathbb{E}[\boldsymbol{\xi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\chi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ a.s. for all t.*

The time-varying distribution Π_t in Assumption 2.1 is a standard modeling assumption in the non-stationary stochastic optimization literature (Sahu et al., 2026; Cutler et al., 2021) and captures several practical settings, including policy optimization and reinforcement learning (Kakade & Langford, 2002; Schulman et al., 2015), online recommendation and contextual bandits (Li et al., 2010), continual learning (Parisi et al., 2019), and federated learning with non-stationary clients (Kairouz & McMahan, 2021).

The Adam update (Kingma & Ba, 2015), with step size $\alpha > 0$, momentum parameters $\beta_1, \beta_2 \in (0, 1)$, and stabilization parameter $\epsilon > 0$, is summarized in Algorithm 1. We write the stochastic gradient and its elementwise square as

$$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) = \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}, \quad \big(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\big)^{\odot 2} = \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t) + \boldsymbol{\chi}_{t+1}.$$

For any positive definite matrix $\boldsymbol{A} \succ 0$, define the $\boldsymbol{A}$-metric projection onto Θ by $\mathcal{P}^{\boldsymbol{A}}_\Theta(\boldsymbol{z}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{z}\|^2_{\boldsymbol{A}}$. Thus, $\mathcal{P}^{\boldsymbol{I}_d}_\Theta$ is the usual Euclidean projection. Throughout a given run of Algorithm 1, the projection convention is fixed. For the tracking results, we take $\boldsymbol{A}_{t+1} = \boldsymbol{I}_d$, whereas for the metric-projected stationarity results, we take $\boldsymbol{A}_{t+1} = \widetilde{\boldsymbol{P}}^{-1}_{t+1}$, where $\widetilde{\boldsymbol{P}}_{t+1}$ is the predictable preconditioner (see Appendix D.3).

Algorithm 1 Adam optimizer

Require: $\boldsymbol{\theta}_0 \in \Theta$, horizon $T \geq 1$, step size $\alpha > 0$, $\beta_1, \beta_2 \in (0, 1)$, and $\epsilon > 0$

Initialize $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \mathbf{0}$
for $t = 0, 1, \ldots, T-1$ **do**
  $\boldsymbol{g}_{t+1} := \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})$
  $\boldsymbol{m}_{t+1} := \beta_1 \boldsymbol{m}_t + (1-\beta_1)\boldsymbol{g}_{t+1}$
  $\boldsymbol{v}_{t+1} := \beta_2 \boldsymbol{v}_t + (1-\beta_2)\boldsymbol{g}^{\odot 2}_{t+1}$
  $\widehat{\boldsymbol{m}}_{t+1} := \boldsymbol{m}_{t+1}/(1-\beta_1^{t+1}), \quad \widehat{\boldsymbol{v}}_{t+1} := \boldsymbol{v}_{t+1}/(1-\beta_2^{t+1})$
  $\boldsymbol{P}_{t+1} := \mathrm{Diag}\left(\left(\sqrt{\widehat{\boldsymbol{v}}_{t+1}} + \epsilon\right)^{-1}\right)$
  $\boldsymbol{\theta}_{t+1} := \mathcal{P}^{\boldsymbol{A}_{t+1}}_\Theta(\boldsymbol{\theta}_t - \alpha \boldsymbol{P}_{t+1}\widehat{\boldsymbol{m}}_{t+1})$ (Adam)
end for

To analyze the convergence of (Adam), we assume the conditional mean gradient map $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta})$ is uniformly L-Lipschitz. Under mild regularity conditions (e.g. sufficient integrability for the Dominated Convergence Theorem), this implies $G_{t+1}(\boldsymbol{\theta})$ is L-smooth.

Assumption 2.2 (Uniform Lipschitz continuity)**.** *There exists a constant $0 < L < \infty$ such that for all $t \geq 0$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $\|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.*

To obtain high-probability guarantees, we will need to make a standard light-tail assumption on the gradient noise (Sahu et al., 2026; Harvey et al., 2019; Cutler et al., 2021):

[1] Since $G_{t+1}(\boldsymbol{\theta}) = \mathbb{E}[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t]$ is $\mathcal{F}_t$-measurable for each fixed $\boldsymbol{\theta}$, standard measurable selection conditions ensure existence of an $\mathcal{F}_t$-measurable minimizer. If the minimizer is a.s. unique (e.g., under strong convexity), measurability follows by the measurable maximum theorem. See Sahu et al. (2026) for details.

Assumption 2.3 (Conditional sub-Gaussian gradient noise along iterates)**.** *There exists a constant* $\sigma > 0$ *such that for all* $t \geq 0$, $\left\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\right\|_{\Psi_2|\mathcal{F}_t} \leq \sigma$ *a.s.*

Remark 2.1. Sub-exponential noise can be handled via Bernstein-type martingale inequalities, while bounded conditional q-th moments for $q > 2$ yield polynomial-tail guarantees. Predictable time-varying noise levels σ_t are also allowed by replacing uniform variance terms with the corresponding sums of σ_t^2. We use the sub-Gaussian formulation for simplicity and consistency with the stochastic tracking literature.

3 Theoretical Results

3.1 Tracking under adaptive strong monotonicity

In this section, we obtain high-probability guarantees on the tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$, deferring expectation bounds to Appendix D.3. Unlike Sahu et al. (2026), which requires only that $\bar{\boldsymbol{g}}_{t+1}$ itself be strongly monotone, Adam's preconditioned update requires a compatibility condition between the preconditioner and the curvature of the conditional risk. Since $\boldsymbol{P}_{t+1}$ depends on the current sample X_{t+1}, it is $\mathcal{F}_{t+1}$-measurable but not $\mathcal{F}_t$-measurable. Moreover, $\bar{\boldsymbol{g}}_{t+1}$ is the conditional mean-gradient map and is $\mathcal{F}_t$-measurable, so imposing monotonicity on $\boldsymbol{P}_{t+1}\bar{\boldsymbol{g}}_{t+1}$ would make the structural contraction condition depend on a single stochastic realization. We therefore work with the predictable proxy $\widetilde{\boldsymbol{P}}_{t+1}$, obtained by replacing $(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2}$ in the second-moment update by its conditional mean $\boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t)$. This separates the predictable preconditioned mean-gradient geometry from the random preconditioner perturbation; the precise construction is given in Appendix D.1.

Assumption 3.1 (Adaptive strong monotonicity)**.** . *There exists* $0 < \mu < \infty$ *such that, almost surely, for all* $t \geq 0$ *and all* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}',\, \widetilde{\boldsymbol{P}}_{t+1}\left(\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}')\right)\right\rangle \geq \mu\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$.

Assumption 3.1 requires the predictable preconditioner to be compatible with the curvature of the conditional risk. It holds in simple settings such as scalar preconditioning and diagonal quadratics, but also for substantially richer models. We refer the reader to Appendix C for examples involving dense generalized linear models, graph-coupled objectives, and locally contracting nonconvex problems. Since $\widetilde{\boldsymbol{P}}_{t+1}$ is positive definite, preconditioning preserves the zeros of the mean-gradient map, and strong monotonicity therefore guarantees uniqueness of the moving optimum. Importantly, the condition is imposed on the preconditioned operator and does not require G_{t+1} itself to be strongly convex.

We can now state the following theorem that establishes the high-probability tracking error for (Adam) in nonstationary stochastic environments. We defer the proofs for this section to Appendix D.

Theorem 3.1 (High probability tracking error bound for (Adam))**.** *Suppose Assumption 2.2, Assumption 2.3, and 3.1 hold and* $\operatorname{diam}(\Theta) \leq R$*. Then, for all* $t \in [T]$, $\delta \in (0,1)$, *and* $\alpha \leq \min\{\mu\epsilon^2/(4L^2),\, \mu^{-1}\}$, *the following tracking error bound holds for (Adam) with probability at least* $1 - \delta$*:*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \rho_\alpha^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{\mathfrak{D}_t}{\alpha\mu} + \frac{\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{T}{\delta}\right)\right]^2\left[\beta_2 t\lambda_2^{t-1} + \frac{1-\beta_2}{\alpha\mu}\right] + \frac{1}{\mu^2\epsilon^2}\left\{\beta_1^2L^2R^2 + \frac{\sigma^2}{1+\beta_1}\left(d + \log\frac{T}{\delta}\right)\left[1 - \beta_1 + \alpha\mu\beta_1 t\lambda_1^{t-1}\right]\right\}$$

where $\rho_\alpha := 1 - \alpha\mu/2$, $\mathfrak{D}_t := \sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\Delta_\ell^2$, $\Delta_\ell := \|\boldsymbol{\theta}_\ell^\star - \boldsymbol{\theta}_{\ell+1}^\star\|$, *and* $\lambda_i := \rho_\alpha \vee \beta_i$ *for* $i \in \{1,2\}$.

Comparison with SGD, SGDM, and implications for nonstationary optimization. To make the comparison with Adam self-contained, we recall the corresponding high-probability tracking guarantee for constant-stepsize SGD from Sahu et al. (2026):

Theorem 3.2 (High probability tracking error bound for SGD)**.** *Assume* G_t *is* μ*-strongly convex and* L*-smooth. Under Assumption 2.3, for all* $t \in [T]$, $\alpha \leq \min\left\{\mu/L^2, 1/L\right\}$, *and* $\delta \in (0,1)$, *the*

following tracking error bound holds for SGD with probability at least $1-\delta$*:*

$$\|\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^\star\|^2 \lesssim \left(1-\frac{\alpha\mu}{2}\right)^t \|\boldsymbol{\theta}_0-\boldsymbol{\theta}_0^\star\|^2+\frac{\mathfrak{D}_t}{\alpha\mu}+\frac{d\sigma^2\alpha}{\mu}+d\sigma^2\alpha^2\log\frac{T}{\delta}+\left(\frac{\sigma^2\alpha}{\mu}+\alpha^2\sigma^2\mathfrak{D}_t^{(2)}\right)\log\frac{T}{\delta},$$

where $\mathfrak{D}_t := \sum_{\ell=0}^{t-1}(1-\alpha\mu/2)^{t-\ell-1}\Delta_\ell^2$ *and* $\mathfrak{D}_t^{(2)} := \sum_{\ell=0}^{t-1}(1-\alpha\mu/2)^{2(t-\ell-1)}\Delta_\ell^2$.

Our Adam bound has the same basic decomposition as the SGD and SGDM guarantees of Sahu et al. (2026): an exponentially decaying initialization error, a discounted drift term, and stochastic terms determining the eventual tracking floor.

Adam separates the effects of first and second-moment memory. The parameter β_1 controls a smoothing–responsiveness tradeoff: its worst-case deterministic memory-bias contribution scales as β_1^2, while its stochastic contribution is reduced by the averaging factor $(1-\beta_1)/(1+\beta_1)$. Thus larger β_1 suppresses gradient noise without causing the bias to diverge, but increases the effective memory horizon, so stale gradients persist longer under nonstationarity. The remaining $\beta_1 t\lambda_1^{t-1}$ term is a finite-time transient, with $\lambda_1 = \rho_\alpha \vee \beta_1$, and vanishes geometrically. Consequently, larger β_1 is most favorable in relatively stable, noise-dominated regimes, whereas smaller β_1 permits faster adaptation when the objective changes rapidly.

The parameter β_2 induces an analogous transient–floor tradeoff through the adaptive preconditioner. Its persistent contribution scales with $1-\beta_2$, while the bias-correction transient scales as $\beta_2 t\lambda_2^{t-1}$, where $\lambda_2 = \rho_\alpha \vee \beta_2$. Hence increasing β_2 reduces the long-run perturbation induced by the stochastic second-moment estimate, but slows the decay of the corresponding transient. Large β_2 is therefore advantageous over long and comparatively stable horizons, while smaller β_2 can be preferable when rapid adaptation is more important.

Our bounds also reveal a tradeoff governed by ϵ. Increasing ϵ permits a larger admissible stepsize through $\alpha \le \mu\epsilon^2/(4L^2)$ and reduces the first and second-moment adaptive penalties, which scale as $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$, respectively. Thus larger ϵ can improve tracking when stochastic adaptive-preconditioning effects dominate, while its overall benefit still depends on how the preconditioned geometry, and hence μ, changes with ϵ. This provides a theoretical mechanism for the empirical observation of Lyle et al. (2023); Lee et al. (2024); Bai et al. (2025); Gogianu et al. (2026) that increasing ϵ can stabilize Adam under nonstationarity.

After the initialization and bias-correction transients decay, the Adam tracking floor consists of four principal contributions: drift of order $\Delta^2/(\mu^2\alpha^2)$, first-moment memory bias of order $\beta_1^2L^2R^2/(\mu^2\epsilon^2)$, first-moment variance of order $\sigma^2(1-\beta_1)(d+\log(N/\delta))/[\mu^2\epsilon^2(1+\beta_1)]$, and second-moment preconditioner perturbation of order $(1-\beta_2)[L^2R^2+\sigma^2(d+\log(N/\delta))]^2/(\mu^2\epsilon^4)$. Thus, increasing β_1 suppresses stochastic gradient noise but increases the persistent bias from stale first-moment information, while increasing β_2 reduces the persistent error induced by fluctuations in the adaptive preconditioner. Both forms of memory can also slow adaptation: the corresponding transients decay at rates governed by $\lambda_{1,\alpha} = \rho_\alpha \vee \beta_1$ and $\lambda_{2,\alpha} = \rho_\alpha \vee \beta_2$, as reflected in the burn-in terms $\mathcal{B}_1$ and $\mathcal{B}_2$. Hence larger β_1 and β_2 improve temporal averaging and preconditioner stability, respectively, at the cost of slower response to nonstationarity.

Theorem 3.3 (Time to reach the tracking floor with high probability for (Adam))**.** *Under the standing assumptions of Theorem 3.1, suppose* $0 < \alpha \le \alpha_{\max} := \min\{\mu\epsilon^2/(4L^2), \mu^{-1}\}$, $\Delta_t \le \Delta$ *almost surely for all* $t \ge 0$*, and fix a deterministic horizon bound* $N \ge 1$ *and* $\delta \in (0,1)$*. Define the tracking-floor*

$$\mathcal{E}_{\mathrm{A}}(N,\delta,\alpha) \asymp \frac{\Delta^2}{\mu^2\alpha^2}+\frac{\beta_1^2L^2R^2}{\mu^2\epsilon^2}+\frac{C^2\sigma^2(1-\beta_1)}{\mu^2\epsilon^2(1+\beta_1)}\left(d+\log\frac{N}{\delta}\right)+\frac{C(1-\beta_2)}{\mu^2\epsilon^4}\left[L^2R^2+\sigma^2\left(d+\log\frac{N}{\delta}\right)\right]^2,$$

where $C > 0$ *is a fixed sufficiently large universal constant. For* $\alpha \in (0, \alpha_{\max}]$*, let* $\rho_\alpha := 1-\alpha\mu/2$ *and* $\lambda_{i,\alpha} := \rho_\alpha \vee \beta_i$ *for* $i \in \{1,2\}$*. For a target floor* $E > 0$*, define*

$$\mathcal{B}_1(\alpha,E) \asymp \frac{1}{1-\lambda_{1,\alpha}}\log_+\left(\frac{C^2\alpha\sigma^2\left(d+\log\frac{N}{\delta}\right)}{\mu\epsilon^2(1+\beta_1)(1-\lambda_{1,\alpha})E}\right)$$

$$\mathcal{B}_2(\alpha,E) \asymp \frac{1}{1-\lambda_{2,\alpha}}\log_+\left(\frac{C\alpha\left[L^2R^2+\sigma^2\left(d+\log\frac{N}{\delta}\right)\right]^2}{\mu\epsilon^4(1-\lambda_{2,\alpha})E}\right).$$

Then we have the following:

1. ***(Constant learning rate).*** *If $\alpha_t \equiv \alpha$, with α chosen deterministically, then for any integer $1 \le T \le N$, with probability $\ge 1-\delta$, for all $t \in [T]$,*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \rho_\alpha^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \mathcal{E}_{\mathrm{A}}(N,\delta,\alpha) + \frac{\alpha\sigma^2}{\mu\epsilon^2(1+\beta_1)}\left(d + \log\frac{N}{\delta}\right)\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1} + \frac{\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d+\log\frac{N}{\delta}\right)\right]^2\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_2^{\ell+1}.$$

Let $\alpha_{\mathrm{A}}^\star \in (0, \alpha_{\max}]$ be a deterministic target stepsize and $\mathcal{E}_{\mathrm{A}}^\star := \mathcal{E}_{\mathrm{A}}(N,\delta,\alpha_{\mathrm{A}}^\star)$. Then, running Adam with $\alpha = \alpha_{\mathrm{A}}^\star$ gives $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \mathcal{E}_{\mathrm{A}}^\star$ after at most

$$t_{\mathrm{A}}^\star \lesssim \frac{1}{\mu\alpha_{\mathrm{A}}^\star}\log_+\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \mathcal{B}_1\left(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star\right) + \mathcal{B}_2\left(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star\right)$$

iterations, provided $t_{\mathrm{A}}^\star \le T$. The bound holds simultaneously for $t_{\mathrm{A}}^\star \le t \le T$ with probability $\ge 1-\delta$.

2. ***(Step-decay with Adam-state restart).*** *Suppose a deterministic upper bound $D \ge \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2$ is available and $\alpha_{\mathrm{A}}^\star < \alpha_{\max}$. Set $\alpha_0 := \alpha_{\max}$, $\alpha_k := (\alpha_{k-1} + \alpha_{\mathrm{A}}^\star)/2$, and $K := 1 + \lceil \log_2(\alpha_0/\alpha_{\mathrm{A}}^\star)\rceil$. For each $k \ge 0$, let $E_k := \mathcal{E}_{\mathrm{A}}(N,\delta,\alpha_k)$, and set*

$$T_0 \asymp \left\lceil \max\left\{\frac{1}{\mu\alpha_0}\log_+\left(\frac{D}{E_0}\right), \mathcal{B}_1(\alpha_0,E_0), \mathcal{B}_2(\alpha_0,E_0)\right\}\right\rceil,$$
$$T_k \asymp \left\lceil \max\left\{\frac{1}{\mu\alpha_k}, \mathcal{B}_1(\alpha_k,E_k), \mathcal{B}_2(\alpha_k,E_k)\right\}\right\rceil, \qquad k \ge 1.$$

Suppose $T := \sum_{k=0}^{K-1} T_k \le N$. Running Adam at constant stepsize α_k for T_k steps per epoch with $(\boldsymbol{m}, \boldsymbol{v})$ and the bias-correction restarted at the beginning of every epoch yields $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim \mathcal{E}_{\mathrm{A}}^\star$ with probability $\ge 1 - K\delta$. Moreover,

$$T \lesssim \frac{1}{\mu\alpha_0}\log_+\left(\frac{D}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \frac{1}{\mu\alpha_{\mathrm{A}}^\star} + \sum_{k=0}^{K-1}\mathcal{B}_1(\alpha_k,E_k) + \sum_{k=0}^{K-1}\mathcal{B}_2(\alpha_k,E_k).$$

4 Metric-projected stationarity under general Adam preconditioning

The preceding analysis establishes Euclidean tracking guarantees under adaptive strong monotonicity. We now drop this condition and instead study stationarity through the $\widetilde{\boldsymbol{P}}_{t+1}^{-1}$-metric projected-gradient mapping $\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)$, the natural constrained analogue of the preconditioned gradient (see Appendix F.1 for details). We derive high-probability bounds on its average squared norm under nonstationary objectives and defer the corresponding expectation bounds to Appendix E.2. Throughout this section, we additionally assume that $G_t(\boldsymbol{\theta})$ is uniformly bounded below by $G^\star$, as standard in nonconvex optimization (Reddi et al., 2018; Défossez et al., 2022; Li et al., 2023).

Assumption 4.1 (Uniform lower boundedness)**.** *There exists a constant $G^\star > -\infty$ such that for all $t \in [T]$ and all $\boldsymbol{\theta} \in \Theta$, $G_t(\boldsymbol{\theta}) \ge G^\star$.*

We now state our main result for Adam with general adaptive preconditioners, which gives a high-probability bound on the average projected stationarity gap of the projected Adam update under non-stationary objectives. All proofs for this section are deferred to Appendix E.

Theorem 4.1 (High-probability projected-gradient rate)**.** *Suppose $\mathrm{diam}(\Theta) \le R$. Then, for all integers $T \ge 1$ and all $\delta \in (0,1)$, if $\alpha \le \epsilon/(4L)$, the iterates generated by the projected Adam update satisfy, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}_{t+1}^{-1}} \lesssim \frac{G_1(\boldsymbol{\theta}_0) - G^\star + \mathfrak{D}_T}{\alpha T} + \frac{\mathsf{Decay}_T(\delta)}{T} + \mathsf{Floor}_T(\delta), \tag{2}$$

where

$$\mathsf{Decay}_T(\delta) \sim \mathcal{O}\left\{\frac{\beta_2}{\epsilon^3(1-\beta_2)}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2+\frac{\sigma^2\left(d+\log\frac{T}{\delta}\right)(1+\log T)}{\epsilon(1+\beta_1)}\right\},$$

$$\mathsf{Floor}_T(\delta) \sim \mathcal{O}\left\{\frac{L^2R^2\beta_1^2}{\epsilon}+\frac{\sigma^2(1-\beta_1)}{\epsilon(1+\beta_1)}\left(d+\log\frac{T}{\delta}\right)+\frac{1-\beta_2}{\epsilon^3}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2\right\}.$$

The explicit rate in Theorem 4.1 separates objective variation, finite-horizon transients, and persistent Adam errors. The term $(G_1(\boldsymbol{\theta}_0) - G^\star + \mathfrak{D}_T)/(\alpha T)$ vanishes whenever $\mathfrak{D}_T = o(T)$, and at the usual $1/T$ rate when $\mathfrak{D}_T = O(1)$. Under persistent nonstationarity with $\mathfrak{D}_T = \Theta(T)$, this contribution need not vanish. The term $\mathsf{Decay}_T(\delta)/T$ collects the remaining bias-correction and finite-horizon effects and vanishes up to logarithmic factors.

The persistent term $\mathsf{Floor}_T(\delta)$ makes the roles of β_1 and β_2 explicit. The first-moment parameter β_1 trades off a memory bias of order β_1^2 against a variance contribution proportional to $(1-\beta_1)/(1+\beta_1)$. Thus larger β_1 improves noise averaging but increases the cost of stale gradients. Similarly, β_2 induces a transient–floor tradeoff: the finite-horizon transient scales as $\beta_2/(1-\beta_2)$, while the persistent preconditioner perturbation scales as $1-\beta_2$. Hence larger β_2 improves the long-run preconditioner floor at the cost of slower finite-horizon adaptation. Together, these terms quantify the bias–variance and stability–adaptivity tradeoffs governing Adam under nonstationarity.

5 Numerical Experiments

We compare SGD and Adam on six online problems: strongly convex least squares, logistic regression, lasso regression, and three non-convex models, namely teacher–student MLP regression, phase retrieval, and matrix factorization. In all experiments, the target evolves via a normalized random walk, $\boldsymbol{\theta}^\star_{t+1} = \boldsymbol{\theta}^\star_t + \Delta_t \frac{\boldsymbol{u}_t}{\|\boldsymbol{u}_t\|}$, $\boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, where Δ_t controls drift and σ_t controls stochastic observation noise. For each problem, we show high-drift/low-noise (left) and low-drift/high-noise (right) regimes. SGD performs better in drift-dominated settings due to faster adaptation, while Adam is more effective in noise-dominated regimes via momentum and adaptive scaling. Additional experimental details are provided in Appendix G.

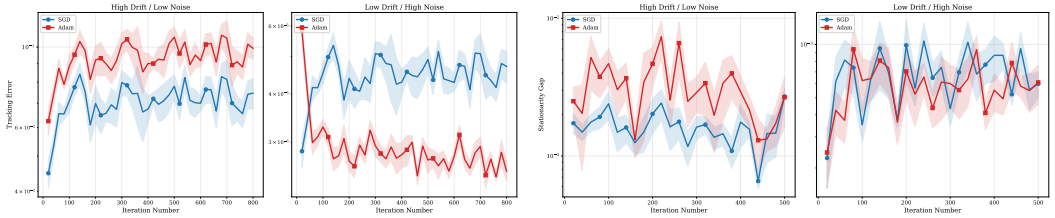

Figure 1: **Left**: Strongly convex least squares with objective $F_t(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star_t)\|^2$, where we report the tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2$. **Right**: Teacher–student MLP regression with objective $F_t(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - f_{\boldsymbol{\theta}^\star_t}(\boldsymbol{x})\right)^2\right]$, where we report the stationarity gap. Within each experiment, the left subplot corresponds to the high-drift, low-noise regime with $\Delta_t = c_2\log(t+2)$ and $\sigma_t = c_1\log(t+2)$, while the right subplot corresponds to the low-drift, high-noise regime with $\Delta_t = c_1\log(t+2)$ and $\sigma_t = c_2\log(t+2)$, for positive constants $c_1 < c_2$.

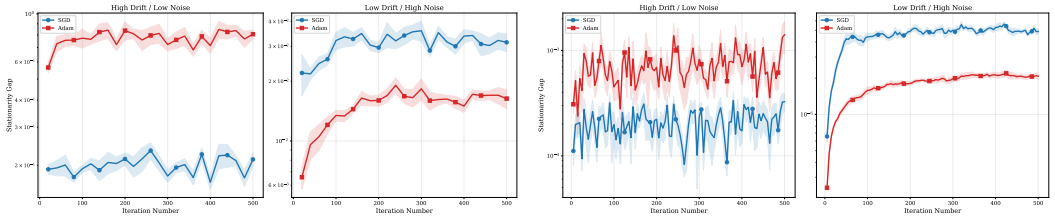

Figure 2: **Left**: Phase retrieval with objective $F_t(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\left((\boldsymbol{x}^\top\boldsymbol{\theta})^2 - (\boldsymbol{x}^\top\boldsymbol{\theta}_t^\star)^2\right)^2\right]$, where we report the stationarity gap. **Right**: Matrix factorization with objective $F_t(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2mn}\left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}_t^\star\right\|_F^2$, where we also report the stationarity gap. Within each experiment, the left subplot corresponds to the high-drift, low-noise regime and the right subplot corresponds to the low-drift, high-noise regime. For phase retrieval, these regimes are given by $\Delta_t = c_2\log(t+2)$, $\sigma_t = c_1\log(t+2)$ and $\Delta_t = c_1\log(t+2)$, $\sigma_t = c_2\log(t+2)$, respectively, for positive constants $c_1 < c_2$. For matrix factorization, these regimes are given by $\Delta_t = c_2$, $\sigma_t = c_1$ and $\Delta_t = c_3$, $\sigma_t = c_4$, respectively, where $0 < c_1 < c_2$ and $0 < c_3 < c_4$.

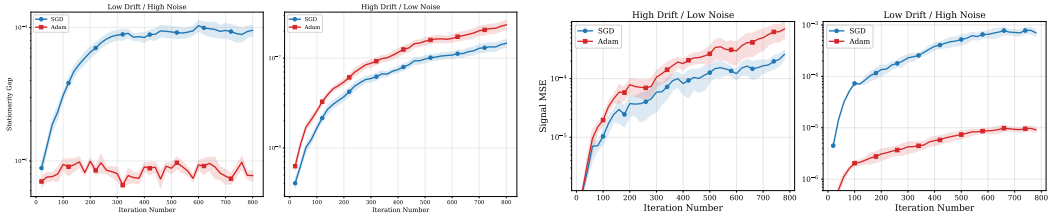

Figure 3: **Left**: Logistic regression, where we report the stationarity gap. **Right**: Lasso regression, where we report prediction MSE. Within each experiment, the left subplot corresponds to the higher-drift, lower-noise regime and the right subplot corresponds to the lower-drift, higher-noise regime. For logistic regression, these regimes are given by $\Delta_t = c_2$, $\sigma_t = c_1$ and $\Delta_t = c_3$, $\sigma_t = c_4$, respectively, where $0 < c_1 < c_2$ and $0 < c_3 < c_4$. For lasso regression, these regimes are given by $\Delta_t = c_2\log(t+2)$, $\sigma_t = c_3\log(t+2)$ and $\Delta_t = c_1\log(t+2)$, $\sigma_t = c_4\log(t+2)$, respectively, where $0 < c_1 < c_2$ and $0 < c_3 < c_4$. In both settings, the same qualitative pattern appears: SGD performs better in the more drift-dominated, lower-noise regime, whereas Adam improves relative performance in the more noise-dominated regime.

6 Conclusion

We provide a finite-time theoretical analysis of Adam under non-stationary stochastic objectives, characterizing how its moment estimates interact with distribution shift. Our bounds decompose the tracking error into four interpretable components (initialization, objective drift, first-moment memory bias, and preconditioner perturbation) and reveal a fundamental noise–drift tradeoff: Adam's adaptive mechanisms can yield more favorable tracking guarantees when gradient noise dominates, whereas under substantial drift, stale first-moment information and preconditioner perturbations can enlarge Adam's tracking bound, potentially allowing vanilla SGD to attain a smaller tracking error. The explicit $(\beta_1, \beta_2, \epsilon)$-dependence of our bounds provides a theoretical explanation for the empirical instability of Adam under distribution shift, including the stabilizing effect of increasing ϵ under task changes, and our experimental results corroborate the predicted tradeoffs across synthetic nonstationary settings.

For future work, it would be interesting to establish minimax lower bounds for Adam under non-stationarity, both to characterize the best achievable tracking rates and to determine whether the dependence of our upper bounds on drift, noise, and $(\beta_1, \beta_2, \epsilon)$ is minimax optimal. The lower bounds of Sahu et al. (2026) for SGD and SGDM rely on strongly convex constructions, and extending such results to the general L-smooth setting considered here would likely require different techniques. Another promising direction is to develop adaptive optimizers that can switch between update rules or parameter regimes so as to minimize cumulative tracking error.

AI USE STATEMENT

In this work, we used generative AI tools to assist with verifying proofs and mathematical statements and running and debugging experiments, including generating synthetic data i.e. covariate matrices and stochastic observational data. We did not use generative AI tools to develop the core theoretical framework, formulate the central hypotheses, or design the primary research methodology. Other required disclosure tasks were not applicable to this work. Additionally, we used generative AI tools to assist with editing the manuscript for clarity, grammar, and readability, as well as with writing and debugging research code.

We reviewed all AI-assisted work. In particular, all AI-assisted experimental code and results were reviewed and tested for correctness. All final scientific claims, theoretical results, experimental conclusions, and manuscript text were determined and approved by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of Optimization. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of both our theoretical and empirical results. All assumptions, problem settings, and algorithmic parameters used in our theoretical analysis are stated explicitly in the main text, and complete proofs of the theoretical claims are provided in the appendix. For the experiments, we provide detailed descriptions of the optimization problems, data-generation procedures, model architectures, hyperparameters, drift and noise settings, evaluation metrics, and training configurations in Appendix G. We also report results across multiple random seeds and use consistent training budgets when comparing SGD and Adam. Together, these details are intended to provide sufficient information to reproduce the theoretical derivations and experimental findings reported in this work.

REFERENCES

Zhiwei Bai, Zhangchen Zhou, Jiajie Zhao, Xiaolong Li, Zhiyu Li, Feiyu Xiong, Hongkang Yang, Yaoyu Zhang, and Zhi-Qin John Xu. Adaptive preconditioners trigger loss spikes in adam. *arXiv preprint arXiv:2506.04805*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Xuanyu Cao, Junshan Zhang, and H. Vincent Poor. On the time-varying distributions of online stochastic optimization. In *2019 American Control Conference (ACC)*, pp. 1494–1500, 2019. doi: 10.23919/ACC.2019.8814889.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11859–11869. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/62e7f2e090fe150ef8deb4466fdc81b3-Paper.pdf.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023. URL http://jmlr.org/papers/v24/21-1410.html.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Maintaining plasticity in deep continual learning. *arXiv preprint arXiv:2306.13812*, 2023.

Yuze Dong and Jinsong Wu. Rethinking Adam for time series forecasting: A simple heuristic to improve optimization under distribution shifts. *Neurocomputing*, 682:133279, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133279. URL https://www.sciencedirect.com/science/article/pii/S0925231226006764.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benjamin Ellis, Matthew T Jackson, Andrei Lupu, Alexander D Goldie, Mattie Fellows, Shimon Whiteson, and Jakob N Foerster. Adam on local time: Addressing nonstationarity in rl with relative Adam timesteps. *Advances in Neural Information Processing Systems*, 37:134567–134590, 2024.

Florin Gogianu, Luțu Adrian-Cătălin, and Razvan Pascanu. Revisiting adam for streaming reinforcement learning. *Reinforcement Learning Journal*, 7, 2026.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1579–1613. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/harvey19a.html.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 08 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 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. In *International Conference on Machine Learning*, pp. 27979–28030. PMLR, 2025.

Ruinan Jin, Yingbin Liang, and Shaofeng Zou. Why Adam can beat sgd: Second-moment normalization yields sharper tails. *arXiv preprint arXiv:2603.03099*, 2026.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Llearning*, pp. 267–274, 2002.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. URL http://jmlr.org/papers/v22/19-466.html.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *11th International Conference on Learning Representations (ICLR)*, 2023.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Harold J. Kushner and G. George Yin. *Applications: Proofs of Convergence*, pp. 251–272. Springer New York, New York, NY, 1997. ISBN 978-1-4899-2696-8. doi: 10.1007/978-1-4899-2696-8_9. URL https://doi.org/10.1007/978-1-4899-2696-8_9.

Hojoon Lee, Hyeonseo Cho, Hyunseung Kim, Donghu Kim, Dugki Min, Jaegul Choo, and Clare Lyle. Slow and steady wins the race: Maintaining plasticity with hare and tortoise networks. *arXiv preprint arXiv:2406.02596*, 2024.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, pp. 23190–23211. PMLR, 2023.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl Akad Nauk SSSR*, 269:543, 1983. URL https://cir.nii.ac.jp/crid/1370862715914709505.

Y. Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer US, 2018. ISBN 9783319915784. URL https://books.google.com/books?id=IPh6DwAAQBAJ.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(63)90382-3. URL https://www.sciencedirect.com/science/article/pii/0041555363903823.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Subhrajyoty Roy, Abir Sarkar, Abhik Ghosh, and Ayanendranath Basu. Asymptotic breakdown point analysis for a general class of minimum divergence estimators. *Bernoulli*, 32(1):698–722, 2026.

Sharan Sahu, Cameron J Hogan, and Martin T Wells. On the provable suboptimality of momentum sgd in nonstationary stochastic optimization. *arXiv preprint arXiv:2601.12238*, 2026.

A.H. Sayed. *Fundamentals of Adaptive Filtering*. IEEE Press. Wiley, 2003. ISBN 9780471461265. URL https://books.google.com/books?id=VaAV4uqMuKYC.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897. PMLR, 2015.

Yinan Shen, Yichen Zhang, and Wen-Xin Zhou. SGD with dependent data: Optimal estimation, regret, and inference, 2026. URL https://arxiv.org/abs/2601.01371. arXiv:2601.01371.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. *Advances in Neural Information Processing Systems*, 37:72438–72474, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. URL http://jmlr.org/papers/v17/16-157.html.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024. URL http://jmlr.org/papers/v25/21-0748.html.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 928–935. AAAI Press, 2003. ISBN 1577351894.

APPENDICES

A Related work 15

A.1 Adam in stationary settings 15

A.2 SGD and momentum in non-stationary settings 15

A.3 Adam and adaptive methods under non-stationarity 15

B Problem setup, standing assumptions, and notation 16

C Examples and interpretation of adaptive strong monotonicity 17

D Proofs for tracking under adaptive strong monotonicity 20

D.1 Proofs for Theorem 3.1 20

D.2 Proof of Theorem 3.3 27

D.3 Proofs for expected tracking error 30

E Proofs for metric-projected stationarity under general Adam preconditioning 33

E.1 Proof of Theorem 4.1 36

E.2 Proof of expected projected stationarity bounds 37

F Technical lemmas 37

F.1 Metric projection and projected-gradient mappings 37

F.2 Martingale concentration inequalities 39

F.3 L-smoothness and descent inequalities 40

G Additional experimental details and results 40

G.1 General setup 40

G.2 Task-specific details 41

G.3 Additional experimental examples 42

G.4 Adam hyperparameter dependence 43

A Related work

Adam in stationary settings. Since its introduction, Adam (Kingma & Ba, 2015) has been the subject of extensive theoretical study in stationary settings. Early work showed that Adam can fail to converge even in simple convex problems due to the exponential moving average of squared gradients, motivating the AMSGrad variant (Reddi et al., 2018). Subsequent analyses established convergence guarantees under progressively weaker assumptions: Défossez et al. (2022) provide a clean proof for smooth nonconvex objectives with bounded gradients, obtaining the tightest known dependence on β_1; Li et al. (2023) prove convergence to ϵ-stationary points under a generalized smoothness condition without requiring globally bounded gradients; and Hong & Lin (2024) extend this to affine variance noise models, showing that Adam is free to tune step sizes without knowledge of problem parameters. Most recently, Taniguchi et al. (2024) propose ADOPT, a modification of Adam that achieves the optimal $\mathcal{O}(1/\sqrt{T})$ nonconvex rate with any choice of β_2 without relying on bounded noise, by removing the current gradient from the second-moment estimate. The adaptivity of Adam relative to SGD has also been studied, with work identifying heavy-tailed gradient distributions and ill-conditioned objectives as settings where Adam's coordinate-wise preconditioning yields provable advantages (Kunstner et al., 2024; Zhang et al., 2024; Kunstner et al., 2023). Complementary recent work by Jin et al. (2026) gives a high-probability separation between Adam and SGD for stationary stochastic optimization under bounded-variance noise, showing that Adam's second-moment normalization can yield sharper confidence dependence than SGD through a stopping-time martingale analysis. All of these results assume a fixed stationary objective and none address the behavior of the preconditioner under distribution shift.

SGD and momentum in non-stationary settings. A complementary line of work studies optimization under time-varying objectives and distribution shift. The foundational dynamic regret framework of Zinkevich (2003) measures performance against a drifting comparator in online convex optimization, and has inspired a large body of follow-up work on efficient non-stationary online learning (Zhao et al., 2024). In stationary settings with constant step sizes, momentum methods are essentially equivalent at steady state to SGD with a rescaled learning rate and do not reduce the MSE floor (Yuan et al., 2016; Kovachki & Stuart, 2021). Under non-stationarity, this equivalence breaks down: past gradients become stale and may point in the wrong direction, making temporal averaging actively harmful. Cao et al. (2019) derive dynamic regret bounds for online SGD under time-varying distributions, while Cutler et al. (2023) establish non-asymptotic tracking guarantees for proximal stochastic gradient methods under time drift. Shen et al. (2026) establish optimal bounds for SGD under temporally dependent data. Most directly related to our work, Sahu et al. (2026) provide finite-time tracking bounds for SGD and its momentum variants, including Polyak's Heavy-Ball and Nesterov acceleration, under strongly convex smooth objectives, showing explicitly that momentum amplifies drift-induced tracking lag and establishing minimax lower bounds confirming this penalty is unavoidable. Nonconvex tracking guarantees under distribution shift remain scarce even for SGD, and our work is the first to establish such guarantees for Adam. Our work extends this line of inquiry to Adam, where the adaptive preconditioner introduces an additional and qualitatively distinct source of non-stationary error.

Adam and adaptive methods under non-stationarity. The failure of Adam under distribution shift has been documented empirically in several settings. Lyle et al. (2023) study plasticity loss in neural networks under non-stationary objectives and find that Adam's moment estimates degrade under task changes, recommending increased ϵ and more aggressive second-moment decay as fixes — a phenomenon our bounds now explain theoretically. Dohare et al. (2023) show empirically that loss of plasticity in continual learning is worse for Adam than for SGD, and Ellis et al. (2024) directly analyze how non-stationary gradient magnitudes in reinforcement learning cause Adam's updates to become excessively large, proposing a local-timestep variant to address this. Despite this empirical evidence, no prior work provides a theoretical framework that precisely characterizes when and why Adam degrades under distribution shift, how its $(\beta_1, \beta_2, \epsilon)$ hyperparameters interact with the drift-noise balance, or how it compares to SGD in a unified tracking setting. This paper fills that gap.

B PROBLEM SETUP, STANDING ASSUMPTIONS, AND NOTATION

We work on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and let $(X_t)_{t\geq 0}$ be an $\mathbb{F}$-adapted process. We take the natural filtration $\mathcal{F}_t := \sigma(X_0, \dots, X_t)$. For each $t \geq 0$, conditional mean gradient and conditional second moment by

$$\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t], \quad \boldsymbol{s}_{t+1}(\boldsymbol{\theta}) := \mathbb{E}\big[(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}))^{\odot 2} \mid \mathcal{F}_t\big].$$

We write $\boldsymbol{\theta}^\star_{t+1} \in \arg\min_{\boldsymbol{\theta}\in\Theta} G_{t+1}(\boldsymbol{\theta})$ for a (measurable) conditional minimizer where $\Theta \subset \mathbb{R}^d$ is compact and convex, and define the tracking error, minimizer drift, and shifted error as

$$\boldsymbol{e}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t, \quad \boldsymbol{\Delta}_t := \boldsymbol{\theta}^\star_t - \boldsymbol{\theta}^\star_{t+1}, \quad \boldsymbol{d}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_{t+1} = \boldsymbol{e}_t + \boldsymbol{\Delta}_t.$$

The associated first and second moment noise terms are

$$\boldsymbol{\xi}_{t+1} := \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t), \quad \boldsymbol{\chi}_{t+1} := (\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2} - \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t),$$

both of which are $\mathcal{F}_{t+1}$-measurable martingale differences satisfying $\mathbb{E}[\boldsymbol{\xi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\chi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ a.s.

Assumption B.1 (Stochastic predictability framework)**.** *There exists a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $(X_t)_{t\geq 0}$ be an $\mathbb{F}$-adapted process, i.e., X_t is $\mathcal{F}_t$-measurable for all t. For each $t \geq 0$, let Π_{t+1} denote the regular conditional law of X_{t+1} given $\mathcal{F}_t$, i.e., $\Pi_{t+1}(A) = \mathbb{P}(X_{t+1} \in A \mid \mathcal{F}_t)$ a.s. for every measurable set A, and assume Π_{t+1} is $\mathcal{F}_t$-measurable. Define the conditional risk*

$$G_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t] = \mathbb{E}_{X\sim\Pi_{t+1}}[g(\boldsymbol{\theta}, X)],$$

and let $\boldsymbol{\theta}^\star_{t+1} \in \arg\min_{\boldsymbol{\theta}\in\Theta} G_{t+1}(\boldsymbol{\theta})$ denote a (measurable) minimizer contained within the interior of Θ where $\Theta \subset \mathbb{R}^d$ is compact and convex. Assume the following hold for all $t \geq 0$:

1. ***(Predictable minimizer)*** *$\boldsymbol{\theta}^\star_{t+1}$ is $\mathcal{F}_t$-measurable.*
2. ***(Algorithm adaptedness)*** *The iterate $\boldsymbol{\theta}_t$ is $\mathcal{F}_t$-measurable.*
3. ***(Martingale difference noise)*** *The noise terms $\boldsymbol{\xi}_{t+1}$ and $\boldsymbol{\chi}_{t+1}$ defined above are $\mathcal{F}_{t+1}$-measurable and satisfy $\mathbb{E}[\boldsymbol{\xi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\chi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ a.s. for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

The Adam update Kingma & Ba (2015) with step size $\alpha > 0$, parameters $\beta_1, \beta_2 \in (0, 1)$, and $\epsilon > 0$, initialized with $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \mathbf{0}$, uses the uncorrected moment recursions $\boldsymbol{m}_{t+1} := \beta_1 \boldsymbol{m}_t + (1 - \beta_1)\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})$ and $\boldsymbol{v}_{t+1} := \beta_2 \boldsymbol{v}_t + (1 - \beta_2)(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2}$. The bias-corrected Adam update can then be written as

$$\begin{aligned}
\widehat{\boldsymbol{m}}_{t+1} &:= \frac{\boldsymbol{m}_{t+1}}{1 - \beta_1^{t+1}}, \\
\widehat{\boldsymbol{v}}_{t+1} &:= \frac{\boldsymbol{v}_{t+1}}{1 - \beta_2^{t+1}}, \\
\boldsymbol{P}_{t+1} &:= \operatorname{Diag}\left(\left(\sqrt{\widehat{\boldsymbol{v}}_{t+1}} + \epsilon\right)^{-1}\right), \\
\boldsymbol{\theta}_{t+1} &:= \mathcal{P}_\Theta^{\boldsymbol{A}_{t+1}}(\boldsymbol{\theta}_t - \alpha \boldsymbol{P}_{t+1} \widehat{\boldsymbol{m}}_{t+1}).
\end{aligned} \tag{Adam}$$

where $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) = \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}$, $(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2} = \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t) + \boldsymbol{\chi}_{t+1}$, and $\mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{z}\|_{\boldsymbol{A}}^2$.

The following three assumptions hold throughout all analyses:

Assumption B.2 (Uniform L-Lipschitz continuity)**.** *There exists $L > 0$ such that for all $t \geq 0$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,*

$$\|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}')\| \;\leq\; L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Assumption B.3 (Conditional sub-Gaussian gradient noise along iterates)**.** *There exists a constant $\sigma > 0$ such that for all $t \geq 0$, $\big\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\big\|_{\Psi_2|\mathcal{F}_t} \leq \sigma$ a.s.*

Bias-correction weights and constants. For $i \in \{1,2\}$ and $0 \le k \le t-1$, define the bias-correction weights

$$w_{i,t,k} := \frac{(1-\beta_i)\beta_i^{t-1-k}}{1-\beta_i^t},$$

which satisfy $\sum_{k=0}^{t-1} w_{i,t,k} = 1$. The following quantities appear throughout the analysis and are collected here for reference:

$$\kappa_{1,t} := \frac{(1-\beta_1)(1+\beta_1^t)}{(1+\beta_1)(1-\beta_1^t)}, \quad \vartheta_{2,t} := \beta_2^t + (1-\beta_2), \quad q_+ := \frac{1}{\epsilon}. \tag{3}$$

Here $\kappa_{1,t} = \sum_{k=0}^{t-1} w_{1,t,k}^2$ governs the variance of the bias-corrected first-moment average, $\vartheta_{2,t}$ captures the transient bias and stochastic fluctuation of the bias-corrected second moment, and q_+ is the deterministic upper spectral bound on the Adam preconditioners.

C Examples and interpretation of adaptive strong monotonicity

We provide several examples to illustrate the scope of adaptive strong monotonicity. In particular, adaptive strong monotonicity is not restricted to diagonal quadratics or objectives whose curvature is aligned with the adaptive preconditioner. It can hold for dense, nonseparable objectives such as generalized linear models and graph-coupled ranking problems, as well as for globally nonconvex objectives when the projected iterates remain in a locally contracting region. We also provide an explicit example showing that the assumption is nonvacuous: ordinary strong convexity alone does not guarantee adaptive strong monotonicity.

Throughout this section, fix t and suppress the time index when there is no ambiguity. When $\bar{\boldsymbol{g}}_{t+1} = \nabla F_{t+1}$, Θ is convex, and F_{t+1} is twice continuously differentiable, a direct sufficient condition for adaptive strong monotonicity is

$$\boldsymbol{S}_{t+1}(\boldsymbol{\theta}) := \frac{\widetilde{\boldsymbol{P}}_{t+1}\nabla^2 F_{t+1}(\boldsymbol{\theta}) + \nabla^2 F_{t+1}(\boldsymbol{\theta})\widetilde{\boldsymbol{P}}_{t+1}}{2} \succeq \mu \boldsymbol{I}$$

uniformly over t and $\boldsymbol{\theta} \in \Theta$. Indeed, if $\boldsymbol{u} := \boldsymbol{\theta} - \boldsymbol{\theta}'$, then the fundamental theorem of calculus gives $\nabla F_{t+1}(\boldsymbol{\theta}) - \nabla F_{t+1}(\boldsymbol{\theta}') = \int_0^1 \nabla^2 F_{t+1}(\boldsymbol{\theta}' + s\boldsymbol{u})\boldsymbol{u}\, ds$. Since $\boldsymbol{u}^\top \widetilde{\boldsymbol{P}}_{t+1}\nabla^2 F_{t+1}\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{S}_{t+1}\boldsymbol{u}$, integration yields

$$\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \widetilde{\boldsymbol{P}}_{t+1}\left(\nabla F_{t+1}(\boldsymbol{\theta}) - \nabla F_{t+1}(\boldsymbol{\theta}')\right)\right\rangle \ge \mu \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2.$$

Thus each example below amounts to establishing a uniform positive lower bound on the symmetric preconditioned Hessian.

Example C.1 (Parameter-separable objectives)**.** Consider $F(\boldsymbol{\theta}) = \sum_{j=1}^d f_j(\theta_j)$, where $f_j''(\theta_j) \ge \mu_0 > 0$, together with the diagonal adaptive preconditioner $\widetilde{\boldsymbol{P}} = \mathrm{Diag}(p_1, \ldots, p_d)$, with $p_j > 0$. Then, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,

$$\begin{aligned}\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \widetilde{\boldsymbol{P}}\left(\nabla F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta}')\right)\right\rangle &= \sum_{j=1}^d p_j(\theta_j - \theta_j')\left(f_j'(\theta_j) - f_j'(\theta_j')\right) \\ &\ge \mu_0 \left(\min_{j\in[d]} p_j\right)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2.\end{aligned}$$

Thus, at time t, the preconditioned map is strongly monotone with modulus $\mu_t = \mu_0 \min_j p_{t+1,j}$. In particular, if $\inf_t \min_j p_{t+1,j} > 0$, then Assumption D.1 holds uniformly with $\mu = \mu_0 \inf_t \min_j p_{t+1,j}$.

Scalar and diagonal strongly convex quadratics are immediate special cases, but the argument is not specific to quadratic losses. It also applies to coordinatewise M-estimation, separable stochastic regression, and other coordinate-separable stochastic programs. The same argument extends directly to block-separable objectives with the corresponding block-diagonal preconditioner.

Example C.2 (Dense generalized linear models)**.** Consider the conditional regularized generalized linear model

$$F_{t+1}(\boldsymbol{\theta}) = \mathbb{E}_t\left[b(\boldsymbol{X}_{t+1}^\top\boldsymbol{\theta}) - Y_{t+1}\boldsymbol{X}_{t+1}^\top\boldsymbol{\theta}\right] + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2.$$

Its Hessian is $\nabla^2 F_{t+1}(\boldsymbol{\theta}) = \mathbb{E}_t[b''(\boldsymbol{X}_{t+1}^\top\boldsymbol{\theta})\boldsymbol{X}_{t+1}\boldsymbol{X}_{t+1}^\top] + \lambda\boldsymbol{I}$. Suppose, uniformly over $\boldsymbol{\theta} \in \Theta$, that $0 \le \underline{b} \le b''(\boldsymbol{X}_{t+1}^\top\boldsymbol{\theta}) \le \overline{b}$ and

$$\nu\boldsymbol{I} \preceq \mathbb{E}_t[\boldsymbol{X}_{t+1}\boldsymbol{X}_{t+1}^\top] \preceq M_X\boldsymbol{I}.$$

Then $m\boldsymbol{I} \preceq \nabla^2 F_{t+1}(\boldsymbol{\theta}) \preceq M\boldsymbol{I}$, where $m := \lambda + \underline{b}\,\nu$ and $M := \lambda + \overline{b}\,M_X$.

The conditional covariance may be fully dense and need not share eigenvectors with $\widetilde{\boldsymbol{P}}_{t+1}$. Let

$$\underline{p}_t := \lambda_{\min}(\widetilde{\boldsymbol{P}}_{t+1}),\ \ \overline{p}_t := \lambda_{\max}(\widetilde{\boldsymbol{P}}_{t+1}),\ \ \kappa_{P,t} := \frac{\overline{p}_t}{\underline{p}_t}.$$

Write $\widetilde{\boldsymbol{P}}_{t+1} = p_{0,t}\boldsymbol{I} + \boldsymbol{E}_t$, where $p_{0,t} := (\underline{p}_t + \overline{p}_t)/2$. Then $\|\boldsymbol{E}_t\| \le (\overline{p}_t - \underline{p}_t)/2$. If $\boldsymbol{H} := \nabla^2 F_{t+1}(\boldsymbol{\theta})$, then for every unit vector $\boldsymbol{v}$,

$$\begin{aligned}\boldsymbol{v}^\top\frac{\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{H} + \boldsymbol{H}\widetilde{\boldsymbol{P}}_{t+1}}{2}\boldsymbol{v} &= p_{0,t}\boldsymbol{v}^\top\boldsymbol{H}\boldsymbol{v} + \boldsymbol{v}^\top\boldsymbol{E}_t\boldsymbol{H}\boldsymbol{v}\\ &\ge p_{0,t}m - \|\boldsymbol{E}_t\|M\\ &= \underline{p}_t\left[m - \frac{\kappa_{P,t}-1}{2}(M-m)\right].\end{aligned}$$

Therefore, the preconditioned mean-gradient map is strongly monotone whenever

$$m > \frac{\kappa_{P,t}-1}{2}(M-m), \tag{4}$$

with local modulus

$$\mu_t = \underline{p}_t\left[m - \frac{\kappa_{P,t}-1}{2}(M-m)\right].$$

A uniform lower bound $\inf_t \mu_t > 0$ verifies Assumption D.1.

Substituting the definitions of m and M, Eq. (4) is equivalent to

$$\lambda + \underline{b}\,\nu > \frac{\kappa_{P,t}-1}{2}\left(\overline{b}\,M_X - \underline{b}\,\nu\right).$$

For linear regression, $b'' \equiv 1$, so the condition becomes $\lambda + \nu > (\kappa_{P,t}-1)(M_X - \nu)/2$. Thus the feature covariance may be fully dense and arbitrarily rotated relative to the adaptive coordinates.

For logistic regression, $b''(z) = e^z/(1+e^z)^2 \le 1/4$. If the projected parameter region and covariates satisfy $|\boldsymbol{X}^\top\boldsymbol{\theta}| \le B$, then $b''(\boldsymbol{X}^\top\boldsymbol{\theta}) \ge e^B/(1+e^B)^2 =: \underline{b}_B > 0$. Consequently, regularized logistic regression, and unregularized logistic regression under sufficiently well-conditioned covariance and adaptive scaling, also satisfy the condition. Linear and logistic regression are among the examples used in our experiments.

This example shows directly that adaptive strong monotonicity does not require the Hessian and adaptive preconditioner to commute; it only requires their anisotropies to be sufficiently compatible.

Example C.3 (Pairwise logistic regression and graph-ranking models)**.** Consider

$$F(\boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \sum_{(i,j)\in E} a_{ij}\log\left(1 + \exp\left(-y_{ij}(\theta_i - \theta_j)\right)\right),$$

where $a_{ij} \ge 0$, $y_{ij} \in \{-1, 1\}$, and $\widetilde{\boldsymbol{P}} = \mathrm{Diag}(p_1, \ldots, p_d)$. This includes regularized Bradley–Terry estimation and pairwise graph-ranking models.

For an edge (i,j), let $h_{ij} := \ell''_{ij}(\theta_i - \theta_j)$, where $\ell_{ij}(z) = a_{ij}\log(1+\exp(-y_{ij}z))$. Since $0 \le h_{ij} \le a_{ij}/4$, the Hessian has weighted graph-Laplacian structure,

$$H_{ii} = \lambda + \sum_{j:(i,j)\in E} h_{ij},\ \ H_{ij} = -h_{ij} \text{ for } (i,j) \in E.$$

For $\boldsymbol{S} = (\widetilde{\boldsymbol{P}}\boldsymbol{H} + \boldsymbol{H}\widetilde{\boldsymbol{P}})/2$, we have $S_{ii} = p_i(\lambda + \sum_j h_{ij})$ and $S_{ij} = -(p_i + p_j)h_{ij}/2$. Gershgorin's theorem therefore gives

$$\lambda_{\min}(\boldsymbol{S}) \geq \min_i \left\{ \lambda p_i + \frac{1}{2} \sum_{j:(i,j)\in E} (p_i - p_j)h_{ij} \right\}$$
$$\geq \min_i \left\{ \lambda p_i - \frac{1}{8} \sum_{j:(i,j)\in E} a_{ij}|p_i - p_j| \right\}.$$

Thus adaptive strong monotonicity with modulus $\mu > 0$ holds whenever

$$\lambda p_i - \frac{1}{8} \sum_{j:(i,j)\in E} a_{ij}|p_i - p_j| \geq \mu \text{ for every } i \in [d]. \tag{5}$$

This condition has a direct interpretation. The edge weights a_{ij} measure the strength of the statistical coupling between coordinates, while $|p_i - p_j|$ measures the discrepancy between their adaptive scales. Adaptive scaling is therefore benign when strongly coupled coordinates receive similar preconditioning. If $p_i = p_j$ along every edge, the graph-induced penalty vanishes regardless of graph density. More generally, dense coupling is allowed provided the adaptive scales vary sufficiently slowly across strongly interacting coordinates.

Example C.4 (A globally nonconvex objective)**.** Adaptive strong monotonicity only needs to hold on the projected tracking region Θ. Thus the objective may be globally nonconvex while the preconditioned gradient remains contracting on the region visited by the algorithm.

Consider $F(x, y) = x^4/4 - x^2/2 + y^2/2 + cxy$ and $\widetilde{\boldsymbol{P}} = \mathrm{Diag}(p_1, p_2)$. The objective is globally nonconvex because

$$\nabla^2 F(0, 0) = \begin{pmatrix} -1 & c \\ c & 1 \end{pmatrix}$$

is indefinite. Restrict instead to $\Theta = [a, b] \times [-R, R]$, where $a > 1/\sqrt{3}$, and let $m := 3a^2 - 1 > 0$. On this region,

$$\nabla^2 F(x, y) = \begin{pmatrix} 3x^2 - 1 & c \\ c & 1 \end{pmatrix}, \quad 3x^2 - 1 \geq m.$$

Hence

$$\frac{\widetilde{\boldsymbol{P}}\nabla^2 F + \nabla^2 F\widetilde{\boldsymbol{P}}}{2} \succeq \begin{pmatrix} p_1 m & (p_1 + p_2)c/2 \\ (p_1 + p_2)c/2 & p_2 \end{pmatrix}.$$

The matrix on the right is positive definite precisely when

$$4p_1 p_2 m > (p_1 + p_2)^2 c^2. \tag{6}$$

Equivalently, $|c| < 2\sqrt{p_1 p_2 m}/(p_1 + p_2)$. Under this condition, its smallest eigenvalue is

$$\gamma = \frac{1}{2}\left[p_1 m + p_2 - \sqrt{(p_1 m - p_2)^2 + (p_1 + p_2)^2 c^2}\right] > 0,$$

so adaptive strong monotonicity holds on Θ with modulus $\mu = \gamma$.

This example is both nonseparable and globally nonconvex. Projection restricts the iterates to a region of positive local curvature, while Eq. (6) controls the strength of the cross-coordinate interaction relative to that curvature and the adaptive scales.

Example C.5 (Strong convexity does not imply adaptive strong monotonicity)**.** Finally, adaptive strong monotonicity imposes a genuine compatibility condition between objective curvature and adaptive scaling. Consider the strongly convex quadratic $F(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta}/2$, where

$$\boldsymbol{H} = \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \quad \widetilde{\boldsymbol{P}} = \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}.$$

Assume $a, b, p_1, p_2 > 0$ and $ab > c^2$, so $\boldsymbol{H} \succ 0$. The symmetric preconditioned Hessian is

$$\boldsymbol{S} = \frac{\widetilde{\boldsymbol{P}}\boldsymbol{H} + \boldsymbol{H}\widetilde{\boldsymbol{P}}}{2} = \begin{pmatrix} p_1 a & (p_1 + p_2)c/2 \\ (p_1 + p_2)c/2 & p_2 b \end{pmatrix}.$$

Although $\boldsymbol{H} \succ 0$, $\boldsymbol{S} \succ 0$ only when

$$4p_1p_2ab > (p_1 + p_2)^2c^2. \tag{7}$$

Define $\rho_H := |c|/\sqrt{ab}$ and $\rho_P := 2\sqrt{p_1p_2}/(p_1 + p_2)$. Then Eq. (7) is exactly $\rho_H < \rho_P$: stronger cross-coordinate coupling requires more balanced adaptive scaling.

For a concrete case where the converse is not true, take

$$\boldsymbol{H} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad \widetilde{\boldsymbol{P}} = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Hessian eigenvalues are $1 - r$ and $1 + r$, so the objective is strongly convex whenever $|r| < 1$. However,

$$\boldsymbol{S} = \begin{pmatrix} 9 & 5r \\ 5r & 1 \end{pmatrix}, \quad \det(\boldsymbol{S}) = 9 - 25r^2.$$

Thus adaptive strong monotonicity requires $|r| < 3/5$. Taking $r = 4/5$, for example, gives a strongly convex objective with Hessian eigenvalues $1/5$ and $9/5$, while $\det(\boldsymbol{S}) = -7 < 0$. Adaptive strong monotonicity therefore does not hold.

This example makes the distinction precise: strong convexity controls the geometry of ∇F, whereas adaptive strong monotonicity controls the geometry after coordinate-dependent preconditioning.

Taken together, these examples show that Assumption D.1 is substantive but not restricted to diagonal or quadratic objectives. It applies to separable problems with uniformly positive adaptive scales, dense generalized linear models, graph-coupled estimation problems, and locally contracting regions of globally nonconvex objectives. At the same time, Example C.5 shows that sufficiently strong cross-coordinate curvature can destroy contraction when it is incompatible with the adaptive scaling.

D Proofs for tracking under adaptive strong monotonicity

For this section, we will assume the preconditioned conditional mean gradient map $\boldsymbol{\theta} \mapsto \widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta})$ is μ-strongly monotone.

Assumption D.1 (Adaptive strong monotonicity)**.** *There exists $0 < \mu < \infty$ such that, almost surely, for all $t \geq 0$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}',\, \widetilde{\boldsymbol{P}}_{t+1}\left(\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}')\right)\right\rangle \geq \mu\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$.*

D.1 Proofs for Theorem 3.1

We first introduce the predictable preconditioner proxy used throughout the Adam analysis. The realized Adam preconditioner $\boldsymbol{P}_{t+1}$ is computed from the fresh sample X_{t+1}, and is therefore $\mathcal{F}_{t+1}$-measurable rather than $\mathcal{F}_t$-measurable. Since our tracking argument imposes contraction on a predictable preconditioned mean-gradient map, we separate the predictable geometry from the random second-moment fluctuation.

To do so, we construct a proxy by replacing the uncorrected second moment with its conditional expectation. Define $\widetilde{\boldsymbol{v}}_{t+1} := \mathbb{E}[\boldsymbol{v}_{t+1} \mid \mathcal{F}_t]$ and $\widetilde{\boldsymbol{P}}_{t+1} := [\mathrm{Diag}(\sqrt{\widetilde{\boldsymbol{v}}_{t+1}}) + \epsilon \boldsymbol{I}]^{-1}$. Since $\boldsymbol{v}_{t+1} = \beta_2\boldsymbol{v}_t + (1-\beta_2)(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2}$, taking conditional expectation given $\mathcal{F}_t$ yields $\widetilde{\boldsymbol{v}}_{t+1} = \beta_2\boldsymbol{v}_t + (1-\beta_2)\boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t)$. Thus both $\widetilde{\boldsymbol{v}}_{t+1}$ and $\widetilde{\boldsymbol{P}}_{t+1}$ are $\mathcal{F}_t$-measurable.

We next rewrite the Adam update around this predictable proxy. Define the first-moment tracking error by $\boldsymbol{r}_{t+1} := \widehat{\boldsymbol{m}}_{t+1} - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)$ and the preconditioner perturbation by $\boldsymbol{\eta}_{t+1} := (\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1})\widehat{\boldsymbol{m}}_{t+1}$. Then

$$\boldsymbol{P}_{t+1}\widehat{\boldsymbol{m}}_{t+1} = \widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) + \widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1} + \boldsymbol{\eta}_{t+1}.$$

Consequently, for the pre-projection point $\boldsymbol{z}_{t+1} := \boldsymbol{\theta}_t - \alpha\boldsymbol{P}_{t+1}\widehat{\boldsymbol{m}}_{t+1}$, we obtain the predictable decomposition

$$\boldsymbol{z}_{t+1} = \boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) - \alpha\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1} - \alpha\boldsymbol{\eta}_{t+1}, \tag{8}$$

with $\boldsymbol{\theta}_{t+1} = \mathcal{P}_\Theta(\boldsymbol{z}_{t+1})$.

We will analyze the exact Adam update through this decomposition: the first term is the predictable preconditioned mean-gradient step, $\boldsymbol{r}_{t+1}$ measures the error from using Adam's first-moment estimate instead of the current conditional mean gradient, and $\boldsymbol{\eta}_{t+1}$ captures the price of replacing the random preconditioner by its predictable proxy.

We first establish uniform high-probability control of both the stochastic gradient and the centered squared-gradient fluctuation. These bounds will then be used to control the bias-corrected second moment and the corresponding preconditioner perturbation.

Lemma D.1 (Uniform high-probability control of the stochastic gradient)**.** *Suppose Assumption B.3 holds and* $\operatorname{diam}(\Theta) \le R$*. Then there exists a universal constant* $C > 0$ *such that, for every* $T \ge 1$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, simultaneously for every* $t \in \{0,\dots,T-1\}$*, the following statements hold:*

1. *The stochastic gradient satisfies*

$$\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\| \le LR + C\sigma\sqrt{d + \log\frac{4T}{\delta}}. \tag{9}$$

2. *The centered coordinatewise squared-gradient fluctuation satisfies*

$$\left\|(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2} - \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t)\right\|_\infty \le C\left(LR\sigma\sqrt{\log\frac{4dT}{\delta}} + \sigma^2\log\frac{4dT}{\delta}\right). \tag{10}$$

3. *The squared gradient coordinates satisfy*

$$\left\|(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2}\right\|_\infty \le C\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right). \tag{11}$$

Proof of Lemma D.1. Since $\boldsymbol{\theta}^\star_{t+1}$ lies in the interior of Θ, the first-order optimality condition gives $\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}^\star_{t+1}) = \mathbf{0}$. Therefore, by the L-Lipschitz continuity of $\bar{\boldsymbol{g}}_{t+1}$ and the fact that $\operatorname{diam}(\Theta) \le R$,

$$\|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\| = \left\|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}^\star_{t+1})\right\| \le L\left\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_{t+1}\right\| \le LR. \tag{12}$$

By definition, $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) = \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)$. Under Assumption B.3, for every $\boldsymbol{u} \in \mathbb{S}^{d-1}$ $\|\langle \boldsymbol{u}, \boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\rangle\|_{\Psi_2|\mathcal{F}_t} \le \sigma$ a.s. Thus, by the one-step conditional form of Theorem F.1 (see also (Vershynin, 2018, Proposition 2.6.1)), there exists a universal constant $c > 0$ such that, conditionally on $\mathcal{F}_t$,

$$\mathbb{P}\left(|\langle \boldsymbol{u}, \boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\rangle| \ge x \mid \mathcal{F}_t\right) \le 2\exp\left(-\frac{cx^2}{\sigma^2}\right).$$

Let $\mathcal{N} \subseteq \mathbb{S}^{d-1}$ be a $1/2$-net of $\mathbb{S}^{d-1}$. Standard volumetric arguments then imply that $|\mathcal{N}| \le 5^d$ (Vershynin, 2018, Corollary 4.2.11). By (Vershynin, 2018, Exercise 4.35), $\|\boldsymbol{x}\| \le 2\max_{\boldsymbol{u}\in\mathcal{N}} |\langle \boldsymbol{u}, \boldsymbol{x}\rangle|$ for every $\boldsymbol{x} \in \mathbb{R}^d$, a union bound over $\mathcal{N}$ gives

$$\mathbb{P}\left(\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\| \ge C\sigma\sqrt{d+x} \,\middle|\, \mathcal{F}_t\right) \le 2e^{-x}$$

for every $x > 0$. Taking expectations, choosing $x = \log(4T/\delta)$, and applying a union bound over $t = 0,\dots,T-1$ yields, with probability at least $1-\delta/2$,

$$\max_{0\le t\le T-1} \|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\| \le C\sigma\sqrt{d + \log\frac{4T}{\delta}}.$$

Together with Eq. (12), this gives

$$\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\| \le LR + C\sigma\sqrt{d + \log\frac{4T}{\delta}},$$

simultaneously for all $t \in \{0,\dots,T-1\}$, which proves Eq. (9).

We next control the centered squared-gradient fluctuation. Fix $t \in \{0, \dots, T-1\}$ and $i \in [d]$. Using $\nabla_{\theta_i} g(\boldsymbol{\theta}_t, X_{t+1}) = \bar{g}_{t+1,i}(\boldsymbol{\theta}_t) + \xi_{t+1,i}(\boldsymbol{\theta}_t)$ and $\mathbb{E}[\xi_{t+1,i}(\boldsymbol{\theta}_t) \mid \mathcal{F}_t] = 0$, we have $s_{t+1,i}(\boldsymbol{\theta}_t) = \bar{g}_{t+1,i}(\boldsymbol{\theta}_t)^2 + \mathbb{E}\big[\xi_{t+1,i}(\boldsymbol{\theta}_t)^2 \mid \mathcal{F}_t\big]$. Consequently,

$$\begin{aligned}&\left(\nabla_{\theta_i} g(\boldsymbol{\theta}_t, X_{t+1})\right)^2 - s_{t+1,i}(\boldsymbol{\theta}_t) \\ &\quad = 2\bar{g}_{t+1,i}(\boldsymbol{\theta}_t)\xi_{t+1,i}(\boldsymbol{\theta}_t) + \xi_{t+1,i}(\boldsymbol{\theta}_t)^2 - \mathbb{E}\big[\xi_{t+1,i}(\boldsymbol{\theta}_t)^2 \mid \mathcal{F}_t\big]\,. \end{aligned} \tag{13}$$

Assumption B.3 implies $\|\xi_{t+1,i}(\boldsymbol{\theta}_t)\|_{\Psi_2|\mathcal{F}_t} \le \sigma$. Hence by (Vershynin, 2018, Exercise 2.44), we have

$$\big\|\xi_{t+1,i}(\boldsymbol{\theta}_t)^2 - \mathbb{E}\big[\xi_{t+1,i}(\boldsymbol{\theta}_t)^2 \mid \mathcal{F}_t\big]\big\|_{\Psi_1|\mathcal{F}_t} \le C\sigma^2.$$

Moreover, by Eq. (12), $|\bar{g}_{t+1,i}(\boldsymbol{\theta}_t)| \le LR$, so $2\bar{g}_{t+1,i}(\boldsymbol{\theta}_t)\xi_{t+1,i}(\boldsymbol{\theta}_t)$ is conditionally sub-Gaussian with scale at most $CLR\sigma$. Therefore, conditional Bernstein's inequality (Theorem F.2) applied to Eq. (13) gives, for every $x \ge 1$,

$$\mathbb{P}\left(\left| \left(\nabla_{\theta_i} g(\boldsymbol{\theta}_t, X_{t+1})\right)^2 - s_{t+1,i}(\boldsymbol{\theta}_t) \right| > C\left(LR\sigma\sqrt{x} + \sigma^2 x\right) \,\middle|\, \mathcal{F}_t \right) \le 2e^{-x}. \tag{14}$$

Taking expectations, choosing $x = \log(4dT/\delta)$, and applying a union bound over $t = 0, \dots, T-1$ and $i \in [d]$ yields, with probability at least $1 - \delta/2$,

$$\max_{\substack{0 \le t \le T-1 \\ i \in [d]}} \left| \left(\nabla_{\theta_i} g(\boldsymbol{\theta}_t, X_{t+1})\right)^2 - s_{t+1,i}(\boldsymbol{\theta}_t) \right| \le C\left(LR\sigma\sqrt{\log\frac{4dT}{\delta}} + \sigma^2 \log\frac{4dT}{\delta} \right),$$

which proves Eq. (10).

It remains to obtain a direct bound on the squared gradient coordinates. Conditional sub-Gaussianity implies $\mathbb{E}\big[\xi_{t+1,i}(\boldsymbol{\theta}_t)^2 \mid \mathcal{F}_t\big] \le C\sigma^2$, and hence $s_{t+1,i}(\boldsymbol{\theta}_t) \le L^2R^2 + C\sigma^2$. Combining this with Eq. (10) gives

$$\begin{aligned}\left|\nabla_{\theta_i} g(\boldsymbol{\theta}_t, X_{t+1})\right|^2 &\le L^2R^2 + C\sigma^2 + CLR\sigma\sqrt{\log\frac{4dT}{\delta}} + C\sigma^2\log\frac{4dT}{\delta} \\ &\le C\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right),\end{aligned}$$

where the last inequality follows from Young's inequality. Taking the maximum over $i \in [d]$ proves Eq. (11). Intersecting the two preceding events completes the proof. □

Having established uniform concentration of the stochastic gradient and its coordinatewise square, we can now control the discrepancy between the realized Adam preconditioner and its predictable proxy. The argument separates naturally into two pieces: the transient error introduced by bias correction of the second moment and the fresh second-moment fluctuation represented by $\boldsymbol{\chi}_{t+1}$.

Lemma D.2 (High-probability control of the preconditioner perturbation)**.** *Suppose Assumption B.3 holds. Then there exists a universal constant $C > 0$ such that, for every $T \ge 1$ and $\delta \in (0,1)$, with probability at least $1 - \delta$, simultaneously for every $t \in \{0, \dots, T-1\}$,*

$$\|\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1}\|_{\mathrm{op}}^2 \le C\epsilon^{-4}\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right)\vartheta_{2,t+1}. \tag{15}$$

Moreover,

$$\|\boldsymbol{\eta}_{t+1}\|^2 \le C\epsilon^{-4}\left(L^2R^2 + \sigma^2\left(d + \log\frac{4T}{\delta}\right)\right)^2\vartheta_{2,t+1}. \tag{16}$$

Proof of Lemma D.2. Since both $\boldsymbol{P}_{t+1}$ and $\widetilde{\boldsymbol{P}}_{t+1}$ are diagonal, it suffices to compare their entries coordinatewise. For any $a, b \ge 0$, $\left|(\sqrt{a}+\epsilon)^{-1} - (\sqrt{b}+\epsilon)^{-1}\right| \le \epsilon^{-2}|\sqrt{a} - \sqrt{b}| \le \epsilon^{-2}\sqrt{|a-b|}$. Applying this inequality coordinatewise yields $\|\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1}\|_{\mathrm{op}}^2 \le \epsilon^{-4}\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty$.

It therefore remains to control the discrepancy between the bias-corrected second moment $\widehat{\boldsymbol{v}}_{t+1}$ and its predictable proxy $\widetilde{\boldsymbol{v}}_{t+1}$. Since $\widehat{\boldsymbol{v}}_{t+1} = \boldsymbol{v}_{t+1}/(1-\beta_2^{t+1})$, we may write $\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1} = \frac{\beta_2^{t+1}}{1-\beta_2^{t+1}}\boldsymbol{v}_{t+1} + \boldsymbol{v}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}$.

The first term is the transient bias-correction error. Since $\boldsymbol{v}_0 = \mathbf{0}$, unrolling the second-moment recursion gives $\boldsymbol{v}_{t+1} = (1-\beta_2)\sum_{j=0}^{t}\beta_2^{t-j}(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_j, X_{j+1}))^{\odot 2}$. By Lemma D.1 (Eq. (11)), with probability at least $1-\delta$, simultaneously for every $j \in \{0,\dots,T-1\}$,

$$\|\boldsymbol{v}_{t+1}\|_\infty \le C(1-\beta_2)\sum_{j=0}^{t}\beta_2^{t-j}\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right) = C(1-\beta_2^{t+1})\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right). \tag{17}$$

For the second term, the definition of $\widetilde{\boldsymbol{v}}_{t+1}$ gives $\boldsymbol{v}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1} = (1-\beta_2)[(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}))^{\odot 2} - \boldsymbol{s}_{t+1}(\boldsymbol{\theta}_t)] = (1-\beta_2)\boldsymbol{\chi}_{t+1}$. Hence, by Lemma D.1 (Eq. (10)),

$$\|\boldsymbol{v}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty \le C(1-\beta_2)\left(LR\sigma\sqrt{\log\frac{4dT}{\delta}} + \sigma^2\log\frac{4dT}{\delta}\right). \tag{18}$$

Combining the preceding two bounds with the decomposition of $\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}$ yields

$$\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty \le C\left[\beta_2^{t+1}\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right) + (1-\beta_2)\left(LR\sigma\sqrt{\log\frac{4dT}{\delta}} + \sigma^2\log\frac{4dT}{\delta}\right)\right].$$

Young's inequality gives $LR\sigma\sqrt{\log(4dT/\delta)} \le \frac{1}{2}L^2R^2 + \frac{1}{2}\sigma^2\log(4dT/\delta)$. Recalling that $\vartheta_{2,t+1} := \beta_2^{t+1} + (1-\beta_2)$, we therefore obtain

$$\|\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1}\|_{\mathrm{op}}^2 \le C\epsilon^{-4}\left(L^2R^2 + \sigma^2\log\frac{4dT}{\delta}\right)\vartheta_{2,t+1},$$

which proves Eq. (15).

It remains to control $\widehat{\boldsymbol{m}}_{t+1}$. Unrolling the bias-corrected first moment gives $\widehat{\boldsymbol{m}}_{t+1} = \sum_{j=0}^{t} w_{1,t+1,j}\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_j, X_{j+1})$, where $w_{1,t+1,j} \ge 0$ and $\sum_{j=0}^{t} w_{1,t+1,j} = 1$. By convexity of the squared norm, $\|\widehat{\boldsymbol{m}}_{t+1}\|^2 \le \sum_{j=0}^{t} w_{1,t+1,j}\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_j, X_{j+1})\|^2$. Therefore, by Lemma D.1 (Eq. (9)), $\|\widehat{\boldsymbol{m}}_{t+1}\|^2 \le C\left(L^2R^2 + \sigma^2(d + \log(4T/\delta))\right)$.

Finally, by definition, $\boldsymbol{\eta}_{t+1} = (\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1})\widehat{\boldsymbol{m}}_{t+1}$, and hence $\|\boldsymbol{\eta}_{t+1}\|^2 \le \|\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1}\|_{\mathrm{op}}^2\|\widehat{\boldsymbol{m}}_{t+1}\|^2$. Combining the preceding two estimates gives

$$\|\boldsymbol{\eta}_{t+1}\|^2 \le C\epsilon^{-4}\left(L^2R^2 + \sigma^2\left(d + \log\frac{4T}{\delta}\right)\right)^2\vartheta_{2,t+1},$$

which proves Eq. (16). □

We now use the predictable decomposition in Eq. (8) to derive the one-step tracking recursion. This proof follows similarly to Sahu et al. (2026); Cutler et al. (2021).

Lemma D.3 (One-step recursive relation)**.** *Assume Assumption D.1 and suppose* $\alpha \le \min\left\{\frac{\mu\epsilon^2}{4L^2}, \frac{1}{\mu}\right\}$. *Then, for every* $t \ge 0$,

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \le \left(1 - \frac{1}{2}\alpha\mu\right)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 + \frac{5}{\alpha\mu}\|\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star\|^2 + \frac{10\alpha\epsilon^{-2}}{\mu}\|\boldsymbol{r}_{t+1}\|^2 + \frac{10\alpha}{\mu}\|\boldsymbol{\eta}_{t+1}\|^2. \tag{19}$$

Proof of Lemma D.3. Define the pre-projection point $\boldsymbol{z}_{t+1} := \boldsymbol{\theta}_t - \alpha\boldsymbol{P}_{t+1}\widehat{\boldsymbol{m}}_{t+1}$. By the predictable decomposition in Eq. (8), $\boldsymbol{z}_{t+1} = \boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t) - \alpha\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1} - \alpha\boldsymbol{\eta}_{t+1}$. Since $\boldsymbol{\theta}_{t+1} =$

$\mathcal{P}_\Theta(\boldsymbol{z}_{t+1})$ and $\boldsymbol{\theta}^\star_{t+1} \in \Theta$, nonexpansiveness of the Euclidean projection (Lemma F.1) gives $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \|\boldsymbol{z}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2$.

Recall $\boldsymbol{d}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_{t+1}$, and define $\boldsymbol{u}_t := \boldsymbol{d}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)$. Then $\boldsymbol{z}_{t+1} - \boldsymbol{\theta}^\star_{t+1} = \boldsymbol{u}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1} - \alpha\boldsymbol{\eta}_{t+1}$. Since $\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}^\star_{t+1}) = \boldsymbol{0}$, adaptive strong monotonicity (Assumption D.1) implies $\left\langle \boldsymbol{d}_t,\, \widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\right\rangle \ge \mu\|\boldsymbol{d}_t\|^2$.

Moreover, by L-Lipschitz continuity of $\bar{\boldsymbol{g}}_{t+1}$, $\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}^\star_{t+1}) = 0$, and $\widetilde{\boldsymbol{P}}_{t+1} \preceq q_+\boldsymbol{I}$, we have $\|\widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\|^2 \le q_+^2L^2\|\boldsymbol{d}_t\|^2$. Consequently,

$$\begin{aligned}\|\boldsymbol{u}_t\|^2 &= \|\boldsymbol{d}_t\|^2 - 2\alpha\left\langle \boldsymbol{d}_t,\, \widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\right\rangle + \alpha^2\|\widetilde{\boldsymbol{P}}_{t+1}\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\|^2\\ &\le \left(1 - 2\alpha\mu + \alpha^2q_+^2L^2\right)\|\boldsymbol{d}_t\|^2 \le (1-\alpha\mu)\|\boldsymbol{d}_t\|^2, \end{aligned} \tag{20}$$

where the last inequality follows from $\alpha \le \mu/(4q_+^2L^2)$.

We now add back the two Adam-specific error terms. Let $\tau := \alpha\mu/4$. By Young's inequality and the projection inequality above,

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 &\le (1+\tau)\|\boldsymbol{u}_t\|^2 + (1+\tau^{-1})\alpha^2\|\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1} + \boldsymbol{\eta}_{t+1}\|^2\\ &\le (1+\tau)\|\boldsymbol{u}_t\|^2 + 2(1+\tau^{-1})\alpha^2\left(q_+^2\|\boldsymbol{r}_{t+1}\|^2 + \|\boldsymbol{\eta}_{t+1}\|^2\right). \end{aligned} \tag{21}$$

Since $\alpha\mu \le 1$, we have $\tau \le 1/4$, and hence $(1+\tau)(1-\alpha\mu) \le 1 - \frac{3}{4}\alpha\mu$. Also, $2(1+\tau^{-1})\alpha^2 = 2\alpha^2 + 8\alpha/\mu \le 10\alpha/\mu$. Combining these bounds with Eqs. (20) and (21) yields

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \left(1 - \frac{3}{4}\alpha\mu\right)\|\boldsymbol{d}_t\|^2 + \frac{10\alpha q_+^2}{\mu}\|\boldsymbol{r}_{t+1}\|^2 + \frac{10\alpha}{\mu}\|\boldsymbol{\eta}_{t+1}\|^2. \tag{22}$$

It remains to express the shifted error $\boldsymbol{d}_t$ in terms of the current tracking error and the minimizer drift. Since $\boldsymbol{d}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t + \boldsymbol{\theta}^\star_t - \boldsymbol{\theta}^\star_{t+1}$, Young's inequality with parameter $\alpha\mu/4$ gives

$$\begin{aligned}\|\boldsymbol{d}_t\|^2 &\le \left(1 + \frac{\alpha\mu}{4}\right)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \left(1 + \frac{4}{\alpha\mu}\right)\|\boldsymbol{\theta}^\star_t - \boldsymbol{\theta}^\star_{t+1}\|^2\\ &\le \left(1 + \frac{\alpha\mu}{4}\right)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \frac{5}{\alpha\mu}\|\boldsymbol{\theta}^\star_t - \boldsymbol{\theta}^\star_{t+1}\|^2, \end{aligned} \tag{23}$$

where the last step again uses $\alpha\mu \le 1$. Finally, $\left(1 - \frac{3}{4}\alpha\mu\right)\left(1 + \frac{1}{4}\alpha\mu\right) \le 1 - \frac{1}{2}\alpha\mu$. Substituting Eq. (23) into Eq. (22) and using $q_+^2 = \epsilon^{-2}$ therefore gives

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \left(1 - \frac{1}{2}\alpha\mu\right)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \frac{5}{\alpha\mu}\|\boldsymbol{\theta}^\star_t - \boldsymbol{\theta}^\star_{t+1}\|^2 + \frac{10\alpha\epsilon^{-2}}{\mu}\|\boldsymbol{r}_{t+1}\|^2 + \frac{10\alpha}{\mu}\|\boldsymbol{\eta}_{t+1}\|^2,$$

which is exactly Eq. (19). □

Having reduced the tracking recursion to the two Adam-specific error terms $\boldsymbol{r}_{t+1}$ and $\boldsymbol{\eta}_{t+1}$, it remains to control the first-moment tracking error. The high-probability control of $\boldsymbol{\eta}_{t+1}$ is provided by Lemma D.2. We now turn to $\boldsymbol{r}_{t+1}$, which has two sources: a bias from averaging stale conditional mean gradients and a stochastic noise term obtained by averaging martingale differences.

Lemma D.4 (Bias–noise decomposition of $\boldsymbol{r}_{t+1}$)**.** *For every $t \ge 0$,*

$$\boldsymbol{r}_{t+1} = \boldsymbol{B}^{(1)}_{t+1} + \boldsymbol{N}^{(1)}_{t+1},$$

where

$$\boldsymbol{B}^{(1)}_{t+1} := \sum_{k=0}^{t} w_{1,t+1,k}\left(\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\right), \tag{24}$$

$$\boldsymbol{N}^{(1)}_{t+1} := \sum_{k=0}^{t} w_{1,t+1,k}\boldsymbol{\xi}_{k+1}. \tag{25}$$

Moreover, if $\operatorname{diam}(\Theta) \le R$*, then* $\|\boldsymbol{B}^{(1)}_{t+1}\| \le 2LR\,\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} \le 2LR\beta_1$.

Proof of Lemma D.4. By unrolling the bias-corrected first moment, we have

$$\widehat{\boldsymbol{m}}_{t+1} = \sum_{k=0}^{t} w_{1,t+1,k} \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_k, X_{k+1}) = \sum_{k=0}^{t} w_{1,t+1,k} \left(\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k) + \boldsymbol{\xi}_{k+1}\right).$$

Subtracting $\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)$ and using $\sum_{k=0}^{t} w_{1,t+1,k} = 1$ gives

$$\boldsymbol{r}_{t+1} = \sum_{k=0}^{t} w_{1,t+1,k} \left(\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\right) + \sum_{k=0}^{t} w_{1,t+1,k} \boldsymbol{\xi}_{k+1},$$

which gives the desired decomposition. It remains to bound the bias term. Since the conditional minimizers lie in the interior of Θ, $\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}^\star_{k+1}) = \mathbf{0}$. Thus, by L-Lipschitz continuity and $\operatorname{diam}(\Theta) \le R$, for every $k \ge 0$,

$$\|\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k)\| = \|\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k) - \bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}^\star_{k+1})\| \le L\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star_{k+1}\| \le LR.$$

Consequently, for $0 \le k \le t-1$,

$$\begin{aligned} \|\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k) - \bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\| &\le \|\bar{\boldsymbol{g}}_{k+1}(\boldsymbol{\theta}_k)\| + \|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\| \\ &\le 2LR. \end{aligned} \tag{26}$$

Since the $k = t$ summand in Eq. (24) is zero and the Adam weights are nonnegative, averaging Eq. (26) gives

$$\|\boldsymbol{B}^{(1)}_{t+1}\| \le 2LR \sum_{k=0}^{t-1} w_{1,t+1,k}.$$

For $t \ge 1$, the geometric-series identity yields

$$\sum_{k=0}^{t-1} w_{1,t+1,k} = \frac{1-\beta_1}{1-\beta_1^{t+1}} \sum_{k=0}^{t-1} \beta_1^{t-k} = \frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} \le \beta_1,$$

where the last inequality uses $1-\beta_1^t \le 1-\beta_1^{t+1}$. Therefore,

$$\|\boldsymbol{B}^{(1)}_{t+1}\| \le 2LR \, \frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} \le 2LR\beta_1.$$

When $t = 0$, $\boldsymbol{B}^{(1)}_1 = \mathbf{0}$, and the same bound holds since $1-\beta_1^0 = 0$. This proves the claim. □

The bias component of $\boldsymbol{r}_{t+1}$ is therefore controlled pathwise by the total weight assigned to past conditional mean gradients, which is at most β_1. The stochastic component is a weighted martingale sum. Under conditional sub-Gaussian gradient noise, this weighted sum remains sub-Gaussian with effective variance determined by the sum of squared Adam weights $\kappa_{1,t+1}$.

Lemma D.5 (High-probability control of $\boldsymbol{r}_{t+1}$)**.** *Suppose Assumption B.3 holds and* $\operatorname{diam}(\Theta) \le R$*. Then there exists a universal constant $C > 0$ such that, for every $T \ge 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $t \in \{0, \dots, T-1\}$,*

$$\|\boldsymbol{r}_{t+1}\| \le 2LR \, \frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} + C\sigma \sqrt{\kappa_{1,t+1}\left(d + \log\frac{2T}{\delta}\right)}. \tag{27}$$

Proof of Lemma D.5. Recall from Lemma D.4 that $\boldsymbol{r}_{t+1} = \boldsymbol{B}^{(1)}_{t+1} + \boldsymbol{N}^{(1)}_{t+1}$, where $\boldsymbol{N}^{(1)}_{t+1} = \sum_{k=0}^{t} w_{1,t+1,k} \boldsymbol{\xi}_{k+1}$. The bias component satisfies

$$\|\boldsymbol{B}^{(1)}_{t+1}\| \le 2LR \, \frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}},$$

so it remains to control $\boldsymbol{N}^{(1)}_{t+1}$.

Fix $t \in \{0, \dots, T-1\}$ and $\boldsymbol{u} \in \mathbb{S}^{d-1}$. Define $Z_{k+1} := w_{1,t+1,k}\langle \boldsymbol{u}, \boldsymbol{\xi}_{k+1}\rangle$, $k = 0, \dots, t$. Since $w_{1,t+1,k}$ is deterministic and $\mathbb{E}[\boldsymbol{\xi}_{k+1} \mid \mathcal{F}_k] = \mathbf{0}$, the sequence $(Z_{k+1})_{k=0}^{t}$ is a martingale difference sequence. Moreover, Assumption B.3 gives $\|Z_{k+1}\|_{\Psi_2|\mathcal{F}_k} \le \sigma |w_{1,t+1,k}|$ a.s. Therefore, by Theorem F.1,

$$\mathbb{P}\left(\left|\left\langle \boldsymbol{u}, \boldsymbol{N}_{t+1}^{(1)}\right\rangle\right| \ge s\right) \le 2\exp\left(-\frac{cs^2}{\sigma^2 \sum_{k=0}^{t} w_{1,t+1,k}^2}\right) = 2\exp\left(-\frac{cs^2}{\sigma^2 \kappa_{1,t+1}}\right).$$

Thus, for every $\boldsymbol{u} \in \mathbb{S}^{d-1}$, $\langle \boldsymbol{u}, \boldsymbol{N}_{t+1}^{(1)}\rangle$ is sub-Gaussian with scale at most $C\sigma\sqrt{\kappa_{1,t+1}}$. Let $\mathcal{N}$ be a $1/2$-net of $\mathbb{S}^{d-1}$ satisfying $|\mathcal{N}| \le 5^d$ (Vershynin, 2018, Corollary 4.2.11). By (Vershynin, 2018, Exercise 4.35), $\|\boldsymbol{x}\| \le 2\max_{\boldsymbol{u}\in\mathcal{N}} |\langle \boldsymbol{u}, \boldsymbol{x}\rangle|$ for every $\boldsymbol{x} \in \mathbb{R}^d$. A union bound over $\mathcal{N}$ therefore gives, for every $x > 0$,

$$\mathbb{P}\left(\|\boldsymbol{N}_{t+1}^{(1)}\| \ge C\sigma\sqrt{\kappa_{1,t+1}(d+x)}\right) \le 2e^{-x}.$$

Taking $x = \log(2T/\delta)$ and applying a union bound over $t = 0, \dots, T-1$ yields, with probability at least $1-\delta$, simultaneously for every $t \in \{0, \dots, T-1\}$,

$$\|\boldsymbol{N}_{t+1}^{(1)}\| \le C\sigma\sqrt{\kappa_{1,t+1}\left(d + \log\frac{2T}{\delta}\right)}.$$

Combining this with the bound on $\|\boldsymbol{B}_{t+1}^{(1)}\|$ proves Eq. (27). □

We can now combine the one-step recursion in Lemma D.3 with the high-probability control of $\boldsymbol{r}_{t+1}$ from Lemma D.5 and $\boldsymbol{\eta}_{t+1}$ from Lemma D.2. To simplify the resulting discounted convolutions, define

$$\rho_\alpha := 1 - \frac{1}{2}\alpha\mu, \quad \lambda_i := \rho_\alpha \vee \beta_i, \ i \in \{1, 2\}.$$

Thus, λ_i is the slower of the tracking contraction and the corresponding Adam moment decay.

Theorem D.1 (High probability tracking error bound for Adam)**.** *Suppose Assumptions B.3 and D.1 hold and* $\mathrm{diam}(\Theta) \le R$*. Then, for all* $t \in [T]$*,* $\delta \in (0,1)$*, and* $\alpha \le \min\{\mu\epsilon^2/(4L^2), \mu^{-1}\}$*, the following tracking error bound holds for (Adam) with probability at least* $1-\delta$*:*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \le \rho_\alpha^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{5\mathfrak{D}_t}{\alpha\mu} + \frac{40}{\mu^2\epsilon^2}\left\{4\beta_1^2 L^2 R^2 + \frac{C^2\sigma^2}{1+\beta_1}\left(d + \log\frac{8T}{\delta}\right)\left[1 - \beta_1 + \alpha\mu\beta_1 t\lambda_1^{t-1}\right]\right\}$$
$$+ \frac{10C\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{8T}{\delta}\right)\right]^2\left[\beta_2 t\lambda_2^{t-1} + \frac{2(1-\beta_2)}{\alpha\mu}\right],$$

where $\rho_\alpha := 1 - \alpha\mu/2$*,* $\mathfrak{D}_t := \sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\Delta_\ell^2$*,* $\Delta_\ell := \|\boldsymbol{\theta}_\ell^\star - \boldsymbol{\theta}_{\ell+1}^\star\|$*, and* $\lambda_i := \rho_\alpha \vee \beta_i$ *for* $i \in \{1, 2\}$*.*

Proof of Theorem D.1. Fix $s \in [T]$. Iterating Lemma D.3 from 0 to $s-1$ gives

$$\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^\star\|^2 \le \rho_\alpha^s\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{5\mathfrak{D}_s}{\alpha\mu} + \frac{10\alpha q_+^2}{\mu}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\|\boldsymbol{r}_{\ell+1}\|^2 + \frac{10\alpha}{\mu}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\|\boldsymbol{\eta}_{\ell+1}\|^2. \tag{28}$$

Apply Lemmas D.2 and D.5 with failure probability $\delta/2$ each. By a union bound, with probability at least $1-\delta$, simultaneously for every $\ell = 0, \dots, T-1$,

$$\|\boldsymbol{r}_{\ell+1}\| \le 2LR\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}} + C\sigma\sqrt{\kappa_{1,\ell+1}\left(d + \log\frac{8T}{\delta}\right)},$$
$$\|\boldsymbol{\eta}_{\ell+1}\|^2 \le C\epsilon^{-4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{8T}{\delta}\right)\right]^2\vartheta_{2,\ell+1}.$$

Hence, by $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|\boldsymbol{r}_{\ell+1}\|^2 \le 8L^2R^2\left(\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}}\right)^2 + 2C^2\sigma^2\kappa_{1,\ell+1}\left(d+\log\frac{8T}{\delta}\right).$$

Since $1-\rho_\alpha = \alpha\mu/2$, $\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1} \le 2/(\alpha\mu)$. Moreover, $\beta_1(1-\beta_1^\ell)/(1-\beta_1^{\ell+1}) \le \beta_1$, so

$$\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\left(\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}}\right)^2 \le \frac{2\beta_1^2}{\alpha\mu}.$$

For the variance term, recall that

$$\kappa_{1,\ell+1} = \frac{1-\beta_1}{1+\beta_1} + \frac{2(1-\beta_1)\beta_1^{\ell+1}}{(1+\beta_1)(1-\beta_1^{\ell+1})} \le \frac{1-\beta_1}{1+\beta_1} + \frac{2\beta_1^{\ell+1}}{1+\beta_1},$$

where we used $1-\beta_1^{\ell+1} \ge 1-\beta_1$. Let $\lambda_1 := \rho_\alpha \vee \beta_1$. Then $\rho_\alpha^{s-\ell-1}\beta_1^{\ell+1} \le \beta_1\lambda_1^{s-1}$, and therefore

$$\begin{aligned}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\kappa_{1,\ell+1} &\le \frac{1-\beta_1}{1+\beta_1}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1} + \frac{2}{1+\beta_1}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\beta_1^{\ell+1} \\ &\le \frac{2}{1+\beta_1}\left[\frac{1-\beta_1}{\alpha\mu} + \beta_1 s\lambda_1^{s-1}\right].\end{aligned}$$

Consequently,

$$\frac{10\alpha q_+^2}{\mu}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\|\boldsymbol{r}_{\ell+1}\|^2 \le \frac{40q_+^2}{\mu^2}\left\{4\beta_1^2L^2R^2 + \frac{C^2\sigma^2}{1+\beta_1}\left(d+\log\frac{8T}{\delta}\right)\left[1-\beta_1+\alpha\mu\beta_1 s\lambda_1^{s-1}\right]\right\}.$$

Next, since $\vartheta_{2,\ell+1} = (1-\beta_2)+\beta_2^{\ell+1}$, let $\lambda_2 := \rho_\alpha \vee \beta_2$. As above, $\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\beta_2^{\ell+1} \le \beta_2 s\lambda_2^{s-1}$, and hence

$$\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\vartheta_{2,\ell+1} \le \frac{2(1-\beta_2)}{\alpha\mu} + \beta_2 s\lambda_2^{s-1}.$$

It follows that

$$\frac{10\alpha}{\mu}\sum_{\ell=0}^{s-1}\rho_\alpha^{s-\ell-1}\|\boldsymbol{\eta}_{\ell+1}\|^2 \le \frac{10C\alpha}{\mu\epsilon^4}\left[L^2R^2+\sigma^2\left(d+\log\frac{8T}{\delta}\right)\right]^2\left[\frac{2(1-\beta_2)}{\alpha\mu}+\beta_2 s\lambda_2^{s-1}\right].$$

Substituting these estimates into Eq. (28) and using $q_+ = \epsilon^{-1}$ gives

$$\begin{aligned}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^\star\|^2 \le \rho_\alpha^s\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_0^\star\|^2 + \frac{5\mathfrak{D}_s}{\alpha\mu} + \frac{160\beta_1^2L^2R^2}{\mu^2\epsilon^2} + \frac{40C^2\sigma^2}{\mu^2\epsilon^2(1+\beta_1)}\left(d+\log\frac{8T}{\delta}\right)\left[1-\beta_1+\alpha\mu\beta_1 s\lambda_1^{s-1}\right] \\ + \frac{10C\alpha}{\mu\epsilon^4}\left[L^2R^2+\sigma^2\left(d+\log\frac{8T}{\delta}\right)\right]^2\left[\frac{2(1-\beta_2)}{\alpha\mu}+\beta_2 s\lambda_2^{s-1}\right].\end{aligned}$$

Since the concentration event is uniform over $\ell = 0,\dots,T-1$, this bound holds simultaneously for every $s \in [T]$, proving the claim. □

D.2 Proof of Theorem 3.3

Using Theorem D.1, we can obtain the following result which gives us an algorithmic guarantee for (Adam):

Theorem D.2 (Time to reach the tracking floor with high probability for (Adam))**.** *Under the standing assumptions of Theorem D.1, suppose* $0 < \alpha \le \alpha_{\max} := \min\{\mu\epsilon^2/(4L^2), \mu^{-1}\}$, $\Delta_t \le \Delta$ *almost surely for all* $t \ge 0$*, and fix a deterministic horizon bound* $N \ge 1$ *and* $\delta \in (0,1)$*. Define the tracking-floor*

$$\mathcal{E}_{\mathrm{A}}(N,\delta,\alpha) \asymp \frac{\Delta^2}{\mu^2\alpha^2} + \frac{\beta_1^2L^2R^2}{\mu^2\epsilon^2} + \frac{C^2\sigma^2(1-\beta_1)}{\mu^2\epsilon^2(1+\beta_1)}\left(d+\log\frac{N}{\delta}\right) + \frac{C(1-\beta_2)}{\mu^2\epsilon^4}\left[L^2R^2+\sigma^2\left(d+\log\frac{N}{\delta}\right)\right]^2,$$

where $C > 0$ is a fixed sufficiently large universal constant. For $\alpha \in (0, \alpha_{\max}]$, let $\rho_\alpha := 1 - \alpha\mu/2$ and $\lambda_{i,\alpha} := \rho_\alpha \vee \beta_i$ for $i \in \{1, 2\}$. For a target floor $E > 0$, define

$$\mathcal{B}_1(\alpha, E) \asymp \frac{1}{1-\lambda_{1,\alpha}} \log_+\left(\frac{C^2\alpha\sigma^2\left(d + \log\frac{N}{\delta}\right)}{\mu\epsilon^2(1+\beta_1)(1-\lambda_{1,\alpha})E}\right)$$
$$\mathcal{B}_2(\alpha, E) \asymp \frac{1}{1-\lambda_{2,\alpha}} \log_+\left(\frac{C\alpha\left[L^2R^2 + \sigma^2\left(d + \log\frac{N}{\delta}\right)\right]^2}{\mu\epsilon^4(1-\lambda_{2,\alpha})E}\right).$$

Then we have the following:

1. ***(Constant learning rate).*** *If $\alpha_t \equiv \alpha$, with α chosen deterministically, then for any integer $1 \le T \le N$, with probability $\ge 1 - \delta$, for all $t \in [T]$,*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \rho_\alpha^t\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \mathcal{E}_{\mathrm{A}}(N, \delta, \alpha) + \frac{\alpha\sigma^2}{\mu\epsilon^2(1+\beta_1)}\left(d + \log\frac{N}{\delta}\right)\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1}$$
$$+ \frac{\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{N}{\delta}\right)\right]^2\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_2^{\ell+1}.$$

   Let $\alpha_{\mathrm{A}}^\star \in (0, \alpha_{\max}]$ be a deterministic target stepsize and $\mathcal{E}_{\mathrm{A}}^\star := \mathcal{E}_{\mathrm{A}}(N, \delta, \alpha_{\mathrm{A}}^\star)$. Then, running Adam with $\alpha = \alpha_{\mathrm{A}}^\star$ gives $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \mathcal{E}_{\mathrm{A}}^\star$ after at most

$$t_{\mathrm{A}}^\star \lesssim \frac{1}{\mu\alpha_{\mathrm{A}}^\star}\log_+\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \mathcal{B}_1\left(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star\right) + \mathcal{B}_2\left(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star\right)$$

   iterations, provided $t_{\mathrm{A}}^\star \le T$. The bound holds simultaneously for $t_{\mathrm{A}}^\star \le t \le T$ with probability $\ge 1 - \delta$.

2. ***(Step-decay with Adam-state restart).*** *Suppose a deterministic upper bound $D \ge \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2$ is available and $\alpha_{\mathrm{A}}^\star < \alpha_{\max}$. Set $\alpha_0 := \alpha_{\max}$, $\alpha_k := (\alpha_{k-1} + \alpha_{\mathrm{A}}^\star)/2$, and $K := 1 + \lceil\log_2(\alpha_0/\alpha_{\mathrm{A}}^\star)\rceil$. For each $k \ge 0$, let $E_k := \mathcal{E}_{\mathrm{A}}(N, \delta, \alpha_k)$, and set*

$$T_0 \asymp \left\lceil \max\left\{\frac{1}{\mu\alpha_0}\log_+\left(\frac{D}{E_0}\right),\, \mathcal{B}_1(\alpha_0, E_0),\, \mathcal{B}_2(\alpha_0, E_0)\right\}\right\rceil,$$
$$T_k \asymp \left\lceil \max\left\{\frac{1}{\mu\alpha_k},\, \mathcal{B}_1(\alpha_k, E_k),\, \mathcal{B}_2(\alpha_k, E_k)\right\}\right\rceil, \quad k \ge 1.$$

   Suppose $T := \sum_{k=0}^{K-1} T_k \le N$. Running Adam at constant stepsize α_k for T_k steps per epoch with $(\boldsymbol{m}, \boldsymbol{v})$ and the bias-correction restarted at the beginning of every epoch yields $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim \mathcal{E}_{\mathrm{A}}^\star$ with probability $\ge 1 - K\delta$ Moreover,

$$T \lesssim \frac{1}{\mu\alpha_0}\log_+\left(\frac{D}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \frac{1}{\mu\alpha_{\mathrm{A}}^\star} + \sum_{k=0}^{K-1}\mathcal{B}_1(\alpha_k, E_k) + \sum_{k=0}^{K-1}\mathcal{B}_2(\alpha_k, E_k).$$

Proof of Theorem D.2. Fix a deterministic stepsize $0 < \alpha \le \alpha_{\max}$ and an integer $1 \le T \le N$. By Lemma D.4, $\|\boldsymbol{B}_{t+1}^{(1)}\| \le 2LR\beta_1(1-\beta_1^t)/(1-\beta_1^{t+1}) \le 2LR\beta_1$. Applying Lemmas D.2 and D.5 with horizon N and failure probability $\delta/2$ each, and taking a union bound, gives an event of probability at least $1 - \delta$ on which, simultaneously for every $\ell = 0, \dots, N-1$,

$$\|\boldsymbol{r}_{\ell+1}\| \le 2LR\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}} + C\sigma\sqrt{\kappa_{1,\ell+1}\left(d + \log\frac{8T}{\delta}\right)},$$
$$\|\boldsymbol{\eta}_{\ell+1}\|^2 \le C\epsilon^{-4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{8T}{\delta}\right)\right]^2\vartheta_{2,\ell+1}.$$

Combining these bounds with Lemma D.3 and using $q_+ = \epsilon^{-1}$, we obtain, on the same event, for every $t \in [T]$,

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \le \rho_\alpha^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{5}{\alpha\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\Delta_\ell^2 + \frac{10C\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{N}{\delta}\right)\right]^2 \sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\vartheta_{2,\ell+1}$$
$$+ \frac{10\alpha}{\mu\epsilon^2}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\left(2LR\beta_1 + C\sigma\sqrt{\kappa_{1,\ell+1}\left(d + \log\frac{N}{\delta}\right)}\right)^2. \tag{29}$$

Since $1 - \rho_\alpha = \alpha\mu/2$, $\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1} \le 2/(\alpha\mu)$. Hence the drift term in Eq. (29) is at most $10\Delta^2/(\alpha^2\mu^2)$. Moreover, $\vartheta_{2,\ell+1} = (1-\beta_2) + \beta_2^{\ell+1}$, so

$$\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\vartheta_{2,\ell+1} \le \frac{2(1-\beta_2)}{\alpha\mu} + \sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_2^{\ell+1}.$$

Similarly, since

$$\kappa_{1,\ell+1} = \frac{1-\beta_1}{1+\beta_1} + \frac{2(1-\beta_1)\beta_1^{\ell+1}}{(1+\beta_1)(1-\beta_1^{\ell+1})} \le \frac{1-\beta_1}{1+\beta_1} + \frac{2\beta_1^{\ell+1}}{1+\beta_1},$$

we have

$$\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\kappa_{1,\ell+1} \le \frac{2(1-\beta_1)}{\alpha\mu(1+\beta_1)} + \frac{2}{1+\beta_1}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1}.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$ in the last term of Eq. (29) therefore yields

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \rho_\alpha^t\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \mathcal{E}_{\mathrm{A}}(N,\delta,\alpha) + \frac{\alpha\sigma^2}{\mu\epsilon^2(1+\beta_1)}\left(d + \log\frac{N}{\delta}\right)\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1}$$
$$+ \frac{\alpha}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{N}{\delta}\right)\right]^2\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_2^{\ell+1}. \tag{30}$$

This proves the finite-time bound in part (i).

We next control the three transient terms in Eq. (30). Set $\alpha = \alpha_{\mathrm{A}}^\star$. Since $1 - x \le e^{-x}$, $\rho_{\alpha_{\mathrm{A}}^\star}^t \le e^{-\mu\alpha_{\mathrm{A}}^\star t/2}$. Thus the initialization term is at most $\mathcal{E}_{\mathrm{A}}^\star$ whenever $t \ge 2(\mu\alpha_{\mathrm{A}}^\star)^{-1}\log_+(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2/\mathcal{E}_{\mathrm{A}}^\star)$. For $i \in \{1,2\}$, let $\lambda_{i,\alpha} = \rho_\alpha \vee \beta_i$. Then

$$\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_i^{\ell+1} \le t\lambda_{i,\alpha}^t \le \frac{2}{1-\lambda_{i,\alpha}}\exp\left(-\frac{1-\lambda_{i,\alpha}}{2}t\right),$$

where the second inequality follows from $\sup_{x\ge 0} x\lambda^{x/2} = 2/[e\log(1/\lambda)]$ and $\log(1/\lambda) \ge 1 - \lambda$. By the definitions of $\mathcal{B}_1$ and $\mathcal{B}_2$, the first and second-moment transients in Eq. (30) are each at most $\mathcal{E}_{\mathrm{A}}^\star$ for $t \ge \mathcal{B}_1(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star)$ and $t \ge \mathcal{B}_2(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star)$, respectively. Consequently, taking

$$t_{\mathrm{A}}^\star \lesssim \frac{1}{\mu\alpha_{\mathrm{A}}^\star}\log_+\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \mathcal{B}_1(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star) + \mathcal{B}_2(\alpha_{\mathrm{A}}^\star, \mathcal{E}_{\mathrm{A}}^\star)$$

gives $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \lesssim \mathcal{E}_{\mathrm{A}}^\star$ simultaneously for every $t_{\mathrm{A}}^\star \le t \le T$, provided $t_{\mathrm{A}}^\star \le T$. This proves part (i).

For part (ii), let $t_k := \sum_{j=0}^{k-1} T_j$, $\boldsymbol{X}_k := \boldsymbol{\theta}_{t_k}$, and $\boldsymbol{X}_k^\star := \boldsymbol{\theta}_{t_k}^\star$. The epoch boundaries and step-sizes are deterministic. Since $(\boldsymbol{m}, \boldsymbol{v})$ and the bias-correction are restarted at every epoch, the zero-initialized moment expansions used above apply with respect to the shifted filtration $(\mathcal{F}_{t_k+j})_{j\ge 0}$. Applying part (i) on each epoch and taking a union bound over the K epochs gives an event of probability at least $1 - K\delta$ on which

$$\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_{k+1}^\star\|^2 \lesssim e^{-\mu\alpha_k T_k/2}\|\boldsymbol{X}_k - \boldsymbol{X}_k^\star\|^2 + E_k + \frac{\alpha_k\sigma^2}{\mu\epsilon^2(1+\beta_1)}\left(d + \log\frac{N}{\delta}\right)\sum_{\ell=0}^{T_k-1}\rho_{\alpha_k}^{T_k-\ell-1}\beta_1^{\ell+1}$$
$$+ \frac{\alpha_k}{\mu\epsilon^4}\left[L^2R^2 + \sigma^2\left(d + \log\frac{N}{\delta}\right)\right]^2\sum_{\ell=0}^{T_k-1}\rho_{\alpha_k}^{T_k-\ell-1}\beta_2^{\ell+1} \tag{31}$$

for every $k = 0, \dots, K-1$.

By the convolution bound above and the definitions of $\mathcal{B}_1(\alpha_k, E_k)$ and $\mathcal{B}_2(\alpha_k, E_k)$, the choice of T_k makes both moment transients in Eq. (31) of order at most E_k. The choice of T_0 also gives $e^{-\mu\alpha_0 T_0/2} D \lesssim E_0$, whereas for every $k \geq 1$, $T_k \gtrsim (\mu\alpha_k)^{-1}$ gives $e^{-\mu\alpha_k T_k/2} \leq c$ for a fixed $c < 1$. Hence $\|\boldsymbol{X}_1 - \boldsymbol{X}_1^\star\|^2 \lesssim E_0$.

The recursion $\alpha_k = (\alpha_{k-1} + \alpha_{\mathrm{A}}^\star)/2$ implies $\alpha_k \leq \alpha_{k-1}$. Since the only α-dependent term in $\mathcal{E}_{\mathrm{A}}(N, \delta, \alpha)$ is proportional to α^{-2}, we have $E_{k-1} \lesssim E_k$. Inductively, if $\|\boldsymbol{X}_k - \boldsymbol{X}_k^\star\|^2 \lesssim E_{k-1}$, then Eq. (31) gives $\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_{k+1}^\star\|^2 \lesssim E_k$. Therefore, $\|\boldsymbol{X}_k - \boldsymbol{X}_k^\star\|^2 \lesssim E_{k-1}$ for every $k = 1, \dots, K$.

The stepsize recursion has the explicit solution $\alpha_k = \alpha_{\mathrm{A}}^\star + 2^{-k}(\alpha_0 - \alpha_{\mathrm{A}}^\star)$. Thus $\alpha_{\mathrm{A}}^\star \leq \alpha_{K-1} \leq 2\alpha_{\mathrm{A}}^\star$ by the definition of K. Since $\mathcal{E}_{\mathrm{A}}(N, \delta, \alpha)$ is nonincreasing in α, $E_{K-1} \lesssim \mathcal{E}_{\mathrm{A}}^\star$, and consequently

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 = \|\boldsymbol{X}_K - \boldsymbol{X}_K^\star\|^2 \lesssim E_{K-1} \lesssim \mathcal{E}_{\mathrm{A}}^\star$$

with probability at least $1 - K\delta$.

It remains to bound T. The definitions of the epoch lengths give

$$T \lesssim \frac{1}{\mu\alpha_0} \log_+\left(\frac{D}{E_0}\right) + \sum_{k=1}^{K-1} \frac{1}{\mu\alpha_k} + \sum_{k=0}^{K-1} \mathcal{B}_1(\alpha_k, E_k) + \sum_{k=0}^{K-1} \mathcal{B}_2(\alpha_k, E_k).$$

Let $r := \alpha_0/\alpha_{\mathrm{A}}^\star \geq 1$. From the form of the tracking floor, $E_0 \lesssim \mathcal{E}_{\mathrm{A}}^\star \lesssim r^2 E_0$, and hence

$$\frac{1}{\mu\alpha_0} \log_+\left(\frac{D}{E_0}\right) \lesssim \frac{1}{\mu\alpha_0} \log_+\left(\frac{D}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \frac{1}{\mu\alpha_{\mathrm{A}}^\star},$$

where we used $(\log r)/r \leq 1/e$. Finally, $\sum_{k=1}^{K-1} \alpha_k^{-1} \lesssim (\alpha_{\mathrm{A}}^\star)^{-1}$. Indeed, this is immediate when $r \leq 2$. If $r > 2$, writing $n := K - 1 = \lceil \log_2 r \rceil$ and using $\alpha_k \geq 2^{-k}(\alpha_0 - \alpha_{\mathrm{A}}^\star)$ gives

$$\sum_{k=1}^{K-1} \frac{1}{\alpha_k} \leq \frac{2^{n+1} - 2}{\alpha_0 - \alpha_{\mathrm{A}}^\star} \leq \frac{8}{\alpha_{\mathrm{A}}^\star}.$$

Substituting these estimates into Section D.2 yields

$$T \lesssim \frac{1}{\mu\alpha_0} \log_+\left(\frac{D}{\mathcal{E}_{\mathrm{A}}^\star}\right) + \frac{1}{\mu\alpha_{\mathrm{A}}^\star} + \sum_{k=0}^{K-1} \mathcal{B}_1(\alpha_k, E_k) + \sum_{k=0}^{K-1} \mathcal{B}_2(\alpha_k, E_k),$$

which is the claimed bound. □

D.3 Proofs for expected tracking error

The high-probability analysis above controls the stochastic first-moment and preconditioner errors uniformly over time. For completeness, we also give their corresponding second-moment bounds. In expectation, the analysis is simpler: orthogonality of the martingale differences controls the first-moment noise, while conditional sub-Gaussianity provides the required fourth-moment control for the preconditioner perturbation.

Lemma D.6 (Second-moment control of the first-moment tracking error)**.** *Under the standing setup, suppose Assumption B.3 holds and* $\mathrm{diam}(\Theta) \leq R$*. Then, for every* $t \geq 0$*,*

$$\mathbb{E}\|\boldsymbol{r}_{t+1}\|^2 \leq 8L^2R^2 \left(\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}\right)^2 + 2Cd\sigma^2\kappa_{1,t+1}, \tag{32}$$

where $C > 0$ *is a universal constant.*

Proof of Lemma D.6. By Lemma D.4,

$$\boldsymbol{r}_{t+1} = \boldsymbol{B}_{t+1}^{(1)} + \boldsymbol{N}_{t+1}^{(1)}, \quad \|\boldsymbol{B}_{t+1}^{(1)}\| \leq 2LR\,\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}.$$

Thus, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbb{E}\|\boldsymbol{r}_{t+1}\|^2 \le 8L^2R^2\left(\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}\right)^2 + 2\mathbb{E}\|\boldsymbol{N}_{t+1}^{(1)}\|^2.$$

Since $\boldsymbol{N}_{t+1}^{(1)} = \sum_{k=0}^{t} w_{1,t+1,k}\boldsymbol{\xi}_{k+1}$, the martingale-difference property gives, for $k < \ell$, $\mathbb{E}\langle \boldsymbol{\xi}_{k+1}, \boldsymbol{\xi}_{\ell+1}\rangle = \mathbb{E}\langle \boldsymbol{\xi}_{k+1}, \mathbb{E}[\boldsymbol{\xi}_{\ell+1} \mid \mathcal{F}_\ell]\rangle = 0$ where $\boldsymbol{\xi}_{k+1}$ is $\mathcal{F}_\ell$-measurable and the noise vectors are square-integrable by conditional sub-Gaussianity. Hence

$$\mathbb{E}\|\boldsymbol{N}_{t+1}^{(1)}\|^2 = \sum_{k=0}^{t} w_{1,t+1,k}^2 \mathbb{E}\|\boldsymbol{\xi}_{k+1}\|^2.$$

Assumption B.3 implies $\mathbb{E}[\|\boldsymbol{\xi}_{k+1}\|^2 \mid \mathcal{F}_k] \le Cd\sigma^2$ a.s., and therefore

$$\mathbb{E}\|\boldsymbol{N}_{t+1}^{(1)}\|^2 \le Cd\sigma^2 \sum_{k=0}^{t} w_{1,t+1,k}^2 = Cd\sigma^2\kappa_{1,t+1}.$$

Substituting this bound above proves Eq. (32). □

The next lemma gives the corresponding second-moment control of the preconditioner perturbation. Without assuming bounded sample gradients, conditional sub-Gaussianity provides sufficient fourth-moment control.

Lemma D.7 (Second-moment control of the preconditioner perturbation)**.** *Under the standing setup, suppose Assumption B.3 holds and* $\operatorname{diam}(\Theta) \le R$*. Then, for every* $t \ge 0$*,*

$$\mathbb{E}\|\boldsymbol{\eta}_{t+1}\|^2 \le C\epsilon^{-4}\left(L^2R^2 + d\sigma^2\right)^2 \vartheta_{2,t+1}, \tag{33}$$

where $C > 0$ *is a universal constant.*

Proof of Lemma D.7. By the deterministic coordinatewise comparison in the proof of Lemma D.2,

$$\|\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1}\|_{\mathrm{op}}^2 \le \epsilon^{-4}\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty.$$

Since $\boldsymbol{\eta}_{t+1} = (\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1})\widehat{\boldsymbol{m}}_{t+1}$, Cauchy–Schwarz therefore gives

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\eta}_{t+1}\|^2 &\le \epsilon^{-4}\mathbb{E}\left[\|\widehat{\boldsymbol{m}}_{t+1}\|^2\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty\right] \\ &\le \epsilon^{-4}\left(\mathbb{E}\|\widehat{\boldsymbol{m}}_{t+1}\|^4\right)^{1/2}\left(\mathbb{E}\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty^2\right)^{1/2}.\end{aligned} \tag{34}$$

Since $\|\bar{\boldsymbol{g}}_{t+1}(\boldsymbol{\theta}_t)\| \le LR$ and $\|\boldsymbol{\xi}_{t+1}\|_{\Psi_2|\mathcal{F}_t} \le \sigma$, standard sub-Gaussian moment bounds imply $\mathbb{E}[\xi_{t+1,i}^4 \mid \mathcal{F}_t] \le C\sigma^4$ for every $i \in [d]$. Using $\|\boldsymbol{\xi}_{t+1}\|^4 \le d\sum_{i=1}^d \xi_{t+1,i}^4$ and $(a+b)^4 \le 8a^4 + 8b^4$, we obtain

$$\mathbb{E}\|\boldsymbol{\xi}_{t+1}\|^4 \le Cd^2\sigma^4, \;\; \mathbb{E}\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\|^4 \le C\left(L^2R^2 + d\sigma^2\right)^2.$$

Because $\widehat{\boldsymbol{m}}_{t+1}$ is a convex combination of the past stochastic gradients, convexity gives $\mathbb{E}\|\widehat{\boldsymbol{m}}_{t+1}\|^4 \le C(L^2R^2 + d\sigma^2)^2$. It remains to control the second-moment discrepancy. Recall that

$$\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1} = \frac{\beta_2^{t+1}}{1-\beta_2^{t+1}}\boldsymbol{v}_{t+1} + (1-\beta_2)\boldsymbol{\chi}_{t+1}.$$

Since $\boldsymbol{v}_{t+1} = (1-\beta_2)\sum_{j=0}^{t}\beta_2^{t-j}(\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_j, X_{j+1}))^{\odot 2}$, the triangle inequality in L^2 gives

$$\begin{aligned}\left(\mathbb{E}\|\boldsymbol{v}_{t+1}\|_\infty^2\right)^{1/2} &\le (1-\beta_2)\sum_{j=0}^{t}\beta_2^{t-j}\left(\mathbb{E}\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_j, X_{j+1})\|^4\right)^{1/2} \\ &\le C(1-\beta_2^{t+1})\left(L^2R^2 + d\sigma^2\right).\end{aligned}$$

Moreover,

$$\|\boldsymbol{\chi}_{t+1}\|_\infty \le \|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\|^2 + \mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1})\|^2 \,\middle|\, \mathcal{F}_t\right],$$

and conditional Jensen's inequality together with the fourth-moment bound above yields

$$\left(\mathbb{E}\|\boldsymbol{\chi}_{t+1}\|_\infty^2\right)^{1/2} \le C\left(L^2R^2 + d\sigma^2\right).$$

Hence, by the triangle inequality in L^2,

$$\begin{aligned}\left(\mathbb{E}\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty^2\right)^{1/2} &\le C\left(L^2R^2 + d\sigma^2\right)\left[\beta_2^{t+1} + (1-\beta_2)\right] \\ &= C\left(L^2R^2 + d\sigma^2\right)\vartheta_{2,t+1}. \end{aligned} \tag{35}$$

Substituting this estimate and the fourth-moment bound for $\widehat{\boldsymbol{m}}_{t+1}$ into Eq. (34) proves Eq. (33). □

Theorem D.3 (Expected tracking bound for Adam)**.** *Suppose Assumptions B.3 and D.1 hold under the standing setup,* $\operatorname{diam}(\Theta) \le R$*, and the deterministic stepsize satisfies* $0 < \alpha \le \min\left\{\frac{\mu\epsilon^2}{4L^2}, \frac{1}{\mu}\right\}$*. If* $\mathbb{E}[\Delta_t^2] \le \Delta^2$ *for all* $t \ge 0$*, then, for every* $t \ge 1$*,*

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \le \rho_\alpha^t \mathbb{E}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{10\Delta^2}{\alpha^2\mu^2} + \frac{160\beta_1^2L^2R^2}{\mu^2\epsilon^2} + \frac{40Cd\sigma^2}{\mu^2\epsilon^2(1+\beta_1)}\left[1 - \beta_1 + \alpha\mu\beta_1 t\lambda_1^{t-1}\right] \\ + \frac{10C\alpha}{\mu\epsilon^4}\left(L^2R^2 + d\sigma^2\right)^2\left[\beta_2 t\lambda_2^{t-1} + \frac{2(1-\beta_2)}{\alpha\mu}\right],\end{aligned} \tag{36}$$

where $\rho_\alpha := 1 - \alpha\mu/2$*,* $\lambda_i := \rho_\alpha \vee \beta_i$ *for* $i \in \{1,2\}$*, and* $C > 0$ *is a sufficiently large universal constant.*

Proof of Theorem D.3. Iterating Lemma D.3 and taking expectations gives

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 \le \rho_\alpha^t \mathbb{E}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{5}{\alpha\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\mathbb{E}[\Delta_\ell^2] \\ + \frac{10\alpha q_+^2}{\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\mathbb{E}\|\boldsymbol{r}_{\ell+1}\|^2 + \frac{10\alpha}{\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\mathbb{E}\|\boldsymbol{\eta}_{\ell+1}\|^2.\end{aligned} \tag{37}$$

Since $1 - \rho_\alpha = \alpha\mu/2$, $\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1} \le 2/(\alpha\mu)$. Thus, using $\mathbb{E}[\Delta_\ell^2] \le \Delta^2$, the drift term is bounded by $10\Delta^2/(\alpha^2\mu^2)$.

By Eq. (32),

$$\mathbb{E}\|\boldsymbol{r}_{\ell+1}\|^2 \le 8L^2R^2\left(\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}}\right)^2 + 2Cd\sigma^2\kappa_{1,\ell+1}.$$

Since $\beta_1(1-\beta_1^\ell)/(1-\beta_1^{\ell+1}) \le \beta_1$,

$$\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\left(\frac{\beta_1(1-\beta_1^\ell)}{1-\beta_1^{\ell+1}}\right)^2 \le \frac{2\beta_1^2}{\alpha\mu}.$$

Moreover,

$$\kappa_{1,\ell+1} = \frac{1-\beta_1}{1+\beta_1} + \frac{2(1-\beta_1)\beta_1^{\ell+1}}{(1+\beta_1)(1-\beta_1^{\ell+1})} \le \frac{1-\beta_1}{1+\beta_1} + \frac{2\beta_1^{\ell+1}}{1+\beta_1},$$

where $1 - \beta_1^{\ell+1} \ge 1 - \beta_1$. Let $\lambda_1 := \rho_\alpha \vee \beta_1$. Since $\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1} \le \beta_1\lambda_1^{t-1}$,

$$\begin{aligned}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\kappa_{1,\ell+1} &\le \frac{1-\beta_1}{1+\beta_1}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1} + \frac{2}{1+\beta_1}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_1^{\ell+1} \\ &\le \frac{2}{1+\beta_1}\left[\frac{1-\beta_1}{\alpha\mu} + \beta_1 t\lambda_1^{t-1}\right].\end{aligned}$$

Recalling $q_+ = \epsilon^{-1}$, it follows that

$$\frac{10\alpha q_+^2}{\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\mathbb{E}\|\boldsymbol{r}_{\ell+1}\|^2 \le \frac{160\beta_1^2L^2R^2}{\mu^2\epsilon^2} + \frac{40Cd\sigma^2}{\mu^2\epsilon^2(1+\beta_1)}\left[1 - \beta_1 + \alpha\mu\beta_1 t\lambda_1^{t-1}\right].$$

Similarly, Lemma D.7 gives $\mathbb{E}\|\boldsymbol{\eta}_{\ell+1}\|^2 \le C\epsilon^{-4}(L^2R^2+d\sigma^2)^2\vartheta_{2,\ell+1}$. Since $\vartheta_{2,\ell+1} = (1-\beta_2)+\beta_2^{\ell+1}$, let $\lambda_2 := \rho_\alpha \vee \beta_2$. Then $\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\beta_2^{\ell+1} \le \beta_2 t\lambda_2^{t-1}$, and hence

$$\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\vartheta_{2,\ell+1} \le \frac{2(1-\beta_2)}{\alpha\mu} + \beta_2 t\lambda_2^{t-1}.$$

Therefore,

$$\frac{10\alpha}{\mu}\sum_{\ell=0}^{t-1}\rho_\alpha^{t-\ell-1}\mathbb{E}\|\boldsymbol{\eta}_{\ell+1}\|^2 \le \frac{10C\alpha}{\mu\epsilon^4}\left(L^2R^2+d\sigma^2\right)^2\left[\frac{2(1-\beta_2)}{\alpha\mu}+\beta_2 t\lambda_2^{t-1}\right].$$

Substituting these estimates into Eq. (37) proves Eq. (36). □

E Proofs for metric-projected stationarity guarantees under general Adam preconditioning

For this section, we will make the following standard assumption in nonconvex optimization:

Assumption E.1 (Uniform lower boundedness)**.** *There exists a constant $G^\star > -\infty$ such that for all $t \in [T]$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$, $G_t(\boldsymbol{\theta}) \ge G^\star$.*

For this section, $\mathcal{P}_\Theta$ denotes projection onto Θ in the $\widetilde{\boldsymbol{P}}_{t+1}^{-1}$-metric. For the projected Adam recursion, we replace the final line of (Adam) by

$$\boldsymbol{\theta}_{t+1} := \mathcal{P}_\Theta\left(\boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\nabla_{\boldsymbol{\theta}}G_{t+1}(\boldsymbol{\theta}_t) - \alpha\left(\widetilde{\boldsymbol{P}}_{t+1}\boldsymbol{r}_{t+1}+\boldsymbol{\eta}_{t+1}\right)\right).$$

We also define the noiseless projected point by $\bar{\boldsymbol{\theta}}_{t+1} := \mathcal{P}_\Theta(\boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\nabla_{\boldsymbol{\theta}}G_{t+1}(\boldsymbol{\theta}_t))$, and the preconditioned projected-gradient mapping by

$$\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t) := \frac{1}{\alpha}\left(\boldsymbol{\theta}_t - \mathcal{P}_\Theta(\boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{P}}_{t+1}\nabla_{\boldsymbol{\theta}}G_{t+1}(\boldsymbol{\theta}_t))\right).$$

The projected-gradient mapping is the constrained analogue of the gradient: $\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = \mathbf{0}$ if and only if $\boldsymbol{\theta}$ satisfies the first-order constrained stationarity condition $-\nabla f(\boldsymbol{\theta}) \in N_\Theta(\boldsymbol{\theta})$ (see Lemma F.2). Thus, small $\|\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta})\|$ measures approximate stationarity for the constrained problem, while in the unconstrained case $\Theta = \mathbb{R}^d$, it reduces to the usual preconditioned gradient $\boldsymbol{P}\nabla f(\boldsymbol{\theta})$.

Before we proceed, we will need to prove a few technical results that we will find useful. The projected stationarity analysis requires slightly different control of the preconditioner perturbation than the tracking analysis above. In particular, the descent argument naturally involves the weighted quantity $\|\widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}$. Controlling this quantity directly avoids introducing a deterministic lower spectral bound on $\widetilde{\boldsymbol{P}}_{t+1}$.

Lemma E.1 (Weighted control of the preconditioner perturbation)**.** *Under the standing setup, suppose Assumption B.3 holds and $\operatorname{diam}(\Theta) \le R$. Then there exists a universal constant $C > 0$ such that, for every $T \ge 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $t \in \{0,\dots,T-1\}$,*

$$\left\|\widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}} \le C\epsilon^{-3}\left[L^2R^2+\sigma^2\left(d+\log\frac{4T}{\delta}\right)\right]^2\vartheta_{2,t+1}. \tag{38}$$

Moreover, for every $t \ge 0$,

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}\right] \le C\epsilon^{-3}\left(L^2R^2+d\sigma^2\right)^2\vartheta_{2,t+1}. \tag{39}$$

Proof of Lemma E.1. For each coordinate, write $P_{t+1,i} = (\sqrt{\widehat{v}_{t+1,i}}+\epsilon)^{-1}$ and $\widetilde{P}_{t+1,i} = (\sqrt{\widetilde{v}_{t+1,i}}+\epsilon)^{-1}$. For any $a, b \ge 0$,

$$\frac{\left[(\sqrt{a}+\epsilon)^{-1}-(\sqrt{b}+\epsilon)^{-1}\right]^2}{(\sqrt{b}+\epsilon)^{-1}} = \frac{(\sqrt{a}-\sqrt{b})^2}{(\sqrt{a}+\epsilon)^2(\sqrt{b}+\epsilon)} \le \epsilon^{-3}|a-b|.$$

Since $\boldsymbol{\eta}_{t+1} = (\boldsymbol{P}_{t+1} - \widetilde{\boldsymbol{P}}_{t+1})\widehat{\boldsymbol{m}}_{t+1}$, applying this inequality coordinatewise gives

$$\left\|\widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|_{\widetilde{\boldsymbol{P}}_{t+1}}^2 \le \epsilon^{-3}\|\widehat{\boldsymbol{m}}_{t+1}\|^2\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty. \tag{40}$$

On the event of Lemma D.1, which has probability at least $1-\delta$, the estimates in the proof of Lemma D.2 give, simultaneously for every $t \in \{0,\dots,T-1\}$,

$$\|\widehat{\boldsymbol{m}}_{t+1}\|^2 \le C\left[L^2R^2 + \sigma^2\left(d + \log\frac{4T}{\delta}\right)\right],$$
$$\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty \le C\left[L^2R^2 + \sigma^2\left(d + \log\frac{4T}{\delta}\right)\right]\vartheta_{2,t+1}.$$

Both estimates hold on the same gradient-concentration event. Substituting them into Eq. (40) proves Eq. (38).

For the expectation bound, Cauchy–Schwarz applied to Eq. (40) gives

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\boldsymbol{m}}_{t+1}\|^2\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty\right] &\le \left(\mathbb{E}\|\widehat{\boldsymbol{m}}_{t+1}\|^4\right)^{1/2}\left(\mathbb{E}\|\widehat{\boldsymbol{v}}_{t+1} - \widetilde{\boldsymbol{v}}_{t+1}\|_\infty^2\right)^{1/2}\\ &\le C\left(L^2R^2 + d\sigma^2\right)^2\vartheta_{2,t+1},\end{aligned}$$

where the last inequality follows from the sub-Gaussian moment estimates used in Lemma D.7. Combining this with Eq. (40) proves Eq. (39). □

With this weighted perturbation estimate in hand, we can now analyze stationarity directly in the geometry induced by the predictable Adam preconditioner. The next lemma bounds the average $\widetilde{\boldsymbol{P}}_{t+1}^{-1}$-metric projected-gradient mapping.

Lemma E.2 (High-probability metric-projected gradient bound under general Adam preconditioning)**.** *Under the standing setup and Assumptions B.2, E.1, and B.3, suppose* $\operatorname{diam}(\Theta) \le R$*. Then, for all integers* $T \ge 1$ *and all* $\delta \in (0,1)$*, if* $0 < \alpha \le \epsilon/(4L)$*, the iterates generated by the metric-projected Adam update satisfy, with probability at least* $1-\delta$*,*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|_{\widetilde{\boldsymbol{P}}_{t+1}}^2 \le \frac{8\left(G_1(\boldsymbol{\theta}_0) - G^\star + \mathfrak{D}_T\right)}{\alpha T} &+ \frac{12C}{\epsilon^3 T}\left[L^2R^2 + \sigma^2\left(d + \log\frac{8T}{\delta}\right)\right]^2\sum_{t=0}^{T-1}\vartheta_{2,t+1}\\ &+ \frac{12}{\epsilon T}\sum_{t=0}^{T-1}\left(2LR\,\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} + C\sigma\sqrt{\kappa_{1,t+1}\left(d + \log\frac{8T}{\delta}\right)}\right)^2.\end{aligned} \tag{41}$$

Here $\mathfrak{D}_T := \sum_{t=0}^{T-2}\left(G_{t+2}(\boldsymbol{\theta}_{t+1}) - G_{t+1}(\boldsymbol{\theta}_{t+1})\right)_+$ *is the pathwise objective-variation budget, with the empty sum interpreted as zero when* $T = 1$*. The matrix* $\widetilde{\boldsymbol{P}}_{t+1}$ *is the predictable Adam preconditioner,* $\kappa_{1,t}$ *and* $\vartheta_{2,t}$ *are defined in (3), and* $C > 0$ *is a sufficiently large universal constant.*

Proof of Lemma E.2. Let $\boldsymbol{s}_{t+1} := \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t$. By the variational inequality for the $\widetilde{\boldsymbol{P}}_{t+1}^{-1}$-metric projection (Lemma F.1), taking $\boldsymbol{y} = \boldsymbol{\theta}_t$ and applying Cauchy–Schwarz and Young's inequality in the $\widetilde{\boldsymbol{P}}_{t+1}/\widetilde{\boldsymbol{P}}_{t+1}^{-1}$ dual norm pair gives

$$\langle\nabla_{\boldsymbol{\theta}}G_{t+1}(\boldsymbol{\theta}_t), \boldsymbol{s}_{t+1}\rangle \le -\frac{1}{2\alpha}\|\boldsymbol{s}_{t+1}\|_{\widetilde{\boldsymbol{P}}_{t+1}^{-1}}^2 + \frac{\alpha}{2}\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|_{\widetilde{\boldsymbol{P}}_{t+1}}^2. \tag{42}$$

Since G_{t+1} is L-smooth, Lemma F.3 implies

$$G_{t+1}(\boldsymbol{\theta}_{t+1}) \le G_{t+1}(\boldsymbol{\theta}_t) - \frac{1}{2\alpha}\|\boldsymbol{s}_{t+1}\|_{\widetilde{\boldsymbol{P}}_{t+1}^{-1}}^2 + \frac{\alpha}{2}\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|_{\widetilde{\boldsymbol{P}}_{t+1}}^2 + \frac{L}{2}\|\boldsymbol{s}_{t+1}\|^2. \tag{43}$$

Because $\widetilde{\boldsymbol{P}}_{t+1} \preceq q_+\boldsymbol{I}$, where $q_+ = \epsilon^{-1}$, $\|\boldsymbol{s}_{t+1}\|^2 \le q_+\|\boldsymbol{s}_{t+1}\|_{\widetilde{\boldsymbol{P}}_{t+1}^{-1}}^2$. The stepsize condition $\alpha \le (4Lq_+)^{-1} = \epsilon/(4L)$ therefore gives

$$G_{t+1}(\boldsymbol{\theta}_{t+1}) \le G_{t+1}(\boldsymbol{\theta}_t) - \frac{1}{4\alpha}\|\boldsymbol{s}_{t+1}\|_{\widetilde{\boldsymbol{P}}_{t+1}^{-1}}^2 + \frac{\alpha}{2}\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}_{t+1}^{-1}\boldsymbol{\eta}_{t+1}\right\|_{\widetilde{\boldsymbol{P}}_{t+1}}^2. \tag{44}$$

Equivalently,

$$\|\boldsymbol{s}_{t+1}\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le 4\alpha\left(G_{t+1}(\boldsymbol{\theta}_t) - G_{t+1}(\boldsymbol{\theta}_{t+1})\right) + 2\alpha^2\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}. \tag{45}$$

We next compare the realized projected step with the noiseless projected-gradient step. By nonexpansiveness of the $\widetilde{\boldsymbol{P}}^{-1}_{t+1}$-metric projection,

$$\|\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}\|_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le \alpha\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|_{\widetilde{\boldsymbol{P}}_{t+1}}.$$

Hence,

$$\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le \frac{2}{\alpha^2}\|\boldsymbol{s}_{t+1}\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} + 2\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}. \tag{46}$$

Combining Eqs. (45) and (46) gives

$$\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le \frac{8}{\alpha}\left(G_{t+1}(\boldsymbol{\theta}_t) - G_{t+1}(\boldsymbol{\theta}_{t+1})\right) + 6\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}. \tag{47}$$

We now sum the one-step descent inequality and account for the temporal variation of the objective. By telescoping,

$$\begin{aligned}\sum_{t=0}^{T-1}\left(G_{t+1}(\boldsymbol{\theta}_t) - G_{t+1}(\boldsymbol{\theta}_{t+1})\right) &= G_1(\boldsymbol{\theta}_0) - G_T(\boldsymbol{\theta}_T) + \sum_{t=0}^{T-2}\left(G_{t+2}(\boldsymbol{\theta}_{t+1}) - G_{t+1}(\boldsymbol{\theta}_{t+1})\right)\\ &\le G_1(\boldsymbol{\theta}_0) - G^\star + \mathfrak{D}_T,\end{aligned}$$

where the last inequality uses $G_T(\boldsymbol{\theta}_T) \ge G^\star$ and the definition of $\mathfrak{D}_T$. Summing Eq. (47) over $t = 0, \dots, T-1$ and dividing by T therefore gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le \frac{8\big(G_1(\boldsymbol{\theta}_0) - G^\star + \mathfrak{D}_T\big)}{\alpha T} + \frac{6}{T}\sum_{t=0}^{T-1}\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}. \tag{48}$$

Finally,

$$\left\|\boldsymbol{r}_{t+1} + \widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}} \le 2q_+\|\boldsymbol{r}_{t+1}\|^2 + 2\left\|\widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}.$$

Applying Lemmas D.5 and E.1 with failure probability $\delta/2$ each and taking a union bound therefore gives, with probability at least $1-\delta$, simultaneously for all $t = 0, \dots, T-1$,

$$\|\boldsymbol{r}_{t+1}\| \le 2LR\,\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}} + C\sigma\sqrt{\kappa_{1,t+1}\left(d + \log\frac{8T}{\delta}\right)}$$

and

$$\left\|\widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}} \le C\epsilon^{-3}\left[L^2R^2 + \sigma^2\left(d + \log\frac{8T}{\delta}\right)\right]^2\vartheta_{2,t+1}.$$

Substituting these two estimates into Eq. (48) and recalling $q_+ = \epsilon^{-1}$ proves Eq. (41). □

The average bound immediately gives a corresponding randomized-iterate statement. This is useful when the optimization algorithm returns an iterate sampled uniformly from its trajectory rather than the entire sequence.

Remark E.1 (Returned-iterate interpretation). Let $\tau \sim \mathrm{Unif}\{0, \dots, T-1\}$ be sampled independently of $\mathcal{F}_T$. Then, on the same high-probability event as Lemma E.2,

$$\mathbb{E}_\tau\left[\|\mathcal{G}_{\alpha,\tau}(\boldsymbol{\theta}_\tau)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{\tau+1}}\right] = \frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}}.$$

Here $\mathbb{E}_\tau$ averages only over the independent index τ, with $\mathcal{F}_T$ fixed. Thus Lemma E.2 also provides a returned-iterate guarantee in the predictable preconditioner geometry. Moreover, if the right-hand side of Eq. (41) is denoted by $B_T(\delta)$, then for every $\eta \in (0,1)$, Markov's inequality gives, on the same event,

$$\mathbb{P}_\tau\left(\|\mathcal{G}_{\alpha,\tau}(\boldsymbol{\theta}_\tau)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{\tau+1}} > \frac{B_T(\delta)}{\eta}\right) \le \eta,$$

where $\mathbb{P}_\tau$ also refers only to the independent choice of τ, with $\mathcal{F}_T$ fixed. Consequently, the corresponding unconditional failure probability is at most $\delta + \eta$.

E.1 Proof of Theorem 4.1

The bound in Lemma E.2 retains the exact bias-correction factors at each iteration. We now average these quantities explicitly to separate the finite-horizon contributions from the persistent terms in the metric-residual upper bound.

Theorem E.2 (Explicit high-probability projected-gradient rate under projected Adam)**.** *Under the conditions of Lemma E.2, with probability at least* $1-\delta$*,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \lesssim \frac{G_1(\boldsymbol{\theta}_0)-G^\star+\mathfrak{D}_T}{\alpha T}+\frac{\mathsf{Decay}_T(\delta)}{T}+\mathsf{Floor}_T(\delta), \tag{49}$$

where

$$\mathsf{Decay}_T(\delta)\sim\mathcal{O}\left\{\frac{\beta_2}{\epsilon^3(1-\beta_2)}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2+\frac{\sigma^2\left(d+\log\frac{T}{\delta}\right)(1+\log T)}{\epsilon(1+\beta_1)}\right\},$$

$$\mathsf{Floor}_T(\delta)\sim\mathcal{O}\left\{\frac{L^2R^2\beta_1^2}{\epsilon}+\frac{\sigma^2(1-\beta_1)}{\epsilon(1+\beta_1)}\left(d+\log\frac{T}{\delta}\right)+\frac{1-\beta_2}{\epsilon^3}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2\right\}.$$

Proof of Theorem E.2. By Lemma E.2 and $(a+b)^2\le 2a^2+2b^2$, the first-moment term is bounded by

$$\frac{96L^2R^2}{\epsilon T}\sum_{t=1}^{T}\left(\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}\right)^2+\frac{24C^2\sigma^2}{\epsilon T}\left(d+\log\frac{8T}{\delta}\right)\sum_{t=1}^{T}\kappa_{1,t}.$$

Since $1-\beta_1^{t-1}\le 1-\beta_1^t$, $\beta_1(1-\beta_1^{t-1})/(1-\beta_1^t)\le\beta_1$, and hence

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}\right)^2\le\beta_1^2.$$

We next bound the average of $\kappa_{1,t}$. For $0<\beta<1$,

$$1-\beta^t=(1-\beta)\sum_{j=0}^{t-1}\beta^j\ge t(1-\beta)\beta^{t-1},$$

so $\beta^t/(1-\beta^t)\le\beta/[t(1-\beta)]\le 1/[t(1-\beta)]$; the same bound is trivial for $\beta=0$. Since

$$\kappa_{1,t}=\frac{1-\beta_1}{1+\beta_1}\left(1+\frac{2\beta_1^t}{1-\beta_1^t}\right),$$

we obtain $\kappa_{1,t}\le(1-\beta_1)/(1+\beta_1)+2/[(1+\beta_1)t]$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\kappa_{1,t}\le\frac{1-\beta_1}{1+\beta_1}+\frac{2(1+\log T)}{(1+\beta_1)T},$$

where $\sum_{t=1}^{T}t^{-1}\le 1+\log T$. Similarly, using $\vartheta_{2,t}=(1-\beta_2)+\beta_2^t$ and the geometric-series identity,

$$\frac{1}{T}\sum_{t=1}^{T}\vartheta_{2,t}=(1-\beta_2)+\frac{\beta_2(1-\beta_2^T)}{T(1-\beta_2)}\le(1-\beta_2)+\frac{\beta_2}{(1-\beta_2)T}.$$

Substituting these estimates into Lemma E.2 yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}}\lesssim\frac{G_1(\boldsymbol{\theta}_0)-G^\star+\mathfrak{D}_T}{\alpha T}+\frac{L^2R^2\beta_1^2}{\epsilon}+\frac{\sigma^2}{\epsilon}\left(d+\log\frac{8T}{\delta}\right)\left[\frac{1-\beta_1}{1+\beta_1}+\frac{1+\log T}{(1+\beta_1)T}\right]\\+\frac{1}{\epsilon^3}\left[L^2R^2+\sigma^2\left(d+\log\frac{8T}{\delta}\right)\right]^2\left[(1-\beta_2)+\frac{\beta_2}{(1-\beta_2)T}\right].\end{aligned}$$

Thus we find that

$$\mathsf{Decay}_T(\delta) := \mathcal{O}\left\{\frac{\beta_2}{\epsilon^3(1-\beta_2)}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2+\frac{\sigma^2\left(d+\log\frac{T}{\delta}\right)(1+\log T)}{\epsilon(1+\beta_1)}\right\},$$

$$\mathsf{Floor}_T(\delta) := \mathcal{O}\left\{\frac{L^2R^2\beta_1^2}{\epsilon}+\frac{\sigma^2(1-\beta_1)}{\epsilon(1+\beta_1)}\left(d+\log\frac{T}{\delta}\right)+\frac{1-\beta_2}{\epsilon^3}\left[L^2R^2+\sigma^2\left(d+\log\frac{T}{\delta}\right)\right]^2\right\}.$$

This proves the claim. □

E.2 Proof of expected projected stationarity bounds

For completeness, we also give the expectation analogue of Lemma E.2. The proof follows the same pathwise descent argument, replacing the uniform high-probability controls of the first-moment and preconditioner errors by their corresponding second-moment bounds. As in Theorem E.2, the sums over $c_{1,t}$, $\kappa_{1,t}$, and $\vartheta_{2,t}$ can be evaluated explicitly to obtain a decay–floor decomposition; we omit this additional step for brevity.

Theorem E.3 (Expected projected-gradient stationarity gap under projected Adam)**.** *Under Assumptions B.2, E.1, and B.3, suppose* $\operatorname{diam}(\Theta) \le R$*. Then, for all integers* $T \ge 1$*, if* $\alpha \le \epsilon/(4L)$*, the iterates generated by the projected Adam update satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}}\right] \le \frac{8\big(G_1(\boldsymbol{\theta}_0)-G^\star+\mathbb{E}[\mathfrak{D}_T]\big)}{\alpha T}+\frac{12C}{\epsilon^3 T}\left(L^2R^2+d\sigma^2\right)^2\sum_{t=0}^{T-1}\vartheta_{2,t+1}$$
$$+\frac{12}{\epsilon T}\sum_{t=0}^{T-1}\left(18L^2R^2\left(\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}\right)^2+2Cd\sigma^2\kappa_{1,t+1}\right). \tag{50}$$

Here $\mathfrak{D}_T := \sum_{t=0}^{T-2}\big(G_{t+2}(\boldsymbol{\theta}_{t+1})-G_{t+1}(\boldsymbol{\theta}_{t+1})\big)_+$ *is the pathwise objective-variation budget,* $\widetilde{\boldsymbol{P}}_{t+1}$ *is the predictable Adam preconditioner, and* $\kappa_{1,t}$*, and* $\vartheta_{2,t}$ *are defined in (3).*

Proof of Theorem E.3. From the proof of Lemma E.2, before applying the high-probability error bounds, we have the pathwise inequality

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)\|^2_{\widetilde{\boldsymbol{P}}^{-1}_{t+1}} \le \frac{8\big(G_1(\boldsymbol{\theta}_0)-G^\star+\mathfrak{D}_T\big)}{\alpha T}+\frac{12q_+}{T}\sum_{t=0}^{T-1}\|\boldsymbol{r}_{t+1}\|^2$$
$$+\frac{12}{T}\sum_{t=0}^{T-1}\left\|\widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}. \tag{51}$$

Taking expectations and applying Lemma D.6 gives $\mathbb{E}\|\boldsymbol{r}_{t+1}\|^2 \le 18L^2R^2\left(\frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}\right)^2 + 2Cd\sigma^2\kappa_{1,t+1}$. Similarly, the expectation part of Lemma E.1 gives

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{P}}^{-1}_{t+1}\boldsymbol{\eta}_{t+1}\right\|^2_{\widetilde{\boldsymbol{P}}_{t+1}}\right] \le C\epsilon^{-3}\left(L^2R^2+d\sigma^2\right)^2\vartheta_{2,t+1}.$$

Substituting these two estimates into Eq. (51) and using $q_+ = \epsilon^{-1}$ yields Eq. (50). □

F Technical lemmas

F.1 Metric projection and projected-gradient mappings

We collect a few standard facts about projected-gradient mappings used in Lemma E.2. Let $\Theta \subset \mathbb{R}^d$ be nonempty, closed, and convex. For a positive definite matrix $\boldsymbol{A}$, define the $\boldsymbol{A}$-metric projection by

$$\mathcal{P}^{\boldsymbol{A}}_\Theta(\boldsymbol{z}) := \arg\min_{\boldsymbol{\theta}\in\Theta}\frac{1}{2}\|\boldsymbol{\theta}-\boldsymbol{z}\|^2_{\boldsymbol{A}}.$$

When the metric is clear from context, we suppress the superscript and write $\mathcal{P}_\Theta$. For $\boldsymbol{\theta} \in \Theta$, we write

$$N_\Theta(\boldsymbol{\theta}) := \left\{ \boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{\theta} \rangle \le 0 \ \text{ for all } \boldsymbol{y} \in \Theta \right\}$$

for the normal cone of Θ at $\boldsymbol{\theta}$. Thus $-\nabla f(\boldsymbol{\theta}) \in N_\Theta(\boldsymbol{\theta})$ is equivalent to $\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{y} - \boldsymbol{\theta} \rangle \ge 0$ for all $\boldsymbol{y} \in \Theta$, which says that no feasible first-order direction decreases f. When $\boldsymbol{\theta}$ lies in the interior of Θ, this reduces to the usual unconstrained condition $\nabla f(\boldsymbol{\theta}) = \mathbf{0}$.

Lemma F.1 (Metric projection facts)**.** *Let $\Theta \subset \mathbb{R}^d$ be nonempty, closed, and convex, and let $\boldsymbol{A} \succ 0$. Then, for every $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z})$ exists and is unique. Moreover, it satisfies the variational inequality*

$$\left\langle \boldsymbol{A}\left(\mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z}) - \boldsymbol{z}\right), \boldsymbol{y} - \mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z}) \right\rangle \ge 0, \quad \forall \boldsymbol{y} \in \Theta.$$

In addition, the projection is nonexpansive in the $\boldsymbol{A}$-metric:

$$\left\| \mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z}) - \mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z}') \right\|_{\boldsymbol{A}} \le \|\boldsymbol{z} - \boldsymbol{z}'\|_{\boldsymbol{A}}, \quad \forall \boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d.$$

Proof. The claim follows by viewing $\mathbb{R}^d$ as a Hilbert space with inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} := \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y}$. Since Θ is closed and convex, the Hilbert-space projection theorem gives existence, uniqueness, and nonexpansiveness of the projection. The variational inequality is the first-order optimality condition for the strongly convex problem defining $\mathcal{P}_\Theta^{\boldsymbol{A}}(\boldsymbol{z})$. □

For the Adam analysis, we take $\boldsymbol{A} = \widetilde{\boldsymbol{P}}_{t+1}^{-1}$. Thus $\mathcal{P}_\Theta$ denotes projection onto Θ in the $\widetilde{\boldsymbol{P}}_{t+1}^{-1}$-metric. Given a differentiable function f, a positive definite preconditioner $\boldsymbol{P}$, and a stepsize $\alpha > 0$, define the preconditioned projected-gradient mapping

$$\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) := \frac{1}{\alpha}\left[\boldsymbol{\theta} - \mathcal{P}_\Theta^{\boldsymbol{P}^{-1}}\left(\boldsymbol{\theta} - \alpha \boldsymbol{P} \nabla f(\boldsymbol{\theta})\right) \right].$$

In Lemma E.2, this becomes $\mathcal{G}_{\alpha,t}(\boldsymbol{\theta}_t)$ with $\boldsymbol{P} = \widetilde{\boldsymbol{P}}_{t+1}$ and $f = G_{t+1}$.

Lemma F.2 (Projected-gradient mapping and constrained stationarity)**.** *Let $\Theta \subset \mathbb{R}^d$ be nonempty, closed, and convex, let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, let $\boldsymbol{P} \succ 0$, and let $\alpha > 0$. Then*

$$\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = \mathbf{0} \iff -\nabla f(\boldsymbol{\theta}) \in N_\Theta(\boldsymbol{\theta}).$$

In particular, $\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = 0$ is exactly the first-order stationarity condition for the constrained problem $\min_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{\theta})$. Moreover, when $\Theta = \mathbb{R}^d$,

$$\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = \boldsymbol{P} \nabla f(\boldsymbol{\theta}), \quad \left\| \mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) \right\|_{\boldsymbol{P}^{-1}}^2 = \|\nabla f(\boldsymbol{\theta})\|_{\boldsymbol{P}}^2.$$

Proof. Let

$$\bar{\boldsymbol{\theta}} := \mathcal{P}_\Theta^{\boldsymbol{P}^{-1}}\left(\boldsymbol{\theta} - \alpha \boldsymbol{P} \nabla f(\boldsymbol{\theta})\right).$$

If $\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = \mathbf{0}$, then $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Applying Lemma F.1 with $\boldsymbol{A} = \boldsymbol{P}^{-1}$ and $\boldsymbol{z} = \boldsymbol{\theta} - \alpha \boldsymbol{P} \nabla f(\boldsymbol{\theta})$ gives, for every $\boldsymbol{y} \in \Theta$,

$$\left\langle \boldsymbol{P}^{-1}\left(\boldsymbol{\theta} - \boldsymbol{z}\right), \boldsymbol{y} - \boldsymbol{\theta} \right\rangle = \alpha \left\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{y} - \boldsymbol{\theta} \right\rangle \ge 0.$$

Equivalently, $\langle -\nabla f(\boldsymbol{\theta}), \boldsymbol{y} - \boldsymbol{\theta} \rangle \le 0$ for all $\boldsymbol{y} \in \Theta$, which is precisely $-\nabla f(\boldsymbol{\theta}) \in N_\Theta(\boldsymbol{\theta})$.

Conversely, if $-\nabla f(\boldsymbol{\theta}) \in N_\Theta(\boldsymbol{\theta})$, then $\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{y} - \boldsymbol{\theta} \rangle \ge 0$ for every $\boldsymbol{y} \in \Theta$. Hence

$$\left\langle \boldsymbol{P}^{-1}\left(\boldsymbol{\theta} - [\boldsymbol{\theta} - \alpha \boldsymbol{P} \nabla f(\boldsymbol{\theta})]\right), \boldsymbol{y} - \boldsymbol{\theta} \right\rangle \ge 0, \quad \forall \boldsymbol{y} \in \Theta.$$

By the variational characterization of the metric projection, $\boldsymbol{\theta} = \mathcal{P}_\Theta^{\boldsymbol{P}^{-1}}(\boldsymbol{\theta} - \alpha \boldsymbol{P} \nabla f(\boldsymbol{\theta}))$, and therefore $\mathcal{G}_\alpha^{\boldsymbol{P}}(\boldsymbol{\theta}) = \mathbf{0}$. The unconstrained identities follow immediately from $\mathcal{P}_{\mathbb{R}^d}^{\boldsymbol{P}^{-1}}$ being the identity map. □

F.2 MARTINGALE CONCENTRATION INEQUALITIES

Theorem F.1 (Martingale subgaussian Hoeffding inequality; cf. (Vershynin, 2018, Theorem 2.7.3))**.** *Let $(\mathcal{F}_t)_{t=0}^T$ be a filtration and let $Z_1, \dots, Z_T$ be an adapted martingale difference sequence, so that $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ almost surely. Suppose that $\|Z_t \mid \mathcal{F}_{t-1}\|_{\psi_2} \le K_t$ almost surely for deterministic constants $K_t \ge 0$. Then, for every $s \ge 0$,*

$$\mathbb{P}\left\{\sum_{t=1}^T Z_t \ge s\right\} \le \exp\left(-\frac{cs^2}{\sum_{t=1}^T K_t^2}\right),$$

where $c > 0$ is an absolute constant.

Proof. Let $S_T := \sum_{t=1}^T Z_t$. We use the exponential moment method. For any $\lambda > 0$, Markov's inequality gives

$$\mathbb{P}\{S_T \ge s\} \le e^{-\lambda s}\mathbb{E}e^{\lambda S_T}.$$

Since S_{T-1} is $\mathcal{F}_{T-1}$-measurable, the tower property gives

$$\mathbb{E}e^{\lambda S_T} = \mathbb{E}\left[e^{\lambda S_{T-1}}\mathbb{E}\left[e^{\lambda Z_T} \mid \mathcal{F}_{T-1}\right]\right].$$

By the standard subgaussian MGF bound in conditional form, $\mathbb{E}[e^{\lambda Z_t} \mid \mathcal{F}_{t-1}] \le \exp(C\lambda^2 K_t^2)$ almost surely for every $\lambda \in \mathbb{R}$. Iterating the tower-property argument therefore yields

$$\mathbb{E}e^{\lambda S_T} \le \exp\left(C\lambda^2 \sum_{t=1}^T K_t^2\right).$$

Hence,

$$\mathbb{P}\{S_T \ge s\} \le \exp\left(-\lambda s + C\lambda^2 \sum_{t=1}^T K_t^2\right).$$

Optimizing in λ by taking $\lambda = s/(2C\sum_{t=1}^T K_t^2)$ gives the claimed bound after adjusting the absolute constant. □

Theorem F.2 (Martingale subexponential Bernstein inequality; cf. (Vershynin, 2018, Theorem 2.9.1))**.** *Let $(\mathcal{F}_t)_{t=0}^T$ be a filtration and let $Z_1, \dots, Z_T$ be an adapted martingale difference sequence, so that $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ almost surely. Suppose that $\|Z_t \mid \mathcal{F}_{t-1}\|_{\psi_1} \le K_t$ almost surely for deterministic constants $K_t \ge 0$. Then, for every $s \ge 0$,*

$$\mathbb{P}\left\{\sum_{t=1}^T Z_t \ge s\right\} \le \exp\left[-c\min\left\{\frac{s^2}{\sum_{t=1}^T K_t^2}, \frac{s}{\max_{1\le t\le T} K_t}\right\}\right],$$

where $c > 0$ is an absolute constant.

Proof. Again let $S_T := \sum_{t=1}^T Z_t$. For any $\lambda > 0$, the exponential moment method gives

$$\mathbb{P}\{S_T \ge s\} \le e^{-\lambda s}\mathbb{E}e^{\lambda S_T}.$$

The conditional subexponential MGF bound implies that there exist absolute constants $c, C > 0$ such that

$$\mathbb{E}\left[e^{\lambda Z_t} \mid \mathcal{F}_{t-1}\right] \le \exp(C\lambda^2 K_t^2)$$

almost surely whenever $\lambda \le c/K_t$. Therefore, using the tower property recursively, for

$$0 \le \lambda \le \frac{c}{\max_{1\le t\le T} K_t},$$

we obtain

$$\mathbb{E}e^{\lambda S_T} \le \exp\left(C\lambda^2 \sum_{t=1}^T K_t^2\right).$$

Consequently,

$$\mathbb{P}\{S_T \ge s\} \le \exp\left(-\lambda s + C\lambda^2 \sum_{t=1}^T K_t^2\right).$$

Writing $\sigma^2 := \sum_{t=1}^{T} K_t^2$, we minimize the exponent subject to the preceding constraint. Taking

$$\lambda := \min\left\{\frac{s}{2C\sigma^2}, \frac{c}{\max_{1\le t\le T} K_t}\right\}$$

yields

$$\mathbb{P}\{S_T \ge s\} \le \exp\left[-c' \min\left\{\frac{s^2}{\sum_{t=1}^{T} K_t^2}, \frac{s}{\max_{1\le t\le T} K_t}\right\}\right]$$

for an absolute constant $c' > 0$, which proves the claim. □

F.3 L-SMOOTHNESS AND DESCENT INEQUALITIES

Lemma F.3 (Descent lemma for L-smooth functions; Nesterov (2018))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and L-smooth, meaning that*

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\| \ \text{ for all } x, y \in \mathbb{R}^n.$$

Then, for any $x, y \in \mathbb{R}^n$, we have

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \le \frac{L}{2}\|y - x\|^2.$$

G ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

We provide additional details for the numerical experiments in Appendix G.3. For each problem and regime, we tune the learning rate over a small fixed grid and choose the run with the best learning rate predictive performance.

G.1 GENERAL SETUP

Optimizers and tuning. We compare vanilla SGD and Adam. For SGD, we tune the learning rate over $\{10^{-4}, 3\cdot 10^{-4}, 10^{-3}, 3\cdot 10^{-3}, 10^{-2}, 3\cdot 10^{-2}\}$. For Adam, we tune the learning rate over the same grid as SGD with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. For each optimizer, problem, and regime, we select the learning rate that minimizes the average metric over the second half of the trajectory. Curves are then reported for the selected learning rate, averaged over seeds $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, with shaded regions denoting one standard error of the mean.

Drift and noise schedules. For vector-valued targets, the drift is generated through normalized Gaussian directions:

$$\boldsymbol{\theta}^\star_{t+1} = \boldsymbol{\theta}^\star_t + \Delta_t \frac{\boldsymbol{u}_t}{\|\boldsymbol{u}_t\|}, \ \ \boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d).$$

Thus Δ_t sets the drift magnitude, while σ_t sets the stochastic observation noise. In the two-layer MLP experiment, the same idea is implemented in function space: the teacher parameters are perturbed in a random normalized parameter direction and rescaled so that the induced change in teacher predictions has magnitude controlled by Δ_t.

Table 1: Drift and noise schedules used in the main comparison figures.

Problem	Regime	Δ_t	σ_t
Least squares	High drift / low noise	$\log(t+2)$	$5 \cdot 10^{-4} \log(t+2)$
	Low drift / high noise	$5 \cdot 10^{-4} \log(t+2)$	$10 \log(t+2)$
Logistic regression	High drift / low noise	100	1
	Low drift / high noise	10^{-4}	10
Teacher–student MLP	High drift / low noise	$5 \log(t+2)$	$10^{-3} \log(t+2)$
	Low drift / high noise	$5 \cdot 10^{-4} \log(t+2)$	$100 \log(t+2)$
Phase retrieval	High drift / low noise	$10 \log(t+2)$	$10^{-2} \log(t+2)$
	Low drift / high noise	$10^{-4} \log(t+2)$	$10 \log(t+2)$
Matrix factorization	High drift / low noise	0.1	0.01
	Low drift / high noise	$5 \cdot 10^{-4}$	1.0
Lasso	Higher drift / low noise	$2 \cdot 10^{-3} \log(t+2)$	$0.1 \log(t+2)$
	Lower drift / high noise	$10^{-3} \log(t+2)$	$10 \log(t+2)$

Evaluation. For least squares, logistic regression, and Lasso, we run $T = 800$ iterations and evaluate every 20 iterations. For teacher–student MLP and phase retrieval, we run $T = 500$ iterations and evaluate every 20 iterations, while for matrix factorization we run $T = 500$ iterations and evaluate every 5 iterations. We use mini-batches of size 1 for least squares, 32 for Lasso, and 256 for teacher–student MLP and phase retrieval.

For least squares and Lasso, we report the parameter tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2$. For logistic regression, teacher–student MLP, phase retrieval, and matrix factorization, we report the stationarity gap $\|\nabla F_t(\boldsymbol{\theta}_t)\|^2$, where for matrix factorization the gradient is taken jointly with respect to both factor matrices.

G.2 Task-specific details

Strongly convex least squares. We use dimension $d = 50$ and $n = 100$ observations. The population objective is

$$F_t(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_t^\star)\|^2,$$

where $\boldsymbol{A}^\top \boldsymbol{A}$ has eigenvalues logarithmically spaced between $\mu = 1$ and $L = 10$. Stochastic observations are generated as

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{\theta}_t^\star + \boldsymbol{\varepsilon}_t,$$

where the observation noise is scaled by $\sigma_t/\sqrt{nL}$. We report the tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2$, since the moving optimum is identifiable in parameter space.

Teacher–student MLP regression. We use a two-layer ReLU network

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}_2 \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + b_2,$$

with input dimension 100 and hidden width 128. Teacher and student networks are initialized with weights drawn at scale 0.04, and the student is warm-started from the teacher at $t = 0$. At each step, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and labels are generated as

$$y_t = f_{\boldsymbol{\theta}_t^\star}(\boldsymbol{x}) + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1).$$

We evaluate on a fixed validation set of size 1024 and report stationarity gap $\|\nabla F_t(\boldsymbol{\theta}_t)\|^2$.

Phase retrieval. We use dimension $d = 50$. Given $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, the time-varying teacher produces

$$y_t = (\boldsymbol{x}^\top \boldsymbol{w}_t^\star)^2 + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1).$$

The corresponding population objective is

$$F_t(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\left((\boldsymbol{x}^\top\boldsymbol{\theta})^2 - (\boldsymbol{x}^\top\boldsymbol{w}_t^\star)^2\right)^2\right].$$

The target $\boldsymbol{w}_t^\star$ evolves by normalized Gaussian drift, and the model is warm-started at the initial target. We use batch size 256, a validation set of size 1024, and report stationarity gap $\|\nabla F_t(\boldsymbol{\theta}_t)\|^2$.

Matrix factorization. We use a rank-5 matrix factorization problem with $m = n = 60$. We write the time-varying target matrix as

$$\boldsymbol{M}_t^\star = \boldsymbol{P}_t^\star\boldsymbol{\Sigma}(\boldsymbol{Q}_t^\star)^\top,\ \ \boldsymbol{U}_t^\star = \boldsymbol{P}_t^\star\boldsymbol{\Sigma}^{1/2},\ \ \boldsymbol{V}_t^\star = \boldsymbol{Q}_t^\star\boldsymbol{\Sigma}^{1/2},$$

where $\boldsymbol{P}_t^\star, \boldsymbol{Q}_t^\star \in \mathbb{R}^{60\times 5}$ have orthonormal columns and $\boldsymbol{\Sigma}$ is fixed. The model predicts $\boldsymbol{U}\boldsymbol{V}^\top$, with population objective $F_t(\boldsymbol{U},\boldsymbol{V}) = (2mn)^{-1}\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}_t^\star\|_F^2$. We normalize the initial target so that $\|\boldsymbol{M}_0^\star\|_F/\sqrt{mn} = 1$.

To generate nonstationarity while preserving the rank and singular values of the target, we draw fixed skew-symmetric matrices satisfying $\boldsymbol{S}_U^\top = -\boldsymbol{S}_U$ and $\boldsymbol{S}_V^\top = -\boldsymbol{S}_V$, and define the corresponding orthogonal rotations

$$\boldsymbol{R}_U = \exp(\omega\boldsymbol{S}_U),\ \ \boldsymbol{R}_V = \exp(\omega\boldsymbol{S}_V).$$

The target subspaces then evolve according to

$$\boldsymbol{P}_{t+1}^\star = \boldsymbol{R}_U\boldsymbol{P}_t^\star,\ \ \boldsymbol{Q}_{t+1}^\star = \boldsymbol{R}_V\boldsymbol{Q}_t^\star.$$

Since matrix exponentials of skew-symmetric matrices are orthogonal, $\boldsymbol{P}_t^\star$ and $\boldsymbol{Q}_t^\star$ retain orthonormal columns, so the rank and singular values of $\boldsymbol{M}_t^\star$ remain fixed. Moreover,

$$\boldsymbol{M}_t^\star = \boldsymbol{R}_U^t\boldsymbol{M}_0^\star(\boldsymbol{R}_V^t)^\top,$$

and hence, by orthogonal invariance of the Frobenius norm,

$$\|\boldsymbol{M}_{t+1}^\star - \boldsymbol{M}_t^\star\|_F = \|\boldsymbol{R}_U\boldsymbol{M}_0^\star\boldsymbol{R}_V^\top - \boldsymbol{M}_0^\star\|_F.$$

Thus the per-step drift $\Delta_t = \|\boldsymbol{M}_{t+1}^\star - \boldsymbol{M}_t^\star\|_F/\sqrt{mn}$ is constant over time, and we choose ω to attain the desired drift level.

We inject noise directly into the gradient oracle. Writing $\boldsymbol{E}_t = \boldsymbol{U}_t\boldsymbol{V}_t^\top - \boldsymbol{M}_t^\star$, we use

$$\widehat{\nabla}_{\boldsymbol{U}}F_t = \frac{1}{mn}(\boldsymbol{E}_t + \sigma_t\boldsymbol{Z}_t)\boldsymbol{V}_t,\ \ \widehat{\nabla}_{\boldsymbol{V}}F_t = \frac{1}{mn}(\boldsymbol{E}_t + \sigma_t\boldsymbol{Z}_t)^\top\boldsymbol{U}_t,$$

where $\boldsymbol{Z}_t$ is independent and mean zero with $\|\boldsymbol{Z}_t\|_F/\sqrt{mn} = 1$. Thus Δ_t and σ_t measure target drift and gradient noise on the same per-entry RMS scale. We report the stationarity gap $\|\nabla_{\boldsymbol{U}}F_t\|_F^2 + \|\nabla_{\boldsymbol{V}}F_t\|_F^2$.

G.3 Additional experimental examples

We include logistic regression and lasso regression as additional examples to test whether the same qualitative noise–drift tradeoff appears beyond the main comparison tasks.

Logistic regression. We consider convex rank-deficient logistic regression with dimension $d = 100$, rank $r = 20$, and $n = 1000$ samples. Covariates are generated as $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ has orthonormal columns and $\boldsymbol{Z} \in \mathbb{R}^{n\times r}$ has i.i.d. Gaussian entries scaled by $1/\sqrt{r}$. Let $\varsigma(z) = (1+e^{-z})^{-1}$ denote the sigmoid function and define $\boldsymbol{p}_t = \varsigma(\boldsymbol{A}\boldsymbol{\theta}_t^\star)$. The population objective is

$$F_t(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n\left[\log\big(1+\exp(\boldsymbol{a}_i^\top\boldsymbol{\theta})\big) - p_{t,i}\boldsymbol{a}_i^\top\boldsymbol{\theta}\right],$$

so that $\nabla F_t(\boldsymbol{\theta}_t^\star) = \boldsymbol{0}$.

Since $\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{Z}\boldsymbol{U}^\top\boldsymbol{\theta}$, the objective depends on $\boldsymbol{\theta}$ only through its active coordinates $\boldsymbol{U}^\top\boldsymbol{\theta}$. In particular,

$$\nabla F_t(\boldsymbol{\theta}) = \boldsymbol{U}\left[\frac{1}{n}\boldsymbol{Z}^\top\left(\varsigma(\boldsymbol{Z}\boldsymbol{U}^\top\boldsymbol{\theta}) - \boldsymbol{p}_t\right)\right] \in \mathrm{range}(\boldsymbol{U}),$$

and the objective is flat on $\mathrm{range}(\boldsymbol{U})^{\perp}$, with $\mathrm{rank}(\nabla^2 F_t) \le r$. We therefore restrict stochastic-gradient noise to the same identifiable subspace and use

$$\boldsymbol{g}_{t+1}(\boldsymbol{\theta}) = \nabla F_t(\boldsymbol{\theta}) + \frac{\sigma_t}{\sqrt{r}} \boldsymbol{U} \boldsymbol{\varepsilon}_{t+1}, \;\; \boldsymbol{\varepsilon}_{t+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r).$$

This oracle is conditionally unbiased, with covariance $(\sigma_t^2/r)\boldsymbol{U}\boldsymbol{U}^\top$ and $\mathbb{E}\|\boldsymbol{g}_{t+1}(\boldsymbol{\theta}) - \nabla F_t(\boldsymbol{\theta})\|^2 = \sigma_t^2$. Restricting the perturbation to $\mathrm{range}(\boldsymbol{U})$ avoids injecting noise into the $d - r$ null-space directions, where the loss provides no restoring gradient and isotropic noise would induce an unidentifiable random walk.

The target likewise drifts only within the active subspace according to $\boldsymbol{\theta}^\star_{t+1} = \boldsymbol{\theta}^\star_t + \Delta_t \boldsymbol{U} \boldsymbol{u}_t$, where $\boldsymbol{u}_t = \boldsymbol{h}_t / \|\boldsymbol{h}_t\|_2$ and $\boldsymbol{h}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. Thus Δ_t controls target drift while σ_t independently controls stochastic-gradient noise. We report the stationarity gap $\|\nabla F_t(\boldsymbol{\theta}_t)\|^2$.

Lasso regression. We consider Lasso regression with dimension $d = 100$, sample size $n = 200$, sparsity level $s = 10$, and regularization parameter $\lambda = 0.02$. The design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ has independent Gaussian entries scaled by $1/\sqrt{d}$. The initial target $\boldsymbol{\beta}^\star_0$ is supported on a randomly selected set $S \subset [d]$ with $|S| = s$ and is normalized so that $\|\boldsymbol{\beta}^\star_0\|_2 = 3$. The support S remains fixed throughout the experiment.

The population objective at time t is

$$F_t(\boldsymbol{\theta}) = \frac{1}{2n} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\beta}^\star_t\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1.$$

At each stochastic step, we sample a mini-batch of 32 rows from $\boldsymbol{X}$ and observe $y_{i,t} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star_t + \sigma_t \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ are independent standard Gaussian random variables. Thus the observation noise is mean zero and sub-Gaussian.

The target evolves only on its fixed support S. Writing $\boldsymbol{\beta}^\star_{t,S}$ for its active coordinates, we take a normalized random step of magnitude Δ_t along the sphere $\|\boldsymbol{\beta}^\star_{t,S}\|_2 = 3$, so that

$$\mathrm{supp}(\boldsymbol{\beta}^\star_t) = S, \;\; \|\boldsymbol{\beta}^\star_t\|_2 = 3$$

for all t. Hence the target remains s-sparse and in a fixed compact set throughout the experiment. We warm-start the model at $\boldsymbol{\beta}^\star_0$ and report the parameter tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\beta}^\star_t\|_2^2$.

G.4 Adam hyperparameter dependence

We next isolate how Adam's hyperparameters affect the noise–drift tradeoff. The first set of experiments uses the online quadratic tracking problem $F_t(\boldsymbol{\theta}) = \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star_t\|^2$. The second uses phase retrieval and reports $\|\nabla F_t(\boldsymbol{\theta}_t)\|^2$.

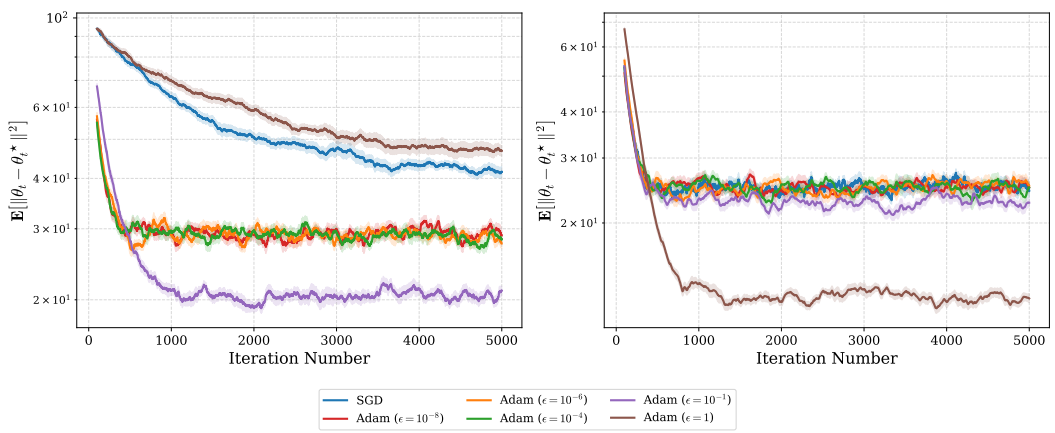

Figure 4: Dependence on the Adam stabilization parameter ϵ for the online quadratic tracking problem. The **left panel** corresponds to high drift and low noise ($\Delta = 0.2$, $\sigma^2 = 0.01$), while the **right panel** corresponds to low drift and high noise ($\Delta = 0.01$, $\sigma^2 = 1$). Across both panels, we fix $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\mu = 0.01$, and $d = 100$. Consistent with our theory, moderate stabilization performs best under high drift, while larger ϵ improves tracking in the noise-dominated regime by suppressing stochastic adaptive-preconditioning effects.

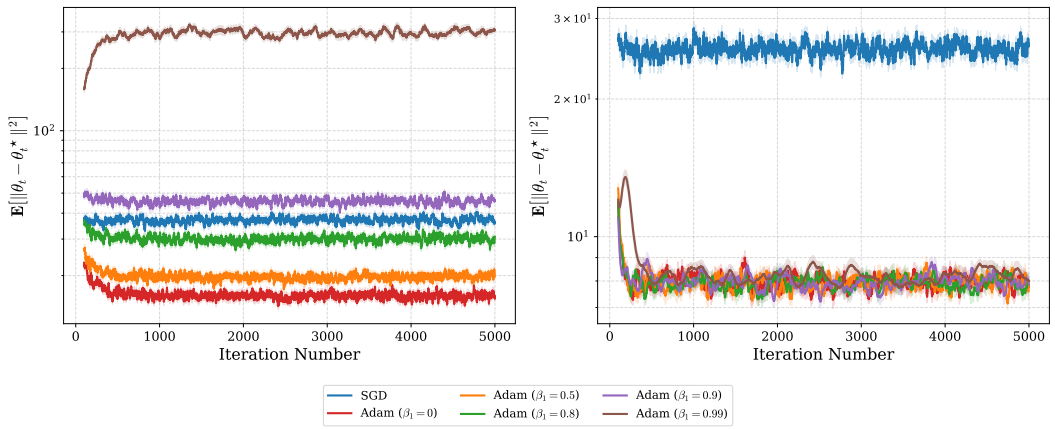

Figure 5: Dependence on the Adam first-moment parameter β_1 for the online quadratic tracking problem. The **left panel** corresponds to high drift and low noise ($\Delta = 2$, $\sigma^2 = 0.01$), while the **right panel** corresponds to low drift and high noise ($\Delta = 0.05$, $\sigma^2 = 10$). Across both panels, we fix $\beta_2 = 0.99$, $\epsilon = 10^{-8}$, $d = 100$, and $\gamma = 0.05$. Consistent with our theory, increasing β_1 degrades tracking in the drift-dominated regime by slowing adaptation, whereas under high noise the dependence on β_1 is much weaker, reflecting the tradeoff between noise averaging and stale-gradient bias.

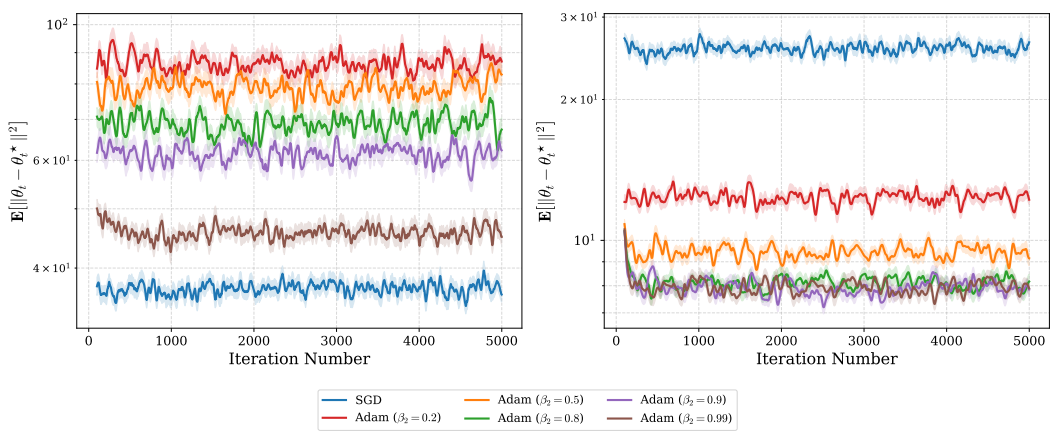

Figure 6: Dependence on the Adam second-moment parameter β_2 for the online quadratic tracking problem. The **left panel** corresponds to high drift and low noise ($\Delta = 2$, $\sigma^2 = 0.01$), while the **right panel** corresponds to low drift and high noise ($\Delta = 0.02$, $\sigma^2 = 10$). Across both panels, we fix $\beta_1 = 0.9$, $\epsilon = 10^{-8}$, $d = 100$, and $\gamma = 0.05$. Consistent with our theory, increasing β_2 improves tracking in both regimes by reducing persistent perturbations from the stochastic second-moment estimate. Over this horizon, this stability benefit dominates the slower adaptation associated with longer second-moment memory.

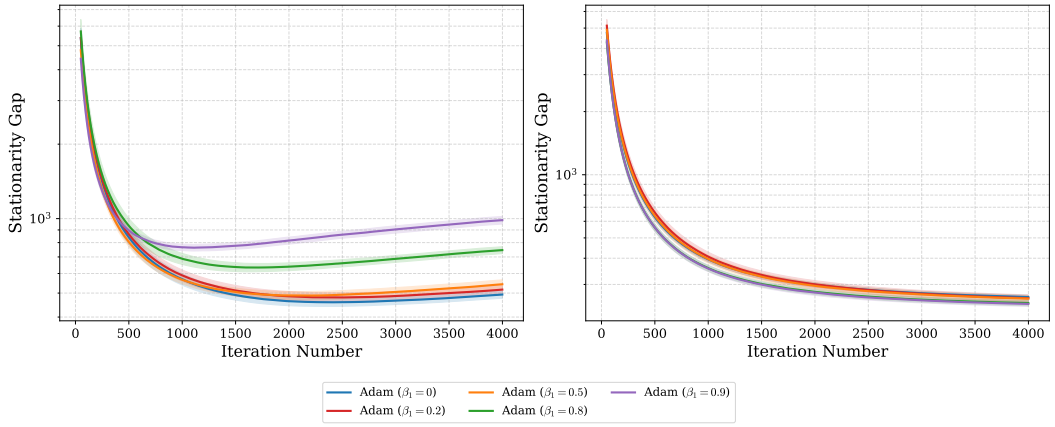

Figure 7: Dependence on the Adam first-moment parameter β_1 for the phase-retrieval problem, measured by the stationarity gap. The **left panel** corresponds to high drift and low noise ($\Delta = 0.10$, $\sigma^2 = 0.01$), while the **right panel** corresponds to no drift and high noise ($\Delta = 0$, $\sigma^2 = 1000$). Across both panels, we fix $\beta_2 = 0.999$, $\epsilon = 10^{-4}$, $d = 100$, and $\gamma = 0.01$. Consistent with our theory, increasing β_1 worsens stationarity under high drift due to stale first-moment information, while under high noise larger β_1 improves stationarity through stronger gradient averaging.

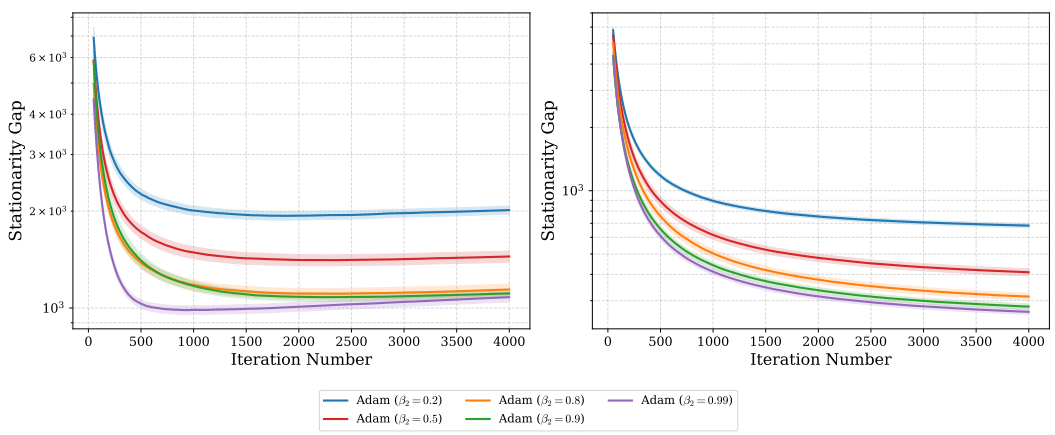

Figure 8: Dependence on the Adam second-moment parameter β_2 for the phase-retrieval problem, measured by the stationarity gap. The **left panel** corresponds to high drift and low noise ($\Delta = 0.10$, $\sigma^2 = 0.01$), while the **right panel** corresponds to no drift and high noise ($\Delta = 0$, $\sigma^2 = 1000$). Across both panels, we fix $\beta_1 = 0.9$, $\epsilon = 10^{-4}$, $d = 100$, and $\gamma = 0.01$. Consistent with our theory, increasing β_2 improves long-run stationarity in both regimes by reducing persistent perturbations from the stochastic second-moment estimate. Over this horizon, this stability benefit dominates the slower adaptation induced by longer second-moment memory.