COORDINATEWISE EPSILON FOR ADAM UNDER HEAVY-TAILED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam estimates second moments coordinatewise, but typically adds the same scalar ε to every denominator. We study how this shared ε affects convergence when noise and curvature vary across coordinates. Under coordinatewise finite-p noise moments, $p \in (1, 2]$, choosing ε_i from the noise and curvature bounds for each coordinate attains the heavy-tailed ℓ_1-stationarity benchmark $O(T^{-\frac{p-1}{3p-2}})$, matching the known optimal time exponent for smooth nonconvex stochastic optimization. For $1 < p < 2$, Adam with a shared ε_t admits a lower bound of order $\Omega(T^{-\frac{p-1}{2p}})$. This lower bound holds for every predictable scalar choice of η_t, $\beta_{1,t}$, $\beta_{2,t}$, and ε_t, precluding a dimension-uniform benchmark rate for this class.

1 INTRODUCTION

Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) are standard optimizers in large-scale language-model training (Brown et al., 2020; Touvron et al., 2023). Heavy-tailed stochastic-gradient noise has been observed in deep networks and attention-model training (Simsekli et al., 2019; Zhang et al., 2020), including recent language-model pretraining runs (Yu et al., 2026b). For linear models with cross-entropy loss, class imbalance produces heterogeneous gradients and Hessians (Kunstner et al., 2024). These observations motivate studying how Adam's coordinatewise normalization behaves when noise and curvature vary across coordinates.

Adam estimates the second moment separately in each coordinate. Its update adds the same scalar to every denominator:

$$x_{t+1,i} = x_{t,i} - \eta_t \frac{m_{t,i}}{\sqrt{v_{t,i}} + \varepsilon_t}.$$

The additive term ε_t prevents small denominators and determines how strongly the update responds to the second-moment estimate (Zaheer et al., 2018; Tong et al., 2022). If ε_t dominates, differences between the second-moment estimates have little effect on the denominators. If ε_t is small, those differences directly shape the updates.

Yu et al. (2026a, Eq. (2)) ask whether AdamW can attain the heavy-tailed ℓ_1-stationarity benchmark of sign-based methods. We adopt the same curvature and noise assumptions as their gradient-independent setting, with one stochastic gradient per step and zero weight decay. Write $L = \sum_i L_i$ for the total coordinatewise curvature bound and $\Sigma = \sum_i \sigma_i$ for the total finite-p noise scale, where $p \in (1, 2]$. We compare shared and coordinatewise choices of ε at fixed L, Σ, allowing the dimension and the allocation of these scales to vary. Our target is their benchmark uniformly over these choices. We refer to $T^{-\frac{p-1}{3p-2}}$ as the sharp heavy-tailed rate, since this time exponent is known to be optimal for smooth nonconvex stochastic optimization (Liu & Zhou, 2025).

For $1 < p < 2$, Adam with a shared ε_t has a lower bound with the slower exponent $(p-1)/(2p)$ on instances whose dimension grows with T, even when all scalar hyperparameters may depend on the past. In contrast, choosing ε_i according to the coordinatewise noise and curvature bounds attains the benchmark for every $p \in (1, 2]$.

The upper bound balances two goals: choosing ε_i large enough to control momentum noise and gradient drift in each coordinate, while keeping $\sum_i \varepsilon_i$ small. Given the coordinatewise bounds, we

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.

take, with $\rho_2 = 1 - \beta_2$,

$$\varepsilon_i \asymp \rho_2^{-\frac{2-p}{2p}} \max\left\{\sigma_i, \frac{L_i}{L}\Sigma\right\}. \tag{1}$$

The maximum combines the noise and curvature requirements, and $\sum_i \max\left\{\sigma_i, \frac{L_i}{L}\Sigma\right\} \le 2\Sigma$ controls the total.

The lower-bound construction pairs coordinates and adds a control coordinate. On an event of constant probability, the control coordinate offsets the progress of all pairs, keeping the iterates in a region with constant nonzero gradient.

Contributions.

1. **Coordinatewise upper bound.** Theorem 4.1 attains the benchmark dependence on Δ, L, Σ, T, and in particular matches the known optimal heavy-tailed time exponent. A self-normalized tracking estimate controls heavy-tailed momentum noise while retaining the random denominator for large observations.
2. **Lower bound for a shared ε.** Theorem 4.2 gives a lower bound for $1 < p \le 2$ through a pathwise comparison of normalized updates. The oracle outputs are gradients of scalar sample functions.

2 Related work

Adaptive normalization and Adam theory. AdaGrad introduced diagonal scaling by accumulated coordinatewise gradients; Adam uses exponential moving averages to estimate first and second moments (Duchi et al., 2011; Kingma & Ba, 2015). The classical counterexample of Reddi et al. (2018) concerns exponential forgetting and nonmonotone effective stepsizes. Heilman & Mohanty (2026) extend this phenomenon to fixed $\beta_1, \beta_2, \varepsilon$, while Zhou et al. (2019) study same-time correlation between the gradient and its second-moment estimate. Convergence analyses address bounded gradients, finite or affine variance, generalized smoothness, and suitable schedules (Chen et al., 2019; Défossez et al., 2022; Zhang et al., 2022; Hong & Lin, 2024). Under Euclidean smoothness and coordinatewise bounded variance, Li et al. (2025) establish an $O(\sqrt{d}\,T^{-1/4})$ expected ℓ_1-gradient rate for AdamW, with Adam as a special case.

Liu et al. (2025) analyze AdaGrad under anisotropic smoothness and noise conditions, while Jiang et al. (2025) combine coordinatewise smoothness and variance with ℓ_1 stationarity to compare AdaGrad and SGD. Together, these works motivate expressing guarantees in a coordinatewise geometry and tracking how dimension enters when the coordinate scales are aggregated. We use this viewpoint to compare shared and coordinatewise choices of ε at fixed $L = \sum_i L_i$ and $\Sigma = \sum_i \sigma_i$.

Heavy-tailed nonconvex optimization. For $1 < p < 2$, a finite pth noise moment need not imply finite variance. Gradient clipping gives expectation and high-probability guarantees in this regime (Zhang et al., 2020; Cutkosky & Mehta, 2021). Zhang et al. (2020) establish the classical heavy-tailed time exponent $(p-1)/(3p-2)$ under a global-norm nonconvex model and also study coordinatewise clipping. Later work develops refined clipping and normalization methods (Liu et al., 2023; Hübler et al., 2025; Liu & Zhou, 2025). Models in this literature vary in whether moments are imposed on stochastic gradients or centered noise, and in whether control is global or coordinatewise. Sign-based methods provide the coordinatewise ℓ_1 comparison relevant here (Yu et al., 2026b;a).

Adaptive methods under heavy-tailed noise. Heavy-tailed guarantees for adaptive methods cover several algorithms and stationarity notions. Chezhegov et al. (2025) show that explicit clipping improves the high-probability confidence dependence of Adam-Norm and AdaGrad-Norm. Liu (2026a) proves an expected rate for coordinatewise AdaGrad when $p > 4/3$ and gives an algorithm-dependent lower bound. Recent online-to-nonconvex analyses establish guarantees for Adam under relaxed stationarity criteria (Liu, 2026b; Pang, 2026). The setting considered here uses the expected ℓ_1 norm of the true gradient, coordinatewise smoothness, and centered coordinatewise noise moments.

The choice of ε. The effect of ε has also been studied: it can suppress adaptivity and enter convergence constants (Zaheer et al., 2018), while softplus transformations and alternative placements

change anisotropic effective steps (Tong et al., 2022; Yuan & Gao, 2020). Yu et al. (2026a) formulate the AdamW benchmark and give a corridor obstruction under specified conditions on β_1, β_2 and the stepsize. This motivates treating ε as part of the optimizer's geometry and comparing a shared stabilization scale with coordinatewise ones.

3 Preliminaries and Problem Setup

3.1 Notation and Objective Geometry

For a positive definite matrix D, write $\|x\|_D = (x^\top D x)^{\frac{1}{2}}$; $\odot$ denotes coordinatewise multiplication. We use $a \lesssim b$ and $a \asymp b$ for one-sided and two-sided bounds up to positive constants. Unless stated otherwise, constants may depend on p and the fixed compact set $\mathcal{K}$ of ratios ρ_2/ρ_1 introduced below, but not on d, T, Δ, L, Σ or the allocation of the coordinatewise scales.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and lower bounded. Fix a diagonal geometry

$$D_L = \mathrm{Diag}(L_1, \dots, L_d), \qquad L_i > 0, \qquad L = \sum_{i=1}^{d} L_i.$$

Assumption 3.1 (Coordinatewise smoothness)**.** We assume

$$\|\nabla f(y) - \nabla f(x)\|_{D_L^{-1}} \le \|y - x\|_{D_L}, \qquad x, y \in \mathbb{R}^d. \tag{2}$$

For twice continuously differentiable f, this condition is equivalent to

$$-D_L \preceq \nabla^2 f(x) \preceq D_L, \qquad x \in \mathbb{R}^d.$$

Assumption 3.1 implies the descent inequality

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|_{D_L}^2. \tag{3}$$

Assumption 3.2 (Initial gap)**.** The initial point satisfies

$$f(x_1) - \inf_x f(x) \le \Delta. \tag{4}$$

3.2 Stochastic Oracle and Stationarity

Let $\mathcal{F}_{t-1}$ denote the history before the tth oracle query, including the initialization, past oracle observations, and algorithmic randomness available before that query. We require x_t and all time-t parameters to be $\mathcal{F}_{t-1}$-measurable and call such parameter choices predictable.

Assumption 3.3 (Finite-moment oracle)**.** At time t, the oracle returns

$$g_t = \nabla f(x_t) + \xi_t,$$

where, coordinatewise,

$$\mathbb{E}[\xi_{t,i} \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[|\xi_{t,i}|^p \mid \mathcal{F}_{t-1}] \le \sigma_i^p, \qquad p \in (1, 2]. \tag{5}$$

Write $\Sigma = \sum_i \sigma_i$ and assume $\Delta, L, \Sigma > 0$. The moment condition controls the centered noise and allows unbounded gradients. Coordinates may be dependent, and the noise distribution may depend on the past. At $p = 2$ this is a coordinatewise conditional variance bound; for $p < 2$ it permits infinite variance.

Stationarity criterion. We use the ℓ_1-stationarity criterion in the open problem of Yu et al. (2026a),

$$R_T = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(x_t)\|_1. \tag{6}$$

3.3 ADAM FAMILIES

Starting from $m_0 = v_0 = 0$, consider the family

$$m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t, \tag{7}$$

$$v_t = \beta_{2,t} v_{t-1} + (1 - \beta_{2,t}) g_t^{\odot 2}, \tag{8}$$

$$x_{t+1,i} = x_{t,i} - \eta_t \frac{m_{t,i}}{\sqrt{v_{t,i}} + \varepsilon_{t,i}}. \tag{9}$$

The upper bound uses fixed $\eta_t \equiv \eta$, $\beta_{k,t} \equiv \beta_k$, and $\varepsilon_{t,i} \equiv \varepsilon_i$; write $\rho_k = 1 - \beta_k$. For the lower bound, $\eta_t \geq 0$, $\beta_{k,t} \in [0, 1)$, and $\varepsilon_t \geq 0$ may be arbitrary $\mathcal{F}_{t-1}$-measurable scalars, with the restriction $\varepsilon_{t,i} \equiv \varepsilon_t$. This class includes fixed-parameter Adam and standard scalar bias correction after its deterministic factors are absorbed into η_t and ε_t. We require updates to be finite and well defined almost surely on the problem class. We study the zero-weight-decay setting.

4 MAIN RESULTS

4.1 UPPER BOUND WITH COORDINATEWISE ε_i

Define

$$\rho_\star = \left(\frac{\Delta L}{\Sigma^2 T} \right)^{\frac{p}{3p-2}}. \tag{10}$$

Choosing ε_i. Write

$$S = \sum_i \varepsilon_i, \qquad \kappa = \max_i \frac{L_i}{\varepsilon_i}.$$

For the stepsize $\eta = \lambda \rho_1 \Sigma \rho_2^{-\frac{2-p}{2p}} / L$, with $\lambda > 0$ sufficiently small, the upper proof uses three sufficient requirements:

1. *Momentum noise.* Each coordinate satisfies
$$\varepsilon_i \geq \rho_2^{-\frac{2-p}{2p}} \sigma_i.$$
2. *Gradient drift.* The updates satisfy
$$\eta \kappa \leq \lambda \rho_1.$$
The coordinatewise bound $\varepsilon_i \geq \rho_2^{-\frac{2-p}{2p}} (L_i / L) \Sigma$ ensures this inequality and allows the drift contribution to be absorbed.
3. *Total size.* The sum satisfies
$$S \lesssim \rho_2^{-\frac{2-p}{2p}} \Sigma.$$
After balancing the stepsize and averaging parameters, the proof gives $R_T \lesssim \sqrt{\rho_1}\, S$.

Taking the larger of the two coordinatewise lower bounds meets the first two requirements. Their sums are both $\rho_2^{-\frac{2-p}{2p}} \Sigma$, so this choice also meets the third.

Theorem 4.1 (Coordinatewise choices of ε_i attain the sharp rate)**.** *Fix $p \in (1, 2]$. Let $\mathcal{K}$ be a compact subset of $(0, 4/(2-p))$ when $p < 2$, or of $(0, \infty)$ when $p = 2$. There exist constants $c, C, \lambda > 0$, depending only on p and $\mathcal{K}$, such that the following holds whenever $\rho_\star \leq c$. Choose*

$$\rho_1 = \rho_\star, \qquad \frac{\rho_2}{\rho_1} \in \mathcal{K}, \qquad \eta = \lambda \frac{\rho_1}{L} \rho_2^{-\frac{2-p}{2p}} \Sigma,$$

using the constant λ above, and set

$$\varepsilon_i = \rho_2^{-\frac{2-p}{2p}} \max\left\{ \sigma_i, \frac{L_i}{L} \Sigma \right\}. \tag{11}$$

Then

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(x_t)\|_1 \leq C (\Delta L)^{\frac{p-1}{3p-2}} \Sigma^{\frac{p}{3p-2}} T^{-\frac{p-1}{3p-2}}. \tag{12}$$

The bound is uniform over all objectives and oracles satisfying Assumptions 3.1–3.3.

The choice (11) gives

$$\sum_i \varepsilon_i \le 2\rho_2^{-\frac{2-p}{2p}}\Sigma. \tag{13}$$

4.2 Lower bound for a shared ε

Theorem 4.2 (Lower bound for a shared ε)**.** *Fix $1 < p \le 2$. There are constants $c, \delta_0 > 0$, depending only on p, such that for every horizon T, every rule in the predictable Adam class with a shared ε_t, and every*

$$0 < \delta \le \min\{\delta_0, T^{-\frac{p-1}{p}}\},$$

there exist a dimension d, a convex continuously differentiable objective, an initial point, and a stochastic first-order oracle whose realizations are gradients of convex continuously differentiable scalar sample functions, satisfying

$$f(x_1) - \inf f \le 1, \qquad L \le 1, \qquad \Sigma \le 1,$$

for which

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(x_t)\|_1 \ge c\sqrt{\delta}. \tag{14}$$

For sufficiently large T, taking $\delta = T^{-\frac{p-1}{p}}$ gives a fixed-gap lower bound with exponent $(p-1)/(2p)$.

4.3 Consequences

Corollary 4.3 (Shared ε under balanced coordinate scales)**.** *Under the assumptions of Theorem 4.1, suppose*

$$\max_i \sigma_i \le C_0 \frac{\Sigma}{d}, \qquad \max_i L_i \le C_0 \frac{L}{d} \tag{15}$$

for a constant C_0. Set

$$\varepsilon_i \equiv \varepsilon = C_0 \rho_2^{-\frac{2-p}{2p}} \frac{\Sigma}{d}.$$

Then Adam with a shared ε satisfies, for a constant C depending only on p, $\mathcal{K}$, and C_0,

$$R_T \le C(\Delta L)^{\frac{p-1}{3p-2}} \Sigma^{\frac{p}{3p-2}} T^{-\frac{p-1}{3p-2}}.$$

Remark 4.4 (Fixed dimension versus dimension-uniform rates)**.** *The lower-bound construction uses $d = 1 + 2N = O(\delta^{-1})$ coordinates. For $\delta = T^{-\frac{p-1}{p}}$, this gives $d = O(T^{\frac{p-1}{p}})$. Growth of the dimension with T is unavoidable for obtaining a strictly slower time exponent with a positive T-independent prefactor uniformly over dimensions and coordinate scales. Indeed, for arbitrary coordinate scales, (15) always holds with $C_0 = d$, since $\max_i \sigma_i \le \Sigma$ and $\max_i L_i \le L$. Thus Corollary 4.3 applies with the shared choice $\varepsilon = \Sigma\rho_2^{-\frac{2-p}{2p}}$ and the remaining parameters of Theorem 4.1. The proof in Appendix C gives a linear dependence on C_0, yielding*

$$R_T \le Cd(\Delta L)^{\frac{p-1}{3p-2}} \Sigma^{\frac{p}{3p-2}} T^{-\frac{p-1}{3p-2}}$$

whenever $\rho_\star \le c$, with C depending only on p and $\mathcal{K}$. Hence, at every fixed dimension, a shared ε attains the benchmark time exponent uniformly over coordinate scales, albeit with a dimension-dependent factor. Consequently, for fixed positive Δ, L, Σ and fixed d, no lower bound with a strictly slower time exponent and a positive T-independent prefactor can hold for every rule in our predictable Adam class as $T \to \infty$.

Table 1 summarizes the coordinate conditions and tuning information for these comparisons.

Table 1: Comparison of the main results under Assumptions 3.1–3.3, using the ℓ_1 stationarity criterion.

ε / result	Coordinate scales	Tuning information or freedom	Guarantee
Coordinatewise upper	Arbitrary	Known $p, \Delta, T, (L_i)_i, (\sigma_i)_i$	(12), uniformly in d and scales
Shared upper	Balanced as in (15)	Known $p, \Delta, T, L, \Sigma, d, C_0$	Same rate; constant depends on C_0
Shared lower	d grows with T; $\Delta, L, \Sigma \le 1$	Predictable scalar hyperparameters	$\Omega(T^{-\frac{p-1}{2p}})$ for $1 < p \le 2$

5 Proof Overview

The upper bound uses the random denominator to control heavy-tailed momentum noise. The lower bound uses paired coordinates and a control coordinate to bound progress under shared scalar tuning. We give the key estimates here; the appendices contain the full proofs.

5.1 Upper bound: controlling momentum noise

Under the tuning of Theorem 4.1, the proof first bounds the weighted gradient term, then the average denominator, and finally the true gradient norm. Write $h_t = \nabla f(x_t)$ and $D_{t,i} = \sqrt{v_{t,i}} + \varepsilon_i$. With $S = \sum_i \varepsilon_i$, the three key relations are

$$\frac{1}{T}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} \lesssim \frac{\Delta}{\eta T} + \rho_1 S, \tag{16}$$

$$\frac{1}{T}\sum_{t,i} \mathbb{E} D_{t,i} \le 2\left(\frac{1}{T}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}}\right) + CS, \tag{17}$$

$$\begin{aligned} R_T^2 &\le \left(\frac{1}{T}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}}\right)\left(\frac{1}{T}\sum_{t,i} \mathbb{E} D_{t,i}\right) \\ &\le \left(\frac{1}{T}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}}\right)\left(2\frac{1}{T}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + CS\right). \end{aligned} \tag{18}$$

The first follows from momentum tracking and smooth descent; the second controls the denominators without an additional dimension factor; the third is Cauchy–Schwarz over time, coordinates, and probability. The prescribed parameters balance $\Delta/(\eta T)$ with $\rho_1 S$. Thus (16)–(18) give

$$R_T^2 \lesssim \rho_1 S(\rho_1 S + S) \lesssim \rho_1 S^2, \qquad R_T \lesssim \sqrt{\rho_1}\, S \asymp \Sigma \rho_1^{\frac{p-1}{p}}.$$

Tracking under heavy-tailed noise. The main difficulty in (16) is controlling $e_t = m_t - h_t$ when the noise may have infinite variance. Fix a coordinate and suppress its index. The noise part of the first moment is

$$n_t = \rho_1 \sum_{s=1}^{t} \beta_1^{t-s} \xi_s.$$

The coordinatewise choice of ε permits the self-normalized estimate

$$\sum_{t=1}^{T} \mathbb{E}\frac{n_t^2}{D_t} \le C\rho_1 T\varepsilon + C\rho_1\rho_2 \sum_{t=1}^{T} \mathbb{E}\frac{h_t^2}{D_t}. \tag{19}$$

Its right-hand side contains the same weighted gradient term as (16), with a coefficient that can be absorbed. Initialization and gradient drift form the other parts of e_t. Lemma A.1 controls them, using $\eta\kappa \le \lambda\rho_1$ for the drift.

For $1 < p < 2$, the proof splits the noise at $\varepsilon/\sqrt{\rho_2}$. At this threshold, the term $\sqrt{\rho_2}|\xi_t|$ becomes comparable to ε. Below this threshold, truncated second moments and conditional centering control the contribution. Above it, the random denominator controls large observations: on $|\xi_t| > 2|h_t|$,

$$D_t \geq \frac{\sqrt{\rho_2}|\xi_t|}{2}, \qquad \frac{\xi_t^2}{\varepsilon^{p-1} D_t^{2-p}} \leq \frac{C|\xi_t|^p}{\varepsilon^{p-1}\rho_2^{(2-p)/2}}.$$

On the complementary event, the bound contributes to the weighted gradient term in (19).

Propagation through the moment recursions. For $t \geq s$, $D_t \geq \beta_2^{(t-s)/2} D_s$. Combined with the first-moment weights, this produces a geometric series with ratio $\beta_1^2/\beta_2^{(2-p)/2}$. Write

$$\delta_p = 1 - \frac{\beta_1^2}{\beta_2^{(2-p)/2}}. \tag{20}$$

For $\rho_2/\rho_1 \in \mathcal{K}$ and sufficiently small ρ_1, $\delta_p \asymp \rho_1 > 0$; the time sum is therefore at most $1/\delta_p$. Appendix D gives the centered noise decomposition, cross-term bounds, and full time summation proving (19). At $p = 2$, the same estimate follows directly from conditional variance and orthogonality.

The average denominator. To obtain (17), the recursion bounds past gradient contributions by the weighted gradient term in (16). The finite-p noise moments contribute at most $C\sigma_i \rho_2^{-(2-p)/(2p)} \leq C\varepsilon_i$ in coordinate i. Summing over coordinates gives CS, controlled by (13). Appendix A proves the descent and denominator bounds in (16)–(17).

5.2 Lower bound: a pathwise comparison

We construct an objective

$$f(z) = \phi(Y(z)), \qquad \phi(0) = 0, \qquad \phi'(y) = \mathrm{clip}(y, 0, 1),$$

where Y is affine with a nonzero coefficient vector. On $Y(z) \geq 1$, the gradient of f equals this fixed vector. Starting from $Y(z_1) = 1$, the construction maintains, on an event of constant probability,

$$Y(z_{t+1}) \geq Y(z_t) \geq 1 \qquad (1 \leq t \leq T).$$

This gives a lower bound on the gradient norm at every iterate.

The construction has three parts. A control coordinate receives wrong-sign observations throughout this event and moves in a direction that increases Y. Coordinate pairs have a uniformly bounded total decrease under arbitrary shared averaging weights. Pairing makes this bound depend only on the totals of the common first- and second-moment weights, allowing the same comparison to cover arbitrary predictable scalar tuning. The affine function Y combines these contributions so that the increase offsets the decrease.

Objective and coordinate scales. Fix δ as in Theorem 4.2. With a fixed positive constant γ chosen in Appendix B, set

$$N = \lceil \delta^{-1} \rceil, \qquad A = \frac{\gamma}{N}, \qquad b = \frac{\gamma\sqrt{\delta}}{N}.$$

Use one control coordinate u and N pairs $(w_{i,+}, w_{i,-})$. With $s_i = (w_{i,+} + w_{i,-})/2$, define

$$Y(z) = 1 + \gamma\delta\, u + b\sum_{i=1}^{N} s_i.$$

At $z_1 = 0$, $Y(z_1) = 1$ and the initial gap is $1/2$. On $Y(z) \geq 1$, the gradient is the coefficient vector ϱ of Y and

$$\|\nabla f(z)\|_1 = \gamma\delta + Nb = \gamma(\delta + \sqrt{\delta}) \geq \gamma\sqrt{\delta}.$$

Oracle and averaging weights. Let $\theta = 1/(8T)$ and draw fresh randomness independently at each step. The control observation is $-\gamma\delta$ with probability $1-\theta$ and $\gamma\delta(2/\theta - 1)$ otherwise. Each pair receives a uniformly random permutation of $\{A+b, -A\}$, using one common sign independent of the control draw. These observations form a vector q_t with $\mathbb{E}q_t = \varrho$. Globally, set

$$g_t(z) = \nabla f(z) + q_t - \varrho = \nabla\big[f(z) + \langle q_t - \varrho, z\rangle\big].$$

Let $\mathcal{A}$ be the event that all control observations take their common value; independent draws give $\Pr(\mathcal{A}) = (1-\theta)^T \geq 7/8$.

Along any realized history, the recursions express m_t and v_t as nonnegative weighted sums of past gradients and squared gradients. Write the common scalar weights as $\alpha_{t,s}, \nu_{t,s}$ and their totals as

$$P_t = \sum_{s\leq t} \alpha_{t,s}, \qquad V_t = \sum_{s\leq t} \nu_{t,s} > 0.$$

Lemma 5.1 (Sum of the two normalized updates)**.** *Suppose a coordinate pair receives a permutation of $\{A+b, -A\}$ at each step up to time t. For $A, b > 0$, the averaging weights above, and a shared $\varepsilon_t \geq 0$, let $D^{\pm}_{t,i} = \sqrt{v^{\pm}_{t,i}} + \varepsilon_t$. Then*

$$0 \leq \frac{m^{+}_{t,i}}{D^{+}_{t,i}} + \frac{m^{-}_{t,i}}{D^{-}_{t,i}} \leq \frac{2bP_t}{A\sqrt{V_t} + \varepsilon_t}. \tag{21}$$

The pair has a positive first-moment sum. The absolute difference between its denominators, divided by their sum, is at most $b/(2A+b)$. This limits the effect of unequal normalization on the sum of the two updates. The proof is in Appendix B.4.

The pathwise invariant. On $\mathcal{A}$, suppose $Y(z_s) \geq 1$ up to time t, so the observations equal q_s. The control moments are $m_{t,0} = -\gamma\delta P_t$ and $v_{t,0} = \gamma^2\delta^2 V_t$. Since $V_t > 0$ and each pair has $v^{\pm}_{t,i} \geq A^2 V_t$, all denominators are positive even when $\varepsilon_t = 0$. The control contribution and the total contribution of the pairs satisfy

$$\gamma\delta(u_{t+1} - u_t) = \frac{\gamma^2\delta\eta_t P_t}{\gamma\sqrt{V_t} + \frac{\varepsilon_t}{\delta}},$$
$$0 \leq -b\sum_i (s_{t+1,i} - s_{t,i}) \leq \frac{Nb^2\eta_t P_t}{A\sqrt{V_t} + \varepsilon_t} = \frac{\gamma^2\delta\eta_t P_t}{\gamma\sqrt{V_t} + N\varepsilon_t}.$$

Consequently,

$$Y(z_{t+1}) - Y(z_t) \geq \gamma^2\delta\eta_t P_t \left[\frac{1}{\gamma\sqrt{V_t} + \frac{\varepsilon_t}{\delta}} - \frac{1}{\gamma\sqrt{V_t} + N\varepsilon_t}\right] \geq 0, \tag{22}$$

because $N \geq \delta^{-1}$. Starting from $Y(z_1) = 1$, induction keeps every iterate in the affine region on $\mathcal{A}$. Hence $R_T \geq (7/8)\gamma\sqrt{\delta}$.

Appendix B gives the full construction and induction, and Appendix E verifies $L \leq 1$ and $\Sigma \leq 1$. Choosing $\delta = T^{-\frac{p-1}{p}}$ for sufficiently large T gives the stated time exponent with $d = 1 + 2\lceil\delta^{-1}\rceil$.

6 Conclusion

Adam is usually viewed as a coordinatewise adaptive method because its second-moment estimate is coordinatewise. The results here show that this view is incomplete. The denominator contains two scales, $\sqrt{v_{t,i}}$ and ε. When ε is shared, the normalization is only as coordinatewise as a single stabilization scale allows it to be. If the relevant noise and curvature scales differ substantially across coordinates, one scalar is asked to regularize several incompatible local regimes.

This points to a distinction that is easy to miss in adaptive methods. Coordinatewise statistics do not by themselves imply coordinatewise tuning. A global regularization applied after those statistics can reintroduce a common scale and limit the adaptivity that the coordinatewise estimator was meant to provide. For Adam, ε is a particularly simple instance of this mismatch.

Reproducibility statement

All assumptions and tuning rules are stated in Sections 3–4. Section 5 gives the key estimates and the pathwise lower-bound argument. Complete proofs and verification of the assumptions are provided in the appendices.

AI use statement

Generative AI tools were used to assist with manuscript drafting and editing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-norm and AdaGrad-norm when the noise is heavy-tailed. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10269–10333, 2025.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Steven Heilman and Sampad Mohanty. On the convergence of Adam, revisited. *arXiv preprint arXiv:2607.03519*, 2026.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421, 2025.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3124–3158, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Frederik Kunstner, Alan Milligan, Robin Yadav, 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*, volume 37, pp. 30106–30148, 2024.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\sqrt{d}/K^{1/4})$ convergence rate of AdamW measured by ℓ_1 norm. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under anisotropic smoothness. In *International Conference on Learning Representations*, 2025.

Zijian Liu. Can adaptive gradient methods converge under heavy-tailed noise? A case study of AdaGrad. *arXiv preprint arXiv:2605.18694*, 2026a. Accepted at the 43rd International Conference on Machine Learning (ICML 2026).

Zijian Liu. Adam converges in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2606.22326*, 2026b.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2266–2290, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yijiang Pang. The convergence behavior of Adam under heavy-tailed noise. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5244–5265, 2026.

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

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837, 2019.

Qianqian Tong, Guannan Liang, and Jinbo Bi. Calibrating the adaptive learning rate to improve convergence of Adam. *Neurocomputing*, 481:333–356, 2022. doi: 10.1016/j.neucom.2022.01.014.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Open problem: Is AdamW effective under heavy-tailed noise? *arXiv preprint arXiv:2606.23676*, 2026a.

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

Wei Yuan and Kai-Xin Gao. EAdam optimizer: How ϵ impact Adam. *arXiv preprint arXiv:2011.02150*, 2020.

Manzil Zaheer, Sashank J. Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28386–28399, 2022.

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

A PROOF OF THE UPPER BOUND WITH COORDINATEWISE ε_i

The proof has two stages. First, we control the accumulated tracking error and use smooth descent to bound $T^{-1}\sum_{t,i}\mathbb{E}[h_{t,i}^2/D_{t,i}]$. Second, we bound the average denominator and apply Cauchy–Schwarz to recover the unweighted ℓ_1 criterion. Appendix D proves the noise estimate used in the first stage.

Throughout the appendices, c and C may change from line to line and may depend on p and $\mathcal{K}$. Additional dependence is stated when it first appears. For brevity, write

$$R_\star = (\Delta L)^{\frac{p-1}{3p-2}}\Sigma^{\frac{p}{3p-2}}T^{-\frac{p-1}{3p-2}}.$$

Also write

$$h_t = \nabla f(x_t), \qquad D_{t,i} = \sqrt{v_{t,i}} + \varepsilon_i.$$

Set $x_0 = x_1$, $h_0 = h_1$, $e_0 = -h_1$, and, for $t \geq 1$,

$$e_t = m_t - h_t.$$

Subtracting h_t from the first-moment update separates gradient drift from momentum noise:

$$\begin{aligned} e_t = m_t - h_t &= \beta_1(m_{t-1} - h_{t-1}) + \beta_1(h_{t-1} - h_t) + \rho_1\xi_t \\ &= \beta_1 e_{t-1} + \beta_1(h_{t-1} - h_t) + \rho_1\xi_t. \end{aligned} \tag{23}$$

Define

$$\kappa = \max_i \frac{L_i}{\varepsilon_i}. \tag{24}$$

The choice of ε_i gives

$$\begin{aligned} \sum_i \varepsilon_i &\leq \rho_2^{-\frac{2-p}{2p}}\left(\sum_i \sigma_i + \frac{\Sigma}{L}\sum_i L_i\right) = 2\Sigma\rho_2^{-\frac{2-p}{2p}}, \\ \kappa &\leq \max_i \frac{L_i}{\rho_2^{-\frac{2-p}{2p}}(L_i/L)\Sigma} = \frac{L}{\Sigma}\rho_2^{\frac{2-p}{2p}}, \\ \eta\kappa &\leq \frac{\lambda\rho_1\Sigma\rho_2^{-\frac{2-p}{2p}}}{L}\frac{L}{\Sigma}\rho_2^{\frac{2-p}{2p}} = \lambda\rho_1. \end{aligned} \tag{25}$$

All expectations below are finite on each fixed horizon. For $0 < \beta_2 < 1$, weighted Cauchy–Schwarz in the moment recursions gives

$$\frac{|m_{t,i}|}{D_{t,i}} \leq \frac{\rho_1}{\sqrt{\rho_2}}\left(\sum_{k=0}^{t-1}\left(\frac{\beta_1^2}{\beta_2}\right)^k\right)^{\frac{1}{2}}.$$

Thus the displacements, true gradients, and objective values are bounded on that horizon. Finite first moments of the noise give $\mathbb{E}\sqrt{v_{t,i}} < \infty$; the same Cauchy–Schwarz bound and $\varepsilon_i > 0$ imply integrability of $D_{t,i}$, $m_{t,i}^2/D_{t,i}$, and $e_{t,i}^2/D_{t,i}$.

A.1 COORDINATEWISE TRACKING

The descent argument needs the weighted tracking error to be a small fraction of the weighted gradient term, up to noise, initialization, and drift terms. The following estimate provides this form by expanding the momentum and splitting its noise at $\varepsilon_i/\sqrt{\rho_2}$.

Lemma A.1 (Coordinatewise tracking estimate)**.** *For every $\tau \in (0,1)$, there exists $\bar{\rho} > 0$, depending on p, $\mathcal{K}$, and τ, such that, whenever $\rho_2/\rho_1 \in \mathcal{K}$ and $0 < \rho_1 \leq \bar{\rho}$,*

$$\begin{aligned} \sum_{t=1}^T\sum_i \mathbb{E}\frac{e_{t,i}^2}{D_{t,i}} \leq \tau\sum_{t=1}^T\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + C\left(\rho_1 T\sum_i \varepsilon_i + \sum_i \frac{e_{0,i}^2}{\varepsilon_i\delta_p}\right) \\ + \frac{C}{\delta_p^2}\sum_{t=1}^T\sum_i \frac{\beta_1^2\mathbb{E}(h_{t-1,i} - h_{t,i})^2}{\varepsilon_i}. \end{aligned} \tag{26}$$

Appendix D derives the lemma from the conditional moment assumption. We next control the initial state and the gradient drift.

A.2 Proof of Theorem 4.1

Proof. We first control tracking, then bound the weighted gradient term, and finally recover the unweighted stationarity criterion.

Tracking error. With $e_0 = -h_1$ and $y = x_1 - D_L^{-1} h_1$, smooth descent gives

$$\inf f \le f(y) \le f(x_1) - \|h_1\|^2_{D_L^{-1}} + \frac{1}{2}\|h_1\|^2_{D_L^{-1}} = f(x_1) - \frac{1}{2}\|h_1\|^2_{D_L^{-1}},$$
$$\sum_i \frac{e_{0,i}^2}{\varepsilon_i} \le \kappa \|h_1\|^2_{D_L^{-1}} \le 2\kappa(f(x_1) - \inf f) \le 2\kappa\Delta.$$

Coordinatewise smoothness, the Adam update, and $D_{t,i} \ge \varepsilon_i$ give

$$\begin{aligned}
\sum_{t=1}^T \sum_i \mathbb{E}\frac{\beta_1^2 (h_{t-1,i} - h_{t,i})^2}{\varepsilon_i} &\le \beta_1^2 \kappa \sum_{t=1}^T \mathbb{E}\|h_{t-1} - h_t\|^2_{D_L^{-1}} \\
&\le \beta_1^2 \eta^2 \kappa \sum_{t=1}^T \mathbb{E}\left\| \left(\frac{m_{t-1,i}}{D_{t-1,i}}\right)_i \right\|^2_{D_L} \\
&\le \beta_1^2 \eta^2 \kappa \sum_{t=1}^T \sum_i \mathbb{E}\left[\frac{L_i}{D_{t,i}} \frac{(h_{t,i} + e_{t,i})^2}{D_{t,i}}\right] \\
&\le 2\beta_1^2 \eta^2 \kappa^2 \sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}^2 + e_{t,i}^2}{D_{t,i}}.
\end{aligned}$$

The shift of the update sum uses $m_0 = 0$.

The drift estimate reintroduces tracking error on the right. We make its coefficient small using $\eta\kappa \le \lambda\rho_1$. Uniformly over $\rho_2/\rho_1 \in \mathcal{K}$, choose $c_\delta > 0$ with $\delta_p \ge c_\delta \rho_1$ for small ρ_1. Fix $\tau = 1/16$ and then λ small enough that $C\eta^2\kappa^2/\delta_p^2 \le C\lambda^2/c_\delta^2 \le 1/16$. Substitution into Lemma A.1 gives

$$\begin{aligned}
\sum_{t=1}^T \sum_i \mathbb{E}\frac{e_{t,i}^2}{D_{t,i}} &\le \frac{\tau + C\eta^2\kappa^2/\delta_p^2}{1 - C\eta^2\kappa^2/\delta_p^2} \sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} \\
&\quad + \frac{C}{1 - C\eta^2\kappa^2/\delta_p^2}\left(\rho_1 T \sum_i \varepsilon_i + \frac{\kappa\Delta}{\delta_p}\right) \\
&\le \frac{2}{15}\sum_{t,i} \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + C\left(\rho_1 T \sum_i \varepsilon_i + \frac{\kappa\Delta}{\rho_1}\right) \\
&\le \frac{1}{4}\sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + C\left(\rho_1 T \sum_i \varepsilon_i + \frac{\kappa\Delta}{\rho_1}\right). \qquad (27)
\end{aligned}$$

Weighted descent. Young's inequality $|ab| \le a^2/4 + b^2$ gives $h_{t,i}(h_{t,i} + e_{t,i}) \ge (3/4)h_{t,i}^2 - e_{t,i}^2$. Therefore (27) gives

$$\begin{aligned}
\sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i} m_{t,i}}{D_{t,i}} &= \sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}(h_{t,i} + e_{t,i})}{D_{t,i}} \\
&\ge \frac{3}{4}\sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} - \sum_{t=1}^T \sum_i \mathbb{E}\frac{e_{t,i}^2}{D_{t,i}} \\
&\ge \frac{1}{2}\sum_{t=1}^T \sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} - C\left(\rho_1 T \sum_i \varepsilon_i + \frac{\kappa\Delta}{\rho_1}\right).
\end{aligned}$$

Similarly,

$$\sum_{t=1}^{T}\sum_i L_i\mathbb{E}\frac{m_{t,i}^2}{D_{t,i}^2} \le 2\kappa\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2+e_{t,i}^2}{D_{t,i}} \le \frac{5}{2}\kappa\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + C\kappa\left(\rho_1 T\sum_i \varepsilon_i + \frac{\kappa\Delta}{\rho_1}\right).$$

Applying (3) to the Adam update and summing over t now gives

$$\begin{aligned}\Delta &\ge \sum_{t=1}^{T}\mathbb{E}[f(x_t)-f(x_{t+1})]\\ &\ge \eta\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}m_{t,i}}{D_{t,i}} - \frac{\eta^2}{2}\sum_{t=1}^{T}\sum_i L_i\mathbb{E}\frac{m_{t,i}^2}{D_{t,i}^2}\\ &\ge \eta\left(\frac{1}{2}-\frac{5}{4}\eta\kappa\right)\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} - C\eta(1+\eta\kappa)\left(\rho_1 T\sum_i\varepsilon_i+\frac{\kappa\Delta}{\rho_1}\right).\end{aligned}$$

For $\lambda \le 1/5$, we have $\eta\kappa \le 1/5$, and therefore

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} &\le \frac{4\Delta}{\eta T} + C\rho_1\sum_i\varepsilon_i + \frac{C\kappa\Delta}{\rho_1 T}\\ &\le \frac{(4+C\lambda)\Delta}{\eta T} + C\rho_1\sum_i\varepsilon_i \le C\left(\frac{\Delta}{\eta T}+\rho_1\sum_i\varepsilon_i\right).\end{aligned} \tag{28}$$

Write

$$S = \sum_i \varepsilon_i.$$

By (11) and (25), $S \asymp \Sigma\rho_2^{-\frac{2-p}{2p}}$, $\eta \asymp \rho_1 S/L$, and $\rho_1 \asymp \rho_2 \asymp \rho_\star$. By the definition of $\rho_\star$,

$$\begin{aligned}\rho_1^2 S^2 \asymp \Sigma^2\rho_\star^{\frac{3p-2}{p}} &= \frac{\Delta L}{T},\\ \frac{\Delta}{\eta T} \asymp \frac{\Delta L}{\rho_1 S T} \asymp \frac{\rho_1^2 S^2}{\rho_1 S} &= \rho_1 S,\\ \frac{1}{T}\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} \le C\left(\frac{\Delta}{\eta T}+\rho_1 S\right) &\le C\rho_1 S.\end{aligned} \tag{29}$$

Average denominator. We bound past gradient contributions using the weighted gradient term from (29). The recursion and $\varepsilon_i > 0$ give, for $s \le t$,

$$\begin{aligned}D_{t,i} = \sqrt{v_{t,i}}+\varepsilon_i &\ge \beta_2^{\frac{t-s}{2}}\sqrt{v_{s,i}}+\varepsilon_i \ge \beta_2^{\frac{t-s}{2}} D_{s,i},\\ \rho_2\sum_{s=1}^{t}\beta_2^{t-s}h_{s,i}^2 = \rho_2\sum_{s=1}^{t}\beta_2^{t-s}D_{s,i}\frac{h_{s,i}^2}{D_{s,i}} &\le D_{t,i}\rho_2\sum_{s=1}^{t}\beta_2^{\frac{t-s}{2}}\frac{h_{s,i}^2}{D_{s,i}}.\end{aligned}$$

Using $v_{0,i} = 0$, the weighted Euclidean triangle inequality, and $2\sqrt{ab} \le a + b$, we obtain

$$\begin{aligned} D_{t,i} &= \left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} (h_{s,i} + \xi_{s,i})^2\right)^{1/2} + \varepsilon_i \\ &\le \left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} h_{s,i}^2\right)^{1/2} + \left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} \xi_{s,i}^2\right)^{1/2} + \varepsilon_i \\ &\le \frac{D_{t,i}}{2} + \frac{\rho_2}{2} \sum_{s=1}^{t} \beta_2^{\frac{t-s}{2}} \frac{h_{s,i}^2}{D_{s,i}} + \left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} \xi_{s,i}^2\right)^{1/2} + \varepsilon_i. \end{aligned}$$

Consequently,

$$D_{t,i} \le \rho_2 \sum_{s=1}^{t} \beta_2^{\frac{t-s}{2}} \frac{h_{s,i}^2}{D_{s,i}} + 2\left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} \xi_{s,i}^2\right)^{1/2} + 2\varepsilon_i. \tag{30}$$

Exchanging the time sums gives

$$\begin{aligned} \rho_2 \sum_{t=1}^{T} \sum_{s=1}^{t} \beta_2^{\frac{t-s}{2}} \frac{h_{s,i}^2}{D_{s,i}} &= \rho_2 \sum_{s=1}^{T} \frac{h_{s,i}^2}{D_{s,i}} \sum_{k=0}^{T-s} \beta_2^{k/2} \\ &\le \frac{\rho_2}{1-\sqrt{\beta_2}} \sum_{s=1}^{T} \frac{h_{s,i}^2}{D_{s,i}} \\ &= (1+\sqrt{\beta_2}) \sum_{s=1}^{T} \frac{h_{s,i}^2}{D_{s,i}} \le 2 \sum_{s=1}^{T} \frac{h_{s,i}^2}{D_{s,i}}. \end{aligned} \tag{31}$$

For the noise term, subadditivity of $z \mapsto z^{\frac{p}{2}}$, (5), and Lyapunov's inequality give

$$\begin{aligned} \mathbb{E}\left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} \xi_{s,i}^2\right)^{1/2} &\le \left[\mathbb{E}\left(\rho_2 \sum_{s=1}^{t} \beta_2^{t-s} \xi_{s,i}^2\right)^{p/2}\right]^{1/p} \\ &\le \left(\rho_2^{p/2} \sum_{s=1}^{t} \beta_2^{\frac{(t-s)p}{2}} \mathbb{E}|\xi_{s,i}|^p\right)^{1/p} \\ &\le \left(\frac{\sigma_i^p \rho_2^{p/2}}{1-\beta_2^{p/2}}\right)^{1/p} \le C\sigma_i \rho_2^{-\frac{2-p}{2p}} \le C\varepsilon_i. \end{aligned}$$

Here concavity gives $1 - \beta_2^{p/2} \ge (p/2)(1-\beta_2) = (p/2)\rho_2$.

Combining this noise bound with (30)–(31) proves

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} D_{t,i} \le 2\left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} \frac{h_{t,i}^2}{D_{t,i}}\right) + CS. \tag{32}$$

Recovering stationarity. Write $|h_{t,i}| = (|h_{t,i}|/\sqrt{D_{t,i}})\sqrt{D_{t,i}}$. Cauchy–Schwarz jointly over time, coordinates, and probability, followed by (32) and (29), gives

$$\begin{aligned} R_T^2 &\le \left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} \frac{h_{t,i}^2}{D_{t,i}}\right)\left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} D_{t,i}\right) \\ &\le \left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} \frac{h_{t,i}^2}{D_{t,i}}\right)\left(2\frac{1}{T} \sum_{t=1}^{T} \sum_{i} \mathbb{E} \frac{h_{t,i}^2}{D_{t,i}} + CS\right) \\ &\le C\rho_1 S(\rho_1 S + S) \le C\rho_1 S^2. \end{aligned}$$

Taking square roots gives

$$R_T \le C\sqrt{\rho_1}\, S \le C\Sigma \rho_\star^{\frac{p-1}{p}}$$
$$= C\Sigma \left(\frac{\Delta L}{\Sigma^2 T}\right)^{\frac{p-1}{3p-2}} = C(\Delta L)^{\frac{p-1}{3p-2}} \Sigma^{\frac{p}{3p-2}} T^{-\frac{p-1}{3p-2}} = C R_\star.$$

This proves the theorem. □

B Proof of the lower bound for a shared ε

We give the objective, oracle, and pathwise invariant for Theorem 4.2.

B.1 Parameters and objective

Fix $0 < \delta \le \min\{\delta_0, T^{-\frac{p-1}{p}}\}$ and set

$$\gamma = \frac{1}{128}, \qquad N = \lceil \delta^{-1} \rceil, \qquad A = \frac{\gamma}{N}, \qquad b = \frac{\gamma\sqrt{\delta}}{N}. \tag{33}$$

Let $z = (u, w_{1,+}, w_{1,-}, \dots, w_{N,+}, w_{N,-})$, write

$$s_i = \frac{w_{i,+} + w_{i,-}}{2},$$

and define

$$Y(z) = 1 + \gamma\delta\, u + b \sum_{i=1}^{N} s_i. \tag{34}$$

Let

$$\phi(y) = \begin{cases} 0, & y \le 0, \\ \frac{y^2}{2}, & 0 < y < 1, \\ y - \frac{1}{2}, & y \ge 1, \end{cases} \qquad f(z) = \phi(Y(z)). \tag{35}$$

The function ϕ is convex and continuously differentiable, with $\phi'(y) = \min\{1, \max\{0, y\}\}$. Since $\phi \ge 0$ and $Y(z)$ can take any real value, $\inf f = 0$. At $z_1 = 0$, $Y(z_1) = 1$, so

$$f(z_1) - \inf f = \frac{1}{2}. \tag{36}$$

Let ϱ be the coefficient vector in (34), with $\varrho_0 = \gamma\delta$ and $\varrho_{i,+} = \varrho_{i,-} = b/2$. The chain rule gives $\nabla f(z) = \phi'(Y(z))\varrho$. Whenever $Y(z) \ge 1$,

$$\nabla f(z) = \varrho, \qquad \|\nabla f(z)\|_1 = \gamma\delta + Nb = \gamma\delta + \gamma\sqrt{\delta} \ge \gamma\sqrt{\delta}. \tag{37}$$

B.2 Sample-function construction

Let $\theta = 1/(8T)$. Independently over time, define a random vector q_t as follows. Its control coordinate is

$$q_{t,0} = \begin{cases} -\gamma\delta, & \text{with probability } 1-\theta, \\ \gamma\delta(\frac{2}{\theta} - 1), & \text{with probability } \theta. \end{cases} \tag{38}$$

Draw one sign $\zeta_t \in \{-1, 1\}$ with equal probabilities, independently of the control coordinate, and use it for all pairs:

$$(q_{t,i}^+, q_{t,i}^-) = \begin{cases} (A+b, -A), & \zeta_t = 1, \\ (-A, A+b), & \zeta_t = -1. \end{cases} \tag{39}$$

The two coordinates therefore receive a uniformly random permutation of $\{A+b, -A\}$. Their means satisfy

$$\mathbb{E}q_{t,0} = (1-\theta)(-\gamma\delta) + \theta\gamma\delta\left(\frac{2}{\theta} - 1\right) = \gamma\delta,$$
$$\mathbb{E}q_{t,i}^{\pm} = \frac{(A+b) + (-A)}{2} = \frac{b}{2}, \qquad \text{hence } \mathbb{E}q_t = \varrho. \tag{40}$$

Put $\omega_t = q_t - \varrho$ and define the random sample function

$$F_t(z) = f(z) + \langle \omega_t, z \rangle. \tag{41}$$

Every F_t is convex and continuously differentiable, and its gradient

$$g_t(z) = \phi'(Y(z))\varrho + \omega_t \tag{42}$$

satisfies $\mathbb{E}[g_t(z)] = \nabla f(z)$. The noise ω_t is independent of z and of the past. Appendix E verifies its coordinatewise pth moments.

Define the event

$$\mathcal{A} = \{q_{t,0} = -\gamma\delta \text{ for every } 1 \le t \le T\}. \tag{43}$$

The union bound gives

$$\Pr(\mathcal{A}) = (1-\theta)^T \ge 1 - T\theta = \frac{7}{8}. \tag{44}$$

B.3 Averaging coefficients for a realized history

For the time-varying $\beta_{1,t}, \beta_{2,t}$ in (9), define, for $1 \le s \le t$,

$$\alpha_{t,s} = (1-\beta_{1,s}) \prod_{r=s+1}^{t} \beta_{1,r}, \qquad \nu_{t,s} = (1-\beta_{2,s}) \prod_{r=s+1}^{t} \beta_{2,r}, \tag{45}$$

$$P_t = \sum_{s=1}^{t} \alpha_{t,s}, \qquad V_t = \sum_{s=1}^{t} \nu_{t,s}. \tag{46}$$

These coefficients are nonnegative. Iterating the two moment recursions from $m_0 = v_0 = 0$ gives

$$m_{t,j} = \sum_{s=1}^{t} \alpha_{t,s} g_{s,j}, \qquad v_{t,j} = \sum_{s=1}^{t} \nu_{t,s} g_{s,j}^2.$$

In particular, $\alpha_{t,t} = 1 - \beta_{1,t} > 0$ and $\nu_{t,t} = 1 - \beta_{2,t} > 0$, so $P_t, V_t > 0$.

Suppose $Y(z_s) \ge 1$ for $s \le t$. Then (42) equals q_s at those iterates. On $\mathcal{A}$, the control coordinate therefore satisfies

$$m_{t,0} = -\gamma\delta P_t, \qquad v_{t,0} = \gamma^2\delta^2 V_t. \tag{47}$$

The resulting contribution to Y is

$$\gamma\delta(u_{t+1} - u_t) = \frac{\eta_t \gamma^2 \delta^2 P_t}{\gamma\delta\sqrt{V_t} + \varepsilon_t} = \frac{\gamma^2 \delta \eta_t P_t}{\gamma\sqrt{V_t} + \frac{\varepsilon_t}{\delta}}. \tag{48}$$

B.4 Proof of the paired-update lemma

Proof of Lemma 5.1. Suppress the pair and time indices. Since each observation pair sums to b and its difference has magnitude $2A + b$,

$$m^+ + m^- = bP, \qquad |m^+ - m^-| \le (2A+b)P.$$

Each squared observation lies between A^2 and $(A+b)^2$, so

$$A\sqrt{V} + \varepsilon_t \le D^{\pm} \le (A+b)\sqrt{V} + \varepsilon_t.$$

Since $A, V > 0$, both denominators are positive even when $\varepsilon_t = 0$. If $\ell \le d_2 \le d_1 \le u$ and $\ell > 0$, then $(d_1 - d_2)/(d_1 + d_2) \le (u - \ell)/(u + \ell)$: the ratio increases with d_1 and decreases with d_2. Applying this to the preceding interval,

$$\frac{|(D^+)^{-1} - (D^-)^{-1}|}{(D^+)^{-1} + (D^-)^{-1}} = \frac{|D^+ - D^-|}{D^+ + D^-} \le \frac{b\sqrt{V}}{(2A+b)\sqrt{V} + 2\varepsilon_t} \le \frac{b}{2A+b}. \tag{49}$$

The potentially negative term in the normalized pair sum satisfies

$$\begin{aligned} \left| \frac{m^+ - m^-}{2} \left(\frac{1}{D^+} - \frac{1}{D^-} \right) \right| &\le \frac{(2A+b)P}{2} \frac{b}{2A+b} \left(\frac{1}{D^+} + \frac{1}{D^-} \right) \\ &= \frac{bP}{2} \left(\frac{1}{D^+} + \frac{1}{D^-} \right). \end{aligned}$$

Consequently,

$$\begin{aligned} 0 &\le \frac{m^+}{D^+} + \frac{m^-}{D^-} \\ &= \frac{bP}{2}\left(\frac{1}{D^+} + \frac{1}{D^-}\right) + \frac{m^+ - m^-}{2}\left(\frac{1}{D^+} - \frac{1}{D^-}\right) \\ &\le bP\left(\frac{1}{D^+} + \frac{1}{D^-}\right) \le \frac{2bP}{A\sqrt{V} + \varepsilon_t}. \end{aligned}$$

□

B.5 The iterates remain in the affine region

On $\mathcal{A}$, suppose $Y(z_s) \ge 1$ for $s \le t$. The mean of pair i changes by

$$s_{t+1,i} - s_{t,i} = -\frac{\eta_t}{2}\left(\frac{m^+_{t,i}}{D^+_{t,i}} + \frac{m^-_{t,i}}{D^-_{t,i}}\right).$$

Lemma 5.1 bounds their total contribution by

$$0 \le -b\sum_{i=1}^{N}(s_{t+1,i} - s_{t,i}) \le \frac{Nb^2\eta_t P_t}{A\sqrt{V_t} + \varepsilon_t} = \frac{\gamma^2\delta\eta_t P_t}{\gamma\sqrt{V_t} + N\varepsilon_t}. \tag{50}$$

Combining this with (48) gives

$$\begin{aligned} Y(z_{t+1}) - Y(z_t) &= \gamma\delta(u_{t+1} - u_t) + b\sum_{i=1}^{N}(s_{t+1,i} - s_{t,i}) \\ &\ge \gamma^2\delta\eta_t P_t\left[\frac{1}{\gamma\sqrt{V_t} + \frac{\varepsilon_t}{\delta}} - \frac{1}{\gamma\sqrt{V_t} + N\varepsilon_t}\right] \ge 0. \end{aligned} \tag{51}$$

The bracket is nonnegative because $N \ge \delta^{-1}$.

At $z_1 = 0$, $Y(z_1) = 1$. Suppose inductively that $Y(z_s) \ge 1$ for $s \le t$. The preceding calculation gives $Y(z_{t+1}) \ge Y(z_t) \ge 1$. On $\mathcal{A}$,

$$Y(z_t) \ge 1 \quad \text{and} \quad \|\nabla f(z_t)\|_1 \ge \gamma\sqrt{\delta} \qquad (1 \le t \le T). \tag{52}$$

Using (44),

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(z_t)\|_1 \ge \Pr(\mathcal{A})\gamma\sqrt{\delta} \ge \frac{7}{8}\frac{\sqrt{\delta}}{128}. \tag{53}$$

Appendix E verifies Assumptions 3.1 and 3.3 with $L \le 1$ and $\Sigma \le 1$. Together with (53), this proves Theorem 4.2.

C Balanced scales and one-dimensional optimality

C.1 Shared ε under balanced scales

For Corollary 4.3, (15) and the common choice of ε_i give

$$\varepsilon_i \ge \rho_2^{-\frac{2-p}{2p}}\max\left\{\sigma_i, \frac{L_i}{L}\Sigma\right\}, \qquad \sum_i \varepsilon_i = C_0\rho_2^{-\frac{2-p}{2p}}\Sigma.$$

These are the two bounds on ε_i used in (25). In particular, $\kappa \le (L/\Sigma)\rho_2^{\frac{2-p}{2p}}$ and $\eta\kappa \le \lambda\rho_1$ still hold. Thus the descent estimate and the denominator estimate in Appendix A give

$$\frac{1}{T}\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} \le C\left(\frac{\Delta}{\eta T} + \rho_1 S\right), \qquad S = C_0\Sigma\rho_2^{-\frac{2-p}{2p}},$$

and

$$R_T^2 \le \left(\frac{1}{T}\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}}\right)\left(2\frac{1}{T}\sum_{t=1}^{T}\sum_i \mathbb{E}\frac{h_{t,i}^2}{D_{t,i}} + CS\right).$$

Summing the balance inequalities shows $C_0 \ge 1$. The same prescribed η and $\rho_1 = \rho_\star$ satisfy $\Delta/(\eta T) \lesssim \rho_1 \Sigma \rho_2^{-\frac{2-p}{2p}} \le \rho_1 S$. Consequently,

$$R_T^2 \le C\rho_1 S(\rho_1 S + S) \le C\rho_1 S^2,$$
$$R_T \lesssim \sqrt{\rho_1} S = C_0 \Sigma \rho_1^{1/2} \rho_2^{-\frac{2-p}{2p}} \asymp C_0 \Sigma \rho_\star^{(p-1)/p} = C_0 R_\star,$$

as claimed.

C.2 One-dimensional fixed-parameter Adam

When $d = 1$, shared and coordinatewise choices of ε coincide. We record the corresponding family-specific minimax statement as a supplementary result.

Proposition C.1 (Minimax rate within a fixed-parameter Adam family)**.** *Fix $p \in (1, 2]$ and let $\mathcal{K}$ be a compact subset of $(0, 4/(2-p))$ when $p < 2$, or of $(0, \infty)$ when $p = 2$. There exist $c, C, \bar\rho, \bar\lambda > 0$, depending only on p and $\mathcal{K}$, such that, for each fixed ratio $r = \rho_2/\rho_1 \in \mathcal{K}$, the admissible parameters satisfy*

$$\rho_2 = r\rho_1 \le \tfrac{1}{2}, \qquad 0 < \rho_1 \le \bar\rho, \qquad 0 < \lambda \le \bar\lambda,$$
$$\varepsilon = \Sigma \rho_2^{-\frac{2-p}{2p}}, \qquad \eta = \frac{\lambda\rho_1\Sigma\rho_2^{-\frac{2-p}{2p}}}{L} = \frac{\lambda\rho_1\varepsilon}{L}. \tag{54}$$

Whenever $\rho_\star \le c$ and $T\rho_\star \ge C$,

$$\inf_{\rho_1,\lambda}\sup_{f,O}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|f'(x_t)| \asymp (\Delta L)^{\frac{p-1}{3p-2}}\Sigma^{\frac{p}{3p-2}}T^{-\frac{p-1}{3p-2}}.$$

The infimum is over (ρ_1, λ) satisfying (54), and the supremum is over scalar objective–oracle pairs satisfying Assumptions 3.1–3.3. The comparison is uniform over $r \in \mathcal{K}$.

In one dimension, (11) gives $\varepsilon = \Sigma\rho_2^{-\frac{2-p}{2p}}$. The upper half of Proposition C.1 follows directly from Theorem 4.1.

For the matching lower bound, we use sparse observations and measure each nonzero observation by the total displacement it induces. Small displacements limit travel, while large displacements permit a two-sign comparison with identical denominators.

C.2.1 Lower-bound construction

Fix $r \in \mathcal{K}$ and a pair (ρ_1, λ) satisfying (54). We show that every such pair has worst-case stationarity at least a constant multiple of $R_\star$.

Choose fixed small constants $\alpha, \mu > 0$ and set

$$\theta = \alpha\rho_\star, \qquad A = \mu\Sigma\theta^{-\frac{1}{p}}, \qquad H = \theta A. \tag{55}$$

Since $R_\star = \Sigma\rho_\star^{(p-1)/p}$,

$$H = \theta A = \mu\Sigma\theta^{\frac{p-1}{p}} = \mu\alpha^{\frac{p-1}{p}}R_\star, \qquad T\rho_\star R_\star^2 = T\Sigma^2\rho_\star^{\frac{3p-2}{p}} = \Delta L.$$

Fix α, μ small enough that $\theta \le 1/2$ and $\mu \le 1/2$ under (54) and $\rho_\star \le c$. Define

$$\psi(y) = \operatorname{clip}\left(y, -\frac{1}{2}, \frac{1}{2}\right), \qquad f'(z) = H\psi\left(\frac{Lz}{H}\right). \tag{56}$$

This derivative is nondecreasing and L-Lipschitz, so f is convex and L-smooth: $|f'(z) - f'(z')| \le H|Lz/H - Lz'/H| = L|z - z'|$. Integrating and normalizing $f(0) = 0$ gives

$$f(z) = \begin{cases} Lz^2/2, & |z| \le H/(2L), \\ H|z|/2 - H^2/(8L), & |z| \ge H/(2L). \end{cases}$$

Thus $\inf f = f(0) = 0$. Start from

$$z_1 = \frac{H}{2L} + W, \qquad W = \frac{\Delta}{H}. \tag{57}$$

Since $f'(z) = H/2$ for $z \geq H/(2L)$,

$$\begin{aligned} f(z_1) - \inf f &= \frac{H^2}{8L} + \frac{HW}{2} = \frac{H^2}{8L} + \frac{\Delta}{2} \\ &= \frac{\mu^2 \alpha^{2(p-1)/p}}{8T\rho_\star}\Delta + \frac{\Delta}{2} \leq \Delta. \end{aligned} \tag{58}$$

The gap bound holds for $T\rho_\star \geq C$ and sufficiently small α and μ.

At time t, independently draw $B_t \sim \text{Bernoulli}(\theta)$. Conditional on $B_t = 1$ and the past, draw $\zeta_t \in \{-1, 1\}$ with

$$\Pr(\zeta_t = 1 \mid \mathcal{F}_{t-1}, B_t = 1) = \frac{1 + \psi\left(\frac{Lz_t}{H}\right)}{2}, \qquad g_t = B_t A \zeta_t. \tag{59}$$

Then

$$\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \theta A \psi\left(\frac{Lz_t}{H}\right) = f'(z_t). \tag{60}$$

The denominator process depends only on the variables B_t, because $g_t^2 = B_t A^2$. Since $|\psi| \leq 1/2$, both signs have conditional probability at least $1/4$ whenever $B_t = 1$.

C.2.2 Displacement from one nonzero observation

Starting from $m = v = 0$, give the algorithm one observation A and zeros thereafter. The total absolute displacement of this trajectory is

$$\mathcal{D} = \eta \sum_{k=0}^{\infty} \frac{\rho_1 A \beta_1^k}{\sqrt{\rho_2 A^2 \beta_2^k + \varepsilon}}. \tag{61}$$

Set

$$\chi = \frac{L\mathcal{D}}{H}. \tag{62}$$

Since every denominator in (61) is at least ε,

$$\begin{aligned} \mathcal{D} &\leq \frac{\eta\rho_1 A}{\varepsilon} \sum_{k=0}^{\infty} \beta_1^k = \frac{\eta A}{\varepsilon}, \\ \chi &= \frac{L\mathcal{D}}{H} \leq \frac{L\eta A}{H\varepsilon} = \frac{L\eta}{\theta\varepsilon} = \lambda\frac{\rho_1}{\theta}. \end{aligned} \tag{63}$$

C.2.3 Small-displacement case

Suppose $\chi \leq \chi_0$.

Let $\tau_1, \ldots, \tau_N$ be the times at which $B_t = 1$. Write the first moment as the sum of the signed contributions from these observations. At every time t,

$$\frac{|m_t|}{\sqrt{v_t} + \varepsilon} \leq \sum_{j=1}^{N} \frac{|m_t^{(j)}|}{\sqrt{v_t^{(j)}} + \varepsilon}. \tag{64}$$

Here, for $t \geq \tau_j$, $m_t^{(j)} = \rho_1 A \zeta_{\tau_j} \beta_1^{t-\tau_j}$ and $v_t^{(j)} = \rho_2 A^2 \beta_2^{t-\tau_j}$; both contributions are zero before τ_j. Thus $m_t = \sum_j m_t^{(j)}$ and $v_t = \sum_j v_t^{(j)}$. The displayed inequality follows because $|m_t| \leq \sum_j |m_t^{(j)}|$ and the full second moment is at least each nonnegative contribution $v_t^{(j)}$. Summing over time and extending each individual displacement contribution to an infinite horizon gives the pathwise bound

$$\sum_{t=1}^{T} |z_{t+1} - z_t| \leq N\mathcal{D}. \tag{65}$$

Since $\mathbb{E}N = T\theta$, Markov's inequality gives

$$\Pr(N\mathcal{D} \geq W) \leq \frac{T\theta\mathcal{D}}{W} = \chi\frac{T\theta H^2}{\Delta L} = \chi\frac{T\alpha\rho_\star\mu^2\alpha^{2(p-1)/p}R_\star^2}{T\rho_\star R_\star^2} = \chi\mu^2\alpha^{1+\frac{2(p-1)}{p}}. \tag{66}$$

Choose α, μ and then χ_0 so that (66) is at most $1/4$. On the complementary event, (65) and (57) imply $z_t \geq H/(2L)$ for every t, so $|f'(z_t)| = H/2$. Hence

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|f'(z_t)| \geq \frac{3}{8}H \geq cR_\star. \tag{67}$$

C.2.4 Large-displacement case

Suppose $\chi > \chi_0$.

Put $q = \rho_2/\theta$. Equation (63) and (54) imply

$$q = r\frac{\rho_1}{\theta} \geq \frac{(\inf\mathcal{K})\chi_0}{\bar{\lambda}} =: q_0 > 0. \tag{68}$$

We first show that a block of length $O(1/\theta)$ captures a fixed fraction of (61).

Lemma C.2 (Finite-horizon displacement bound)**.** *Fix $q_0 > 0$. There are constants $B, c > 0$, depending only on $p, \mathcal{K}, q_0$, with the following property. Let $\theta > 0$, $\rho_2/\theta \geq q_0$, and $A = \mu\Sigma\theta^{-\frac{1}{p}}$. Consider a block of length $\ell = \lceil B/\theta \rceil$. Suppose its initial second-moment state satisfies $v \leq 4\theta A^2$, there is no nonzero observation in the first quarter, exactly one observation ζA in the second quarter, and no other nonzero observation. For every time in the last quarter, the positions obtained with $\zeta = 1$ and $\zeta = -1$ differ by at least $2c\mathcal{D}(\theta, A)$, where $\mathcal{D}(\theta, A)$ is defined as in (61). The bound is independent of the initial position and first-moment state.*

Proof. Let

$$\mathcal{D}_n(\theta, A) = \eta\sum_{j=0}^{n-1}\frac{\rho_1 A\beta_1^j}{\sqrt{\rho_2 A^2\beta_2^j + \varepsilon}}, \qquad \mathcal{D}(\theta, A) = \lim_{n\to\infty}\mathcal{D}_n(\theta, A).$$

Writing $q = \rho_2/\theta$ and using $\varepsilon = \Sigma\rho_2^{-\frac{2-p}{2p}}$ gives

$$\frac{\sqrt{\rho_2}A}{\varepsilon} = \mu\rho_2^{\frac{1}{p}}\theta^{-1/p} = \mu\left(\frac{\rho_2}{\theta}\right)^{1/p} = \mu q^{1/p}. \tag{69}$$

Let

$$j_0 = \left\lceil \frac{2}{\rho_2}\log_+\left(2\mu q^{\frac{1}{p}}\right)\right\rceil.$$

Here $\log_+ u = \max\{0, \log u\}$. Since $\beta_2^{j/2} \leq e^{-\rho_2 j/2}$, for $j \geq j_0$ we have

$$\frac{\sqrt{\rho_2}A\beta_2^{j/2}}{\varepsilon} \leq \mu q^{1/p}e^{-\rho_2 j_0/2} \leq \frac{1}{2}.$$

For $j \geq j_0$, the square-root term in the denominator is at most $\varepsilon/2$. Summing only the terms with $j \geq j_0$ gives

$$\mathcal{D}(\theta, A) \geq \frac{2\eta\rho_1 A}{3\varepsilon}\sum_{j=j_0}^{\infty}\beta_1^j = \frac{2\eta A}{3\varepsilon}\beta_1^{j_0} \geq c\frac{\eta A}{\varepsilon}\beta_1^{j_0}. \tag{70}$$

Also,

$$\mathcal{D}(\theta, A) - \mathcal{D}_n(\theta, A) \leq \frac{\eta\rho_1 A}{\varepsilon}\sum_{j=n}^{\infty}\beta_1^j = \frac{\eta A}{\varepsilon}\beta_1^n. \tag{71}$$

Hence $\mathcal{D}_n(\theta, A) \geq \mathcal{D}(\theta, A)/2$ whenever $n \geq j_0 + C/\rho_1$. Indeed,

$$\frac{\mathcal{D} - \mathcal{D}_n}{\mathcal{D}} \leq \frac{3}{2}\beta_1^{n-j_0} \leq \frac{3}{2}e^{-\rho_1(n-j_0)} \leq \frac{1}{2}$$

as soon as $n - j_0 \geq (\log 3)/\rho_1$.

Both terms in this lower bound for n are $O(1/\theta)$:

$$j_0 \leq 1 + \frac{C(1 + \log_+ q)}{q\theta} \leq \frac{C}{\theta}, \qquad \frac{1}{\rho_1} = \frac{r}{q\theta} \leq \frac{\sup \mathcal{K}}{q_0 \theta}.$$

The function $(1+\log_+ q)/q$ is bounded on $[q_0, \infty)$, and $\theta = \rho_2/q \leq 1/(2q_0)$ absorbs the constant 1 in the bound for j_0. Partition each block into four consecutive sets with endpoints $\lfloor j\ell/4 \rfloor$, $j = 0, \ldots, 4$. Choose B large enough that a nonzero observation in the second quarter precedes every time in the last quarter by at least this many iterations. Such a choice is possible since $t - s \geq \ell/4 - 1$ for these two quarters.

Now account for the inherited second moment. If the observation occurs at time s, then for $j \geq 0$ the two sign trajectories have the same denominator and

$$m_{s+j}^{(\zeta)} = \beta_1^{j+1} m_{s-1} + \rho_1 A \beta_1^j \zeta,$$
$$v_{s+j} = \beta_2^j (\beta_2 v_{s-1} + \rho_2 A^2).$$

Since $v_{s-1} \leq 4\theta A^2$ and $\rho_2/\theta \geq q_0$,

$$\beta_2 v_{s-1} + \rho_2 A^2 \leq \left(1 + \frac{4}{q_0}\right) \rho_2 A^2.$$

The inherited numerator is identical for the two signs and cancels when their positions are subtracted. The inherited second moment satisfies

$$\sqrt{v_{s+j}} + \varepsilon \leq \sqrt{1 + 4/q_0} \left(\sqrt{\rho_2 A^2 \beta_2^j} + \varepsilon\right).$$

Since the trajectories agree at time s, their separation is therefore

$$\begin{aligned} |z_t^{(+)} - z_t^{(-)}| &= 2\eta\rho_1 A \sum_{j=0}^{t-s-1} \frac{\beta_1^j}{\sqrt{v_{s+j}} + \varepsilon} \\ &\geq \frac{2\eta\rho_1 A}{\sqrt{1 + 4/q_0}} \sum_{j=0}^{t-s-1} \frac{\beta_1^j}{\sqrt{\rho_2 A^2 \beta_2^j} + \varepsilon} \\ &= \frac{2}{\sqrt{1 + 4/q_0}} \mathcal{D}_{t-s}(\theta, A) \geq \frac{1}{\sqrt{1 + 4/q_0}} \mathcal{D}(\theta, A). \end{aligned}$$

The truncation estimate applies because $t - s$ exceeds the number of iterations required to capture half of $\mathcal{D}$. This proves the lemma. □

Partition the horizon into complete blocks of length $\ell = \lceil B/\theta \rceil$. Since

$$\mathbb{E}v_t = \beta_2 \mathbb{E}v_{t-1} + \rho_2 \theta A^2, \qquad \mathbb{E}v_t = \rho_2 \theta A^2 \sum_{j=0}^{t-1} \beta_2^j = \theta A^2 (1 - \beta_2^t) \leq \theta A^2,$$

Markov's inequality shows that a block starts with $v \leq 4\theta A^2$ with probability at least $3/4$: for a block starting at step t_0, $\Pr(v_{t_0-1} > 4\theta A^2) \leq \mathbb{E}v_{t_0-1}/(4\theta A^2) \leq 1/4$. The future Bernoulli variables are independent of this event. Let I_2 denote the second quarter of the block. Increase B so that $\ell \geq 8$. Then $|I_2| \geq \ell/8$, $B \leq \ell\theta \leq B + 1$, and $\theta \leq 1/2$. The probability of exactly one nonzero observation in the second quarter and none elsewhere in the block is bounded by

$$\begin{aligned} |I_2|\theta(1-\theta)^{\ell-1} &\geq \frac{\ell\theta}{8} e^{-2\theta(\ell-1)} \\ &\geq \frac{B}{8} e^{-2(B+1)} \geq c > 0. \end{aligned} \tag{72}$$

The first inequality uses $\log(1 - \theta) \geq -2\theta$. Thus the joint event consisting of a small initial second moment and the prescribed observation pattern has probability at least $c > 0$.

Fix the history before a block with $v \le 4\theta A^2$, and condition on its Bernoulli pattern and on the time s of the unique observation. Before s the two sign trajectories coincide, and after s all oracle observations in this block are zero. The remaining randomness relevant to their positions is the sign ζ at time s. For every t in the last quarter, Lemma C.2 gives

$$\frac{Lz_t^{(\zeta)}}{H} = C_t - \zeta a_{s,t}, \qquad a_{s,t} \ge c\frac{L\mathcal{D}}{H} \ge c\chi_0, \tag{73}$$

where C_t is the midpoint of the two scaled positions and is fixed under this conditioning. For every $C \in \mathbb{R}$ and $a \ge 0$,

$$|\psi(C-a)| + |\psi(C+a)| \ge \min\left\{2a, \frac{1}{2}\right\}. \tag{74}$$

To verify this inequality, if either $|C-a|$ or $|C+a|$ is at least $1/2$, its clipped magnitude is $1/2$. Otherwise both terms are unclipped, and $|C-a| + |C+a| \ge |(C+a) - (C-a)| = 2a$. At time s, both signs have conditional probability at least $1/4$. The future Bernoulli variables are independent of this sign, so conditioning on the rest of the pattern preserves that probability bound. Writing $\mathbb{E}_\zeta$ for expectation over these two signs, we therefore have

$$\begin{aligned}\mathbb{E}_\zeta|f'(z_t)| &\ge \frac{H}{4}\big(|\psi(C_t - a_{s,t})| + |\psi(C_t + a_{s,t})|\big) \\ &\ge \frac{H}{4}\min\{2c\chi_0, 1/2\}.\end{aligned}$$

Integrating over the histories and patterns, whose joint probability is at least $c > 0$, yields $\mathbb{E}|f'(z_t)| \ge cH$ at each time in the last quarter of a complete block.

There are $M = \lfloor T/\ell \rfloor$ complete blocks, each with at least $\ell/8$ such times. The assumption $T\theta \ge C$ ensures $T \ge 2\ell$ after increasing C, since $\ell \le (B+1)/\theta$. Hence $M \ge T/(2\ell)$, and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|f'(z_t)| \ge \frac{M\ell}{8T}cH \ge \frac{cH}{16} \ge cR_\star. \tag{75}$$

C.2.5 Sample functions and the moment bound

To construct sample functions for (59), draw $U_t \sim \mathrm{Unif}[-1,1]$ independently and define, for every query point z,

$$\zeta_t(z) = \mathrm{sign}(\psi\left(\frac{Lz}{H}\right) - U_t), \qquad g_t(z) = B_t A\zeta_t(z). \tag{76}$$

The function $z \mapsto g_t(z)$ is nondecreasing. Hence $F_t(z) = \int_0^z g_t(y)\,dy$ is convex and has g_t as a subgradient everywhere (and as its derivative except at finitely many points). Also, $\Pr(U_t < \psi(Lz/H)) = (1+\psi(Lz/H))/2$, and therefore $\mathbb{E}_U\zeta_t(z) = \psi(Lz/H)$. Independence of B_t and U_t proves (60). Finally, $\mathbb{E}[|g_t|^p \mid \mathcal{F}_{t-1}] = \theta A^p$ and $|f'(z_t)| \le H/2 \le H$. Thus

$$\begin{aligned}\mathbb{E}[|g_t - f'(z_t)|^p \mid \mathcal{F}_{t-1}] &\le 2^{p-1}\left(\theta A^p + H^p\right) &(77)\\ &\le 2^p\theta A^p = 2^p\mu^p\Sigma^p \le \Sigma^p &(78)\end{aligned}$$

after choosing $\mu \le 1/2$. Equations (67) and (75) cover all admissible (ρ_1, λ) and prove Proposition C.1.

D Proof of the coordinatewise tracking estimate

We prove Lemma A.1 by expanding the momentum. The noise decomposition controls small observations through their truncated second moments. For large observations, we retain the random denominator and sum over subsequent iterations using the recursions for m_t and v_t. Fix a coordinate and suppress its index. Write $h_t = \partial_i f(x_t)$, $D_t = \sqrt{v_t} + \varepsilon$, and $h_0 = h_1$. The condition on ε is

$$\sigma^p \le \varepsilon^p \rho_2^{\frac{2-p}{2}}. \tag{79}$$

All estimates below are uniform over $\rho_2/\rho_1 \in \mathcal{K}$ for sufficiently small ρ_1, so $\delta_p \asymp \rho_1$. Put $r = \rho_2/\rho_1$. Taylor expansion uniformly for $r \in \mathcal{K}$ gives

$$\frac{\beta_1^2}{\beta_2^{(2-p)/2}} = (1 - 2\rho_1 + \rho_1^2)\left(1 + \frac{(2-p)r}{2}\rho_1 + O(\rho_1^2)\right)$$
$$= 1 - \left(2 - \frac{(2-p)r}{2}\right)\rho_1 + O(\rho_1^2).$$

The coefficient $2 - (2-p)r/2$ has a strictly positive minimum on $\mathcal{K}$; at $p = 2$ it equals 2. Thus there are constants $c_\delta, C_\delta > 0$ such that $c_\delta\rho_1 \le \delta_p \le C_\delta\rho_1$ for sufficiently small ρ_1.

D.1 The accumulated noise

Let

$$n_t = \rho_1 \sum_{s=1}^{t} \beta_1^{t-s}\xi_s.$$

We show that

$$\sum_{t=1}^{T} \mathbb{E}\frac{n_t^2}{D_t} \le C\rho_1 T\varepsilon + C\rho_1\rho_2 \sum_{t=1}^{T} \mathbb{E}\frac{h_t^2}{D_t}. \tag{80}$$

At $p = 2$, conditional centering makes distinct noise terms orthogonal. Indeed, for $s < j$, $\mathbb{E}(\xi_s\xi_j) = \mathbb{E}[\xi_s\mathbb{E}(\xi_j \mid \mathcal{F}_{j-1})] = 0$. We first remove the random denominator using $D_t \ge \varepsilon$, and then apply this orthogonality. Since $\sigma \le \varepsilon$,

$$\mathbb{E}\frac{n_t^2}{D_t} \le \frac{\rho_1^2}{\varepsilon}\sum_{s=1}^{t} \beta_1^{2(t-s)}\mathbb{E}\xi_s^2 \le \frac{\rho_1^2\varepsilon}{1-\beta_1^2} \le \rho_1\varepsilon.$$

This proves (80) at the endpoint. Assume $1 < p < 2$ for the rest of this subsection.

Small and large noise. Set $B = \varepsilon/\sqrt{\rho_2}$ and define

$$a_t = \xi_t\mathbf{1}_{\{|\xi_t|\le B\}} - \mathbb{E}[\xi_t\mathbf{1}_{\{|\xi_t|\le B\}} \mid \mathcal{F}_{t-1}], \qquad b_t = \xi_t - a_t.$$

To see both the centering and the moment bounds, write

$$\mu_t = \mathbb{E}[\xi_t\mathbf{1}_{\{|\xi_t|\le B\}} \mid \mathcal{F}_{t-1}] = -\mathbb{E}[\xi_t\mathbf{1}_{\{|\xi_t|> B\}} \mid \mathcal{F}_{t-1}].$$

Then $b_t = \xi_t\mathbf{1}_{\{|\xi_t|>B\}} + \mu_t$, and both a_t and b_t have conditional mean zero. Their moments satisfy

$$\mathbb{E}[a_t^2 \mid \mathcal{F}_{t-1}] \le \mathbb{E}[\xi_t^2\mathbf{1}_{\{|\xi_t|\le B\}} \mid \mathcal{F}_{t-1}] \le B^{2-p}\sigma^p \le \varepsilon^2,$$
$$\mathbb{E}[|b_t| \mid \mathcal{F}_{t-1}] \le 2\mathbb{E}[|\xi_t|\mathbf{1}_{\{|\xi_t|>B\}} \mid \mathcal{F}_{t-1}] \le 2B^{1-p}\sigma^p \le 2\varepsilon\sqrt{\rho_2}. \tag{81}$$

For the small-noise sum, remove the denominator before using orthogonality:

$$\mathbb{E}\frac{\left(\rho_1\sum_{s\le t}\beta_1^{t-s}a_s\right)^2}{D_t} \le \frac{\rho_1^2}{\varepsilon}\sum_{s\le t}\beta_1^{2(t-s)}\mathbb{E}a_s^2$$
$$\le \frac{\rho_1^2\varepsilon}{1-\beta_1^2} \le \rho_1\varepsilon.$$

For $s < j$, b_s is $\mathcal{F}_{j-1}$-measurable, so

$$\mathbb{E}|b_sb_j| = \mathbb{E}\big[|b_s|\mathbb{E}(|b_j| \mid \mathcal{F}_{j-1})\big] \le 2\varepsilon\sqrt{\rho_2}\,\mathbb{E}|b_s| \le 4\varepsilon^2\rho_2.$$

Using $D_t \ge \varepsilon$, the cross terms in the large-noise square satisfy

$$2\rho_1^2\sum_{s<j\le t}\beta_1^{2t-s-j}\mathbb{E}\frac{|b_sb_j|}{D_t} \le 4\varepsilon\rho_2\left(2\rho_1^2\sum_{s<j\le t}\beta_1^{t-s}\beta_1^{t-j}\right)$$
$$\le 4\varepsilon\rho_2\left(\rho_1\sum_{s\le t}\beta_1^{t-s}\right)^2 \le 4\varepsilon\rho_2. \tag{82}$$

A single diagonal term. Put $\bar{D}_t = \sqrt{\beta_2 v_{t-1}} + \varepsilon$ and $J_t = \mathbb{E}[h_t^2/D_t \mid \mathcal{F}_{t-1}]$. First we bound a single squared noise term by

$$\mathbb{E}\left[\frac{\xi_t^2}{\varepsilon^{p-1} D_t^{2-p}} \,\middle|\, \mathcal{F}_{t-1}\right] \le C\varepsilon \left[1 + \left(\frac{\sqrt{\rho_2}|h_t|}{\bar{D}_t}\right)^{2-p}\right]. \tag{83}$$

Indeed, on $|\xi_t| > 2|h_t|$, the bound $D_t \ge \sqrt{\rho_2}|\xi_t|/2$ gives

$$\frac{\xi_t^2}{\varepsilon^{p-1} D_t^{2-p}} \le \frac{C|\xi_t|^p}{\varepsilon^{p-1} \rho_2^{\frac{2-p}{2}}}.$$

On the complement, use $|\xi_t|^2 \le (2|h_t|)^{2-p}|\xi_t|^p$ and $D_t \ge \bar{D}_t$. Both h_t and $\bar{D}_t$ are $\mathcal{F}_{t-1}$-measurable. Thus the conditional contributions of the two events are bounded, respectively, by

$$\frac{C\sigma^p}{\varepsilon^{p-1}\rho_2^{(2-p)/2}} \le C\varepsilon, \qquad \frac{(2|h_t|)^{2-p}\sigma^p}{\varepsilon^{p-1}\bar{D}_t^{2-p}} \le C\varepsilon \left(\frac{\sqrt{\rho_2}|h_t|}{\bar{D}_t}\right)^{2-p}.$$

Adding these bounds proves (83).

To control its gradient-dependent term, use $D_t \le \bar{D}_t + \sqrt{\rho_2}|h_t + \xi_t|$ together with the conditional first-moment bound $\mathbb{E}[|\xi_t| \mid \mathcal{F}_{t-1}] \le \sigma$ to obtain

$$\begin{aligned}\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] &\le \bar{D}_t + \sqrt{\rho_2}(|h_t| + \sigma) \\ &\le 2\bar{D}_t + \sqrt{\rho_2}|h_t|.\end{aligned}$$

Here the condition on ε implies $\sqrt{\rho_2}\sigma \le \varepsilon \rho_2^{1/p} \le \varepsilon \le \bar{D}_t$. Conditional Jensen and the preceding denominator bound give

$$\begin{aligned}J_t = h_t^2 \mathbb{E}[D_t^{-1} \mid \mathcal{F}_{t-1}] &\ge \frac{h_t^2}{\mathbb{E}[D_t \mid \mathcal{F}_{t-1}]} \\ &\ge \frac{h_t^2}{2\bar{D}_t + \sqrt{\rho_2}|h_t|} = \frac{\bar{D}_t}{\rho_2} \frac{(\sqrt{\rho_2}|h_t|/\bar{D}_t)^2}{2 + \sqrt{\rho_2}|h_t|/\bar{D}_t} \\ &\ge \frac{\bar{D}_t}{2\rho_2} \frac{(\sqrt{\rho_2}|h_t|/\bar{D}_t)^2}{1 + \sqrt{\rho_2}|h_t|/\bar{D}_t}.\end{aligned} \tag{84}$$

For $1 < p < 2$ and every $x \ge 0$,

$$x^{2-p} \le 1 + \frac{2x^2}{1+x} :$$

the constant term suffices for $x \le 1$, and for $x \ge 1$ use $x^{2-p} \le x \le 2x^2/(1+x)$. Applying this inequality with $x = \sqrt{\rho_2}|h_t|/\bar{D}_t$ and using $\varepsilon \le \bar{D}_t$ gives

$$\varepsilon \left(\frac{\sqrt{\rho_2}|h_t|}{\bar{D}_t}\right)^{2-p} \le \varepsilon + \frac{2\varepsilon(\sqrt{\rho_2}|h_t|/\bar{D}_t)^2}{1 + \sqrt{\rho_2}|h_t|/\bar{D}_t} \le \varepsilon + \frac{2\bar{D}_t(\sqrt{\rho_2}|h_t|/\bar{D}_t)^2}{1 + \sqrt{\rho_2}|h_t|/\bar{D}_t} \le \varepsilon + 4\rho_2 J_t.$$

In the decomposition of b_t, conditional centering gives $|\mu_t| \le B^{1-p}\sigma^p \le \varepsilon\sqrt{\rho_2}$. The centering correction contributes at most

$$\frac{\mu_t^2}{\varepsilon^{p-1} D_t^{2-p}} \le \frac{\mu_t^2}{\varepsilon} \le \varepsilon\rho_2.$$

Combining $b_t^2 \le 2\xi_t^2 + 2\mu_t^2$ with (83) and the preceding bounds gives

$$\begin{aligned}\mathbb{E}\left[\frac{b_t^2}{\varepsilon^{p-1} D_t^{2-p}} \,\middle|\, \mathcal{F}_{t-1}\right] &\le 2\mathbb{E}\left[\frac{\xi_t^2}{\varepsilon^{p-1} D_t^{2-p}} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2\mu_t^2}{\varepsilon} \\ &\le C\varepsilon \left[1 + \left(\frac{\sqrt{\rho_2}|h_t|}{\bar{D}_t}\right)^{2-p}\right] + 2\varepsilon\rho_2 \le C\varepsilon + C\rho_2 J_t.\end{aligned} \tag{85}$$

Summing over time. For $t \geq s$, the second-moment recursion gives $D_t \geq \beta_2^{\frac{t-s}{2}} D_s$. Together with $D_t \geq \varepsilon$, this yields

$$\frac{b_s^2}{D_t} = \frac{b_s^2}{D_t^{p-1} D_t^{2-p}} \leq \frac{b_s^2}{\varepsilon^{p-1} D_t^{2-p}} \leq \beta_2^{-\frac{(t-s)(2-p)}{2}} \frac{b_s^2}{\varepsilon^{p-1} D_s^{2-p}}.$$

Using $\beta_1^2/\beta_2^{(2-p)/2} = 1 - \delta_p$ and $\mathbb{E}J_s = \mathbb{E}[h_s^2/D_s]$, summation gives

$$\rho_1^2 \sum_{t=1}^{T} \sum_{s \leq t} \beta_1^{2(t-s)} \mathbb{E}\frac{b_s^2}{D_t} \leq \rho_1^2 \sum_{s=1}^{T} \sum_{k=0}^{T-s} \left(\frac{\beta_1^2}{\beta_2^{\frac{2-p}{2}}} \right)^k \mathbb{E}\frac{b_s^2}{\varepsilon^{p-1} D_s^{2-p}} \leq \frac{C\rho_1^2}{\delta_p} \left(T\varepsilon + \rho_2 \sum_{s=1}^{T} \mathbb{E}\frac{h_s^2}{D_s} \right).$$

To combine the estimates, write n_t as its small- and large-noise sums and use $(x+y)^2 \leq 2x^2 + 2y^2$. Within the large-noise square, bound the cross terms in absolute value by (82). This gives

$$\sum_{t=1}^{T} \mathbb{E}\frac{n_t^2}{D_t} \leq C(\rho_1 + \rho_2)T\varepsilon + \frac{C\rho_1^2}{\delta_p} \left(T\varepsilon + \rho_2 \sum_{s=1}^{T} \mathbb{E}\frac{h_s^2}{D_s} \right) \leq C\rho_1 T\varepsilon + C\rho_1\rho_2 \sum_{s=1}^{T} \mathbb{E}\frac{h_s^2}{D_s}.$$

This proves (80).

D.2 Initialization and gradient drift

Expanding (23) gives

$$e_t = \beta_1^t e_0 + \sum_{s=1}^{t} \beta_1^{t-s+1}(h_{s-1} - h_s) + n_t.$$

The initial contribution obeys

$$\sum_{t=1}^{T} \mathbb{E}\frac{\beta_1^{2t} e_0^2}{D_t} \leq \frac{e_0^2}{\varepsilon} \sum_{t=1}^{T} \beta_1^{2t} \leq \frac{e_0^2}{\varepsilon} \frac{\beta_1^2}{1-\beta_1^2} \leq \frac{e_0^2}{\varepsilon\rho_1}.$$

For the drift, put $d_s = h_{s-1} - h_s$. Using $D_t \geq \varepsilon$ and weighted Cauchy–Schwarz gives

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\frac{\left(\sum_{s \leq t} \beta_1^{t-s+1} d_s\right)^2}{D_t} &\leq \frac{\beta_1^2}{\varepsilon} \sum_{t=1}^{T} \left(\sum_{s \leq t} \beta_1^{t-s} \right) \left(\sum_{s \leq t} \beta_1^{t-s} \mathbb{E} d_s^2 \right) \\
&\leq \frac{\beta_1^2}{\varepsilon\rho_1} \sum_{t=1}^{T} \sum_{s \leq t} \beta_1^{t-s} \mathbb{E} d_s^2 \\
&= \frac{\beta_1^2}{\varepsilon\rho_1} \sum_{s=1}^{T} \mathbb{E} d_s^2 \sum_{t=s}^{T} \beta_1^{t-s} \\
&\leq \frac{\beta_1^2}{\varepsilon\rho_1^2} \sum_{s=1}^{T} \mathbb{E}(h_{s-1} - h_s)^2.
\end{aligned}$$

Apply $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$ to the three terms in e_t. The preceding estimates and (80) yield

$$\sum_{t=1}^{T} \mathbb{E}\frac{e_t^2}{D_t} \leq C\rho_1\rho_2 \sum_{t=1}^{T} \mathbb{E}\frac{h_t^2}{D_t} + C\rho_1 T\varepsilon + \frac{Ce_0^2}{\varepsilon\rho_1} + \frac{C\beta_1^2}{\varepsilon\rho_1^2} \sum_{t=1}^{T} \mathbb{E}(h_{t-1} - h_t)^2.$$

Choose ρ_1 small enough that $C\rho_1\rho_2 \leq \tau$. This is possible uniformly over $\rho_2/\rho_1 \in \mathcal{K}$. Using $\delta_p \asymp \rho_1$ and summing over coordinates proves (26).

E VERIFICATION OF THE LOWER-BOUND ASSUMPTIONS

We verify the two quantitative assumptions in Theorem 4.2. Equation (36) gives the initial-gap bound.

E.1 COORDINATEWISE SMOOTHNESS

Set

$$L_j^\star = \|\varrho\|_1 \varrho_j. \tag{86}$$

All coordinates of ϱ are positive. Substituting the definition of $L_j^\star$ gives

$$\sum_j \frac{\varrho_j^2}{L_j^\star} = \frac{1}{\|\varrho\|_1}\sum_j \varrho_j = 1, \qquad \sum_j L_j^\star = \|\varrho\|_1 \sum_j \varrho_j = \|\varrho\|_1^2. \tag{87}$$

Because ϕ' in (35) is 1-Lipschitz,

$$\begin{aligned}
\|\nabla f(z') - \nabla f(z)\|_{(D_L^\star)^{-1}} &= |\phi'(Y(z')) - \phi'(Y(z))| \|\varrho\|_{(D_L^\star)^{-1}} && (88)\\
&\le |\langle \varrho, z' - z\rangle| && (89)\\
&\le \|\varrho\|_{(D_L^\star)^{-1}} \|z' - z\|_{D_L^\star} && (90)\\
&= \|z' - z\|_{D_L^\star}. && (91)
\end{aligned}$$

If $\delta_0 \le 1$, then $\|\varrho\|_1 = \gamma(\delta + \sqrt{\delta}) \le 2\gamma = 1/64$. Thus $L^\star$ satisfies Assumption 3.1 and $\|L^\star\|_1 \le 1$.

E.2 COORDINATEWISE MOMENT BOUNDS

The oracle noise is $\omega_t = q_t - \varrho$. In the control coordinate it equals $-2\gamma\delta$ on the common outcome and $2\gamma\delta(1/\theta - 1)$ on the rare outcome. Therefore

$$\begin{aligned}
\mathbb{E}|\omega_{t,0}|^p &= (2\gamma\delta)^p \big[(1-\theta) + \theta(\frac{1}{\theta} - 1)^p\big]\\
&= (2\gamma\delta)^p \big[(1-\theta) + \theta^{1-p}(1-\theta)^p\big]\\
&\le (2\gamma\delta)^p (1 + \theta^{1-p}). && (92)
\end{aligned}$$

Choosing $\sigma_0 = (\mathbb{E}|\omega_{t,0}|^p)^{1/p}$, we obtain

$$\begin{aligned}
\sigma_0 &\le 2\gamma\delta(1 + \theta^{1-p})^{1/p} \le 2\gamma\delta(1 + \theta^{-\frac{p-1}{p}})\\
&= \frac{\delta}{64}\left(1 + (8T)^{\frac{p-1}{p}}\right) && (93)\\
&\le \frac{1 + 8^{(p-1)/p}}{64} \le \frac{9}{64}. && (94)
\end{aligned}$$

Here $\delta \le 1$ and $\delta T^{(p-1)/p} \le 1$ by the choice of δ.

In pair i, each coordinate has mean $b/2$. Its two centered outcomes are $(A+b) - b/2 = A + b/2$ and $-A - b/2 = -(A + b/2)$, each with probability $1/2$. Thus its pth absolute moment is exactly $(A + b/2)^p$. We may take $\sigma_{i,+} = \sigma_{i,-} = A + b/2$. From (33),

$$\begin{aligned}
\sum_{i=1}^N (\sigma_{i,+} + \sigma_{i,-}) &= 2NA + Nb && (95)\\
&= \gamma(2 + \sqrt{\delta}) \le \frac{3}{128}. && (96)
\end{aligned}$$

The sampling at time t is independent of $\mathcal{F}_{t-1}$, so these unconditional moment calculations also hold conditionally, as required by Assumption 3.3. Combining (94) and (96) gives

$$\Sigma = \sigma_0 + \sum_{i=1}^N (\sigma_{i,+} + \sigma_{i,-}) \le \frac{9}{64} + \frac{3}{128} = \frac{21}{128} < 1. \tag{97}$$

This completes the proof of Theorem 4.2.