THE PRECISION COMPLEXITY OF DITHERED ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-precision Adam can reduce the memory used by optimizer states. We study the cumulative precision needed to reach a target accuracy when write widths can change over time and respond to observed numerical errors. We first consider bias-corrected Adam on the scalar quadratic objective, with fresh uniform additive dither in the parameter, gradient, and two moment writes. We also extend the complexity bounds to scalar strongly convex smooth objectives and higher-dimensional separable objectives. A causal policy chooses each width using the revealed history, and we measure cost by the sum of all widths used during the run. Under an explicit step-size condition, we establish matching lower and upper bounds of order $T + \log^2(1/\epsilon)$ on the expected cumulative precision required to achieve $\mathbb{E}[x_T^2] \le \epsilon$. These bounds hold for every sufficiently small ϵ and every deterministic horizon T above a threshold of order $\log(1/\epsilon)$. The lower bound applies to all causal policies using the full error history, while a deterministic schedule attains the upper bound by keeping second-moment precision fixed and raising the other three widths near the end of the run. For the same target accuracy and horizon, keeping all four write widths fixed throughout the run gives a minimum cost of $\Theta(T \log(1/\epsilon))$. Thus, for $T \ge \log^2(1/\epsilon)$, temporal precision allocation saves a factor of order $\log(1/\epsilon)$. If the deterministic horizon can also be chosen, a stronger step-size condition gives an optimal cost of $\Theta(\log^2(1/\epsilon))$. Without this additional condition, suitable parameters and initialization can permit finite-time cancellation, reducing the cost to $\Theta(\log(1/\epsilon))$.

1 INTRODUCTION

Lower numerical precision can reduce the memory and arithmetic costs of training. Early fixed-point work used stochastic rounding, while later mixed-precision methods used half-precision arithmetic for much of training and kept higher precision where needed (Gupta et al., 2015; Micikevicius et al., 2018). Adam adds persistent optimizer state: it stores moving averages of gradients and squared gradients, known as the first and second moments (Kingma & Ba, 2015). Blockwise 8-bit and 4-bit optimizers store these moments in substantially smaller formats (Dettmers et al., 2022; Li et al., 2023).

These advances leave open how much precision the full Adam recursion needs when write widths may change during a run. Existing studies of low-precision optimizers often use fixed state formats or give sufficient conditions for convergence (Dettmers et al., 2022; Li et al., 2023; Tang et al., 2026). Other methods vary precision across iterations, but they target different numerical resources or limit which widths can change (Yan et al., 2021; Ding et al., 2024). None of these results characterizes the minimum cumulative precision cost of all four Adam writes when later choices can use observed errors. We measure cumulative precision by the expected sum of all write widths across the run.

An error's effect depends on when it occurs and which Adam state it perturbs. An early error may fade before we measure terminal accuracy, whereas a second-moment error changes the denominator and later effective step sizes. A policy can use observed errors to choose later widths, even though each fresh error has conditional mean zero. The same noisy gradient also enters both moment updates, coupling the two moment states, as Figure 1(b) illustrates. To prove a matching lower bound, we must account for this dependence and the policy's response to earlier errors.

We study these effects in bias-corrected Adam on the scalar quadratic $f(x) = x^2/2$, without stochastic-gradient noise. At each iteration, fresh uniform additive dither perturbs the gradient,

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.

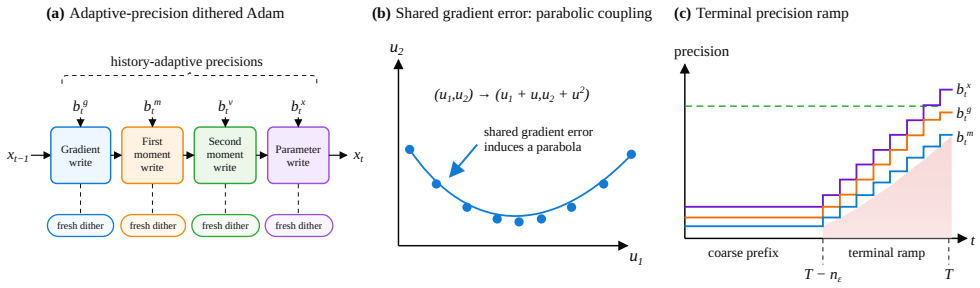

Figure 1: **Precision complexity of dithered Adam.** (a) Four sequential writes each use a width chosen from the revealed history and receive fresh dither. (b) At fixed x, the same gradient-write error u affects both moments, producing the displacement (u, u^2) in transformed moment coordinates. (c) The schedule holds b_t^v fixed and raises the other three widths over the final $n_\varepsilon = \Theta(\log(1/\varepsilon))$ iterations. The shaded area marks the $\Theta(\log^2(1/\varepsilon))$ excess cost. The panels are schematic under the conditions of Theorem 1.

first-moment, second-moment, and parameter writes. All arithmetic between these writes is exact. For a write of type $s \in \{g, m, v, x\}$, the error radius is $r_s 2^{-b_t^s}$, with fixed $r_s > 0$. The policy selects each integer width $b_t^s \geq 1$ before the corresponding error is drawn, using all previously revealed outcomes, including those of earlier writes in the same iteration. The optimizer parameters and error scales are fixed independently of the target accuracy and horizon. We do not assume that the iterates remain in a prescribed bounded region.

Central question. For a prescribed horizon T and target accuracy ϵ, what is the minimum expected cumulative precision cost needed to achieve $\mathbb{E}[x_T^2] \leq \epsilon$ when the policy may choose each width using all outcomes revealed before that write?

Our contributions are as follows.

1. **A matching temporal precision schedule.** We construct a deterministic schedule that uses low precision through most of the run and raises the parameter, gradient, and first-moment widths over the final $O(\log(1/\varepsilon))$ iterations. The second-moment width remains constant throughout the run. A uniform contraction bound shows that this schedule achieves $\mathbb{E}[x_T^2] \leq \varepsilon$ with total cost $O(T + \log^2(1/\varepsilon))$.
2. **A lower bound for all causal adaptive policies.** We prove that no causal policy can improve this order, even when every width may depend on all previously revealed write errors. The proof bounds the probability that the final parameter lies near the minimizer in terms of the expected precision used at earlier iterations. This bound holds uniformly over subsequent adaptive width choices and yields a lower bound on the cumulative precision required to reach the target accuracy. Thus, our deterministic schedule achieves the optimal order of expected cumulative precision within this policy class.
3. **A comparison with fixed write widths.** We show that the minimum cost with all four write widths fixed over time is $\Theta(T \log(1/\varepsilon))$ for the same target accuracy and prescribed horizon. When $T \geq \log^2(1/\varepsilon)$, temporal allocation therefore saves a factor of order $\log(1/\varepsilon)$. For $T = \Theta(\log(1/\varepsilon))$ above the required threshold, both policy classes have cost $\Theta(\log^2(1/\varepsilon))$.

We also consider a setting in which we may choose the deterministic horizon. When the step size satisfies a stronger condition, we obtain $\inf_{T \geq 1} \mathcal{C}(\varepsilon, T) = \Theta\big(\log^2(1/\varepsilon)\big)$. The stronger condition rules out finite-time cancellation. Without this restriction, suitable parameters and initialization can reach the target in one iteration. In that case, the optimized-horizon cost falls to $\Theta(\log(1/\varepsilon))$. Monte Carlo experiments illustrate the terminal precision ramp and compare its finite-accuracy cost with the tested fixed-width formats.

1.1 Related Work

Low-precision optimizer states. Research on reduced-precision training covers fixed-point, floating-point, and mixed-precision formats (Gupta et al., 2015; Micikevicius et al., 2018). For adaptive optimizers, blockwise 8-bit and 4-bit methods reduce the memory used by optimizer states and demonstrate effective training with small state formats (Dettmers et al., 2022; Li et al., 2023). Tang et al. (2026) analyze relative floating-point errors in the parameters, gradients, and both moments of adaptive optimizers. They also derive a sufficient schedule of increasing precision that retains near-full-precision convergence rates. These results provide practical low-bit formats or sufficient precision conditions for convergence. They do not characterize the minimum cumulative precision over an entire run when all four write widths may vary over time and respond to realized numerical errors.

Dynamic precision allocation. Dynamic quantization methods adjust precision as the effect of numerical errors changes over a run. DQ-SGD allocates gradient precision to reduce communication while meeting a convergence-error constraint. The analysis also gives tight bounds in restricted quadratic settings (Yan et al., 2021). Sensitivity-based methods also allocate mixed integer precision during low-bit training (Ding et al., 2024). These methods vary precision over time, but they target different numerical resources or restrict which precision choices can change. Neither determines the minimum total cost of all four Adam writes when each width can depend on the full revealed history.

Precision and information constraints. A separate line of work studies how communication or precision constraints on gradient oracles affect optimization rates. Mayekar & Tyagi (2020) study the precision needed to retain stochastic-optimization rates with fixed-length quantized gradients, while Acharya et al. (2021) allow adaptive processing of gradient information. In these models, the oracle output conveys information about an unknown objective. We fix the objective and compute each gradient exactly before its noisy write; the write errors then enter a prescribed Adam recursion. We therefore measure the cumulative widths of the parameter, gradient, and two moment writes rather than the information supplied by an optimization oracle.

Together, these lines of work provide low-bit formats, sufficient convergence conditions, and dynamic allocation methods. They leave open the minimum cumulative cost of all four Adam writes under causal, history-dependent policies. Appendix F gives a more detailed comparison with related numerical models, including stochastic rounding, error feedback, and adaptive quantization ranges.

2 Dithered Adam and Precision Complexity

In this section, we define the dithered Adam recursion and causal precision policies for a scalar quadratic objective. For this policy class, Theorem 1 characterizes, up to constant factors, the precision complexity of reaching a prescribed terminal mean-square error.

2.1 Numerical Oracle and Causal Policies

Throughout this paper, $\log$ denotes the base-two logarithm, $\ln$ the natural logarithm, and $[u]_+ = \max\{u, 0\}$ the positive part of u.

We study dithered Adam applied to the scalar objective $f(x) = x^2/2$, with learning rate $\eta > 0$, denominator damping $\sigma > 0$, and moment-decay parameters β_1, β_2. The initialization and optimizer parameters satisfy

$$x_0 \neq 0, \qquad m_0 = v_0 = 0, \qquad 0 < \beta_1, \beta_2 < 1. \tag{1}$$

We write $\gamma_i = 1 - \beta_i$ for $i \in \{1, 2\}$. These quantities are fixed independently of the target accuracy ϵ and deterministic horizon T. Arithmetic between writes is exact; only the four stored writes below are perturbed.

Write errors. Each iteration writes the gradient, first moment, second moment, and parameter, in that order. We fix arbitrary error scales $r_x, r_g, r_m, r_v > 0$ across iterations, independently of the target accuracy ε and horizon T. At iteration t, before each write $s \in \{g, m, v, x\}$, a controller

chooses an almost surely finite integer precision $b_t^s \geq 1$ using all information revealed so far. We denote the information available before the error is drawn by $\mathcal{F}_{t,s}$, including earlier write outcomes within the same iteration, the selected precision, and any independent controller randomization. The write then receives the additive error

$$\delta_t^s = r_s 2^{-b_t^s}, \qquad \xi_t^s = \delta_t^s U_t^s, \qquad U_t^s | \mathcal{F}_{t,s} \sim \mathrm{Unif}[-1, 1]. \tag{2}$$

Within iteration t, the decisions and fresh errors occur in the order

$$b_t^g \to \xi_t^g \to b_t^m \to \xi_t^m \to b_t^v \to \xi_t^v \to b_t^x \to \xi_t^x.$$

Thus b_t^x may depend on the realized second-moment write error, whereas b_t^v cannot. Each fresh dither is uniform conditional on the information available before its write. We use b_t^s as a numerical precision index: when a fixed quantization range suffices, its grid point uses $b_t^s + O(1)$ encoded bits. The range qualification and the general-policy cost interpretation are given after Theorem 1.

Dithered Adam. For $t \geq 1$, we denote the bias-corrected moments and effective step size by

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}, \qquad \widehat{v}_t = \frac{v_t}{1 - \beta_2^t}, \qquad s_t = \frac{\eta}{\sigma + \sqrt{\widehat{v}_t}}. \tag{3}$$

The stored quantities follow the Adam recursion

$$\begin{aligned} g_t &= x_{t-1} + \xi_t^g, & m_t &= \beta_1 m_{t-1} + \gamma_1 g_t + \xi_t^m, \\ v_t &= [\beta_2 v_{t-1} + \gamma_2 g_t^2 + \xi_t^v]_+, & x_t &= x_{t-1} - s_t \widehat{m}_t + \xi_t^x. \end{aligned} \tag{4}$$

We refer to (x_t, m_t, v_t) as the state at time t. Note that we project the second-moment onto $[0, \infty)$ to ensure that $\sqrt{\widehat{v}_t}$ remains well defined, since the additive write error may otherwise make v_t negative.

2.2 Precision complexity and main results

Definition 1 (Causal precision policies)**.** Given a deterministic horizon T before the run, a precision policy is called causal if it chooses each almost surely finite integer $b_t^s \geq 1$ using only the information available before that write's fresh error, as specified above. The collection of all causal precision policies is denoted by Π_T.

For each policy $\pi \in \Pi_T$, we denote by P_π the probability law induced by the numerical oracle and the policy, including any controller randomization, and by $\mathbb{E}_\pi$ the corresponding expectation.

Definition 2 (Precision complexity)**.** For a policy $\pi \in \Pi_T$, the terminal mean-square error and expected cumulative precision cost are defined as

$$\mathcal{R}_T(\pi) = \mathbb{E}_\pi x_T^2, \qquad \mathcal{B}_T(\pi) = \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s \in \{x,g,m,v\}} b_t^s, \tag{5}$$

respectively. The precision complexity at accuracy $\varepsilon > 0$ and horizon T is defined as

$$\mathcal{C}(\varepsilon, T) = \inf_{\substack{\pi \in \Pi_T \\ \mathcal{R}_T(\pi) \leq \varepsilon}} \mathcal{B}_T(\pi). \tag{6}$$

We impose the following additional condition to establish the upper bound:

$$0 < \frac{\eta}{\sigma} < \frac{2(1 - \beta_1)}{1 + \beta_1}. \tag{7}$$

Theorem 1 (Precision complexity at prescribed horizons)**.** *For every sufficiently small $\varepsilon > 0$, there exists a threshold $n_\varepsilon = \Theta(\log(1/\varepsilon))$ such that, for every deterministic horizon $T \geq n_\varepsilon$, condition (7) yields*

$$c \left[T + \log^2 \left(\frac{1}{\varepsilon} \right) \right] \leq \mathcal{C}(\varepsilon, T) \leq C \left[T + \log^2 \left(\frac{1}{\varepsilon} \right) \right], \tag{8}$$

where $c, C > 0$ are independent of ε and T. In particular, the upper bound is attained by a deterministic policy, and the lower bound holds without (7).

In (8), the T term accounts for the minimum cost per iteration, while the $\log^2(1/\varepsilon)$ term captures the additional precision needed to reach the target accuracy.

If the horizon can also be chosen, taking $T = n_\varepsilon$ gives a cost of order $\log^2(1/\varepsilon)$. However, Theorem 1 does not rule out the possibility that a policy could attain the same target accuracy at a lower cumulative precision cost by using a shorter horizon. The next corollary shows that this order remains optimal over all deterministic horizons under a stronger step-size condition.

Corollary 2 (Precision complexity over optimized horizons)**.** *Assume the stronger step-size condition*

$$0 < \frac{\eta}{\sigma} < \frac{1-\beta_1}{2(1+\beta_1)}. \tag{9}$$

Then, for all sufficiently small $\varepsilon > 0$, $\inf_{T\geq 1} \mathcal{C}(\varepsilon, T) = \Theta(\log^2(1/\varepsilon))$.

The proof is given in Appendix A.3. The additional step-size restriction cannot be omitted in general. Appendix A.4 gives an example in which a deterministic policy attains the target accuracy in a single iteration despite the write errors. For this example, the optimal cost over horizons is $\Theta(\log(1/\varepsilon))$.

Bit-cost interpretation The write-error model in (2) can be implemented using subtractively dithered quantization (Akyol & Rose, 2013). For the deterministic policy π in Section 3, the state bounds and bounded write errors permit fixed quantization ranges without overflow. A grid with spacing $2r_s 2^{-b_t^s}$ within such a range has $O(2^{b_t^s})$ points, so each write requires $b_t^s + O(1)$ bits to encode its grid index. The total encoding cost therefore differs from $\mathcal{B}_T(\pi)$ by at most $O(T)$, preserving the order in Theorem 1.

For general policies, $\mathcal{B}_T(\pi)$ measures selected write widths and does not include the cost of representing quantization ranges. We exclude computation costs and assume that the dither, quantization formats, and selected widths are shared by the encoder and decoder without charge. The upper schedule has predetermined widths, while the lower bound allows adaptive policies to use their width information for free.

2.3 Comparison with fixed write widths

To measure the benefit of allocating precision over time, let $\Pi_T^{\text{fixed}} \subseteq \Pi_T$ be the class of policies that choose the four write widths before the run and keep them fixed for all T iterations. We define the precision complexity within this class by

$$\mathcal{C}_{\text{fixed}}(\varepsilon, T) = \inf_{\substack{\pi \in \Pi_T^{\text{fixed}} \\ \mathcal{R}_T(\pi) \leq \varepsilon}} \mathcal{B}_T(\pi).$$

Corollary 3 (Precision complexity for fixed write widths)**.** *Assume that the step-size condition in (7) holds. Then, for every sufficiently small $\varepsilon > 0$ and every $T \geq n_\varepsilon$,*

$$\mathcal{C}_{\text{fixed}}(\varepsilon, T) = \Theta\left(T \log \frac{1}{\varepsilon}\right),$$

where $n_\varepsilon = \Theta(\log(1/\varepsilon))$ is the horizon threshold in Theorem 1.

For $T \geq \log^2(1/\varepsilon)$, allowing the widths to vary over time therefore reduces the optimal cumulative precision cost by a factor of order $\log(1/\varepsilon)$.

3 A matching precision schedule

In this section, we construct a precision schedule that achieves the upper bound in Theorem 1. The schedule keeps the second-moment precision fixed and gradually increases the precisions of the parameter, gradient, and first-moment writes during the final iterations. The following lemma bounds how write errors affect the parameter and moments over time.

To construct the deterministic precision schedule for the upper bound, we choose $R = \max\{1, 2|x_0|\}$ and $V = 8R^2$, and use the auxiliary region

$$\mathcal{D} = \{(x, m, v) : |x| \leq R,\ |m| \leq 3R,\ 0 \leq v \leq V\}. \tag{10}$$

Lemma 4. *Assume that the step-size condition in (7) holds. There exist fixed constants $q \in [1/2, 1)$, $H > 0$, $0 < \bar{\delta} \le R$ with the following property. If a causal policy satisfies, almost surely for every $1 \le t \le T$,*

$$\delta_t^s \le \bar{\delta}, \qquad \delta_t^v \le \frac{\gamma_2 V}{2}, \qquad \forall s \in \{x, g, m\}, \tag{11}$$

then $(x_t, m_t, v_t) \in \mathcal{D}$ for every $0 \le t \le T$ almost surely. With the convention $\widehat{m}_0 = 0$, we define

$$N_t = \sqrt{x_t^2 + \beta_1 (\widehat{m}_t - x_t)^2}, \qquad 0 \le t \le T.$$

Then for every $1 \le t \le T$, $|x_t| \le N_t \le R$, $|\widehat{m}_t| \le 2R$, $0 \le v_t \le V(1 - \beta_2^t)$, and

$$N_t \le q N_{t-1} + 2\delta_t^x + H\left(\delta_t^g + \frac{\delta_t^m}{\gamma_1}\right). \tag{12}$$

holds almost surely. Here $N_0 = \sqrt{1 + \beta_1}|x_0|$.

The proof is given in Appendix C.1.

For this construction, we choose fixed positive integers b_0, b_v, independently of ε and T, such that

$$r_s 2^{-b_0} \le \bar{\delta}, \qquad r_v 2^{-b_v} \le \frac{\gamma_2 V}{2}, \qquad \forall s \in \{x, g, m\}. \tag{13}$$

We define $\chi = -\log q$, where q is the contraction factor in Lemma 4.

Proposition 5 (Upper bound for $\mathcal{B}_T(\pi)$)**.** *For all sufficiently small $\varepsilon > 0$, there exists an integer $n_\varepsilon = \Theta(\log(1/\varepsilon))$ such that, for every deterministic $T \ge n_\varepsilon$, the schedule defined by*

$$\begin{aligned} b_t^s &= b_0, && 1 \le t \le T - n_\varepsilon + 1, \\ b_{T-n_\varepsilon+1+j}^s &= b_0 + \lceil \chi j \rceil, && 1 \le j \le n_\varepsilon - 1, \qquad \forall s \in \{x, g, m\}, \\ b_t^v &= b_v, && 1 \le t \le T, \end{aligned} \tag{14}$$

defines a deterministic policy π with states in $\mathcal{D}$, $x_T^2 \le \varepsilon$ almost surely, and $\mathcal{B}_T(\pi) = O(T + \log^2(1/\varepsilon))$, where the implied constant is independent of T.

Proof. Let n_ε be an integer depending on ε, to be specified below. For every $T \ge n_\varepsilon$, the choice in (13) makes the schedule (14) satisfy (11) at every iteration. Therefore, Lemma 4 gives $(x_t, m_t, v_t) \in \mathcal{D}$ for every $0 \le t \le T$ and $N_{T-n_\varepsilon+1} \le R$ almost surely. For $t = T - n_\varepsilon + 1 + j$, the precision schedule and $\lceil \chi j \rceil \ge \chi j$ yield $\delta_{T-n_\varepsilon+1+j}^s \le \bar{\delta} 2^{-\chi j} = \bar{\delta} q^j$ for all $1 \le j \le n_\varepsilon - 1$ and $s \in \{x, g, m\}$. Substituting these bounds into (12), we thus obtain

$$N_{T-n_\varepsilon+1+j} \le q N_{T-n_\varepsilon+j} + C\bar{\delta} q^j, \qquad 1 \le j \le n_\varepsilon - 1.$$

for some fixed $C > 0$. Iterating this inequality gives

$$N_{T-n_\varepsilon+1+j} \le q^j N_{T-n_\varepsilon+1} + C\bar{\delta} \sum_{i=1}^{j} q^{j-i} q^i = q^j \left(N_{T-n_\varepsilon+1} + Cj\bar{\delta}\right), \qquad 1 \le j \le n_\varepsilon - 1,$$

At $j = n_\varepsilon - 1$, we obtain

$$|x_T| \le N_T \le [R + C(n_\varepsilon - 1)\bar{\delta}] q^{n_\varepsilon - 1}. \tag{15}$$

By choosing

$$n_\varepsilon = 1 + \left\lceil \frac{\log(1/\varepsilon)}{\chi} \right\rceil, \tag{16}$$

we obtain $q^{n_\varepsilon - 1} \le \varepsilon$ and $n_\varepsilon - 1 \le \log(1/\varepsilon)/\chi + 1$, since $q = 2^{-\chi}$. Thus (15) gives $|x_T| \le \varepsilon \left[R + C\bar{\delta}\left(\log(1/\varepsilon)/\chi + 1\right)\right] \le \sqrt{\varepsilon}$ for all sufficiently small ε, almost surely. Hence $x_T^2 \le \varepsilon$ and $n_\varepsilon = \Theta(\log(1/\varepsilon))$.

To calculate the expected cumulative precision cost $\mathcal{B}_T(\pi)$ of this schedule, note that the widths are deterministic and $0 \le \lceil \chi j \rceil - \chi j < 1$ for $1 \le j \le n_\varepsilon - 1$. Consequently,

$$\begin{aligned}\mathcal{B}_T(\pi) &= \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in\{x,g,m,v\}} b_t^s = (3b_0 + b_v)T + 3\sum_{j=1}^{n_\varepsilon - 1} \lceil \chi j \rceil \\ &= (3b_0 + b_v)T + 3\chi \sum_{j=1}^{n_\varepsilon - 1} j + O(n_\varepsilon) \\ &= O(T + n_\varepsilon^2) = O\left(T + \log^2\left(\frac{1}{\varepsilon}\right)\right). \end{aligned} \tag{17}$$

This proves the upper bound in Theorem 1. □

4 The cost of concentrating a terminal iterate

In this section, we derive a cumulative precision lower bound for every causal precision policy and every deterministic horizon $T \ge 1$. For $T \ge n_\varepsilon$, it yields the lower bound in Theorem 1. We show that low precision at an earlier iteration limits the probability that x_T lies near zero. Summing the resulting precision requirements over the final iterations gives the cumulative cost bound. The two lemmas below bound state growth and the density of the joint state, respectively.

Lemma 6 (Polynomial state growth). *There exist constants $C_x, C_m, C_v > 0$, independent of the policy and horizon, such that for every deterministic $T \ge 1$, every $\pi \in \Pi_T$, and every $0 \le t \le T$, the bounds*

$$|x_t| \le C_x(1+t)^h, \quad |m_t| \le C_m(1+t)^h, \quad 0 \le v_t \le C_v(1+t)^{2h} \tag{18}$$

hold almost surely, where h is defined as

$$h = \max\left\{1, \frac{\log \beta_2}{\log \beta_1}\right\}. \tag{19}$$

The proof is in Appendix A.1.

For a policy $\pi \in \Pi_T$ and each $1 \le t \le T$, we define the event

$$E_t^+ = \{\xi_j^v > 0 \text{ for every } t \le j \le T\}.$$

The sign of each fresh ξ_j^v is independent of the history before that write and is positive with probability $1/2$. Thus $P_\pi(E_t^+) = 2^{-(T-t+1)}$. We denote the conditional probability given E_t^+ by $Q_t = P_\pi(\cdot \mid E_t^+)$. Notice that on E_t^+, $v_j = \beta_2 v_{j-1} + \gamma_2 g_j^2 + \xi_j^v$ for every $t \le j \le T$. For an integer $K \ge 3$, let A_t^K be the event that the width vector (b_t^x, b_t^m, b_t^v) has total width at most K, that is,

$$A_t^K = \{b_t^x + b_t^m + b_t^v \le K\}.$$

Lemma 7 (Density bound). *For any deterministic $T \ge 1$, policy $\pi \in \Pi_T$, $1 \le t \le T$, and integer $K \ge 3$, there exist nonnegative densities ρ_j, $j = t, \dots, T$, such that for every measurable subset $B \subseteq \mathbb{R}^3$,*

$$\int_B \rho_j(x, m, v)\, \mathrm{d}x\, \mathrm{d}m\, \mathrm{d}v = Q_t\big(A_t^K \cap \{(x_j, m_j, v_j) \in B\}\big).$$

Moreover, there exists a fixed $G > 2$, depending only on β_1, β_2, such that

$$\|\rho_j\|_2 \le \frac{G^{j-t} 2^{\frac{K}{2}}}{\sqrt{4 r_x r_m r_v}}, \qquad \forall j \in \{t, \dots, T\}. \tag{20}$$

These densities have total mass $Q_t(A_t^K)$, rather than one. The lemma is proved in Appendix A.2.

To state the cumulative precision lower bound, we define

$$D = 1 + 4 \log G > 5. \tag{21}$$

Proposition 8 (Lower bound for $\mathcal{B}_T(\pi)$)**.** *There exist fixed constants $c, \zeta > 0$ such that, for every deterministic $T \geq 1$ and every $\pi \in \Pi_T$,*

$$\mathcal{B}_T(\pi) \geq 4T + \frac{1}{16} \sum_{k=0}^{T-1} \left[\log \frac{\zeta}{\mathcal{R}_T(\pi)} - 6h \log(1+T) - Dk \right]_+ . \tag{22}$$

Consequently, for every sufficiently small $\varepsilon > 0$, if $\mathcal{R}_T(\pi) \leq \varepsilon$ and $T \geq \lceil \log(1/\varepsilon)/D \rceil$, then

$$\mathcal{B}_T(\pi) \geq c \left[T + \log^2 \left(\frac{1}{\varepsilon} \right) \right]. \tag{23}$$

The n_ε in Proposition 5 can also be used as the horizon threshold for (23).

Proof of Proposition 8. We fix a policy $\pi \in \Pi_T$. If $\mathcal{B}_T(\pi) = \infty$, the claimed lower bounds are immediate. Otherwise, since all widths are nonnegative, $S_t = \mathbb{E}_\pi(b_t^x + b_t^m + b_t^v) \leq \mathcal{B}_T(\pi) < \infty$ for every $1 \leq t \leq T$.

We first derive a lower bound on the terminal mean-square error from a single iteration. Fix $1 \leq t \leq T$. At iteration t, the policy selects b_t^m, b_t^v before drawing ξ_t^v, but may use this error when selecting b_t^x. Consequently,

$$\begin{aligned} &\mathbb{E}_\pi[b_t^m | E_t^+] = \mathbb{E}_\pi b_t^m, \qquad \mathbb{E}_\pi[b_t^v | E_t^+] = \mathbb{E}_\pi b_t^v, \\ &\mathbb{E}_\pi[b_t^x | E_t^+] = \frac{2^{-(T-t)} \mathbb{E}_\pi[b_t^x \mathbf{1}_{\{\xi_t^v > 0\}}]}{2^{-(T-t+1)}} = 2\mathbb{E}_\pi[b_t^x \mathbf{1}_{\{\xi_t^v > 0\}}] \leq 2\mathbb{E}_\pi b_t^x. \end{aligned}$$

Therefore, we obtain

$$\mathbb{E}_\pi[b_t^x + b_t^m + b_t^v | E_t^+] \leq 2S_t. \tag{24}$$

We choose $K = \lceil 8S_t \rceil$, so $K \leq 8S_t + 1$. By Markov's inequality and (24), we have $Q_t(A_t^K) \geq 1 - 2S_t/K \geq 3/4$. Notice that Lemma 6 gives $|m_T| \leq C_m(1+T)^h$ and $0 \leq v_T \leq C_v(1+T)^{2h}$. For $r > 0$, we define

$$B_r = \left\{ (x, m, v) \in \mathbb{R}^3 : |x| \leq r,\ |m| \leq C_m(1+T)^h,\ 0 \leq v \leq C_v(1+T)^{2h} \right\}.$$

It has volume $4C_mC_v(1+T)^{3h}r$. Combining Lemma 7 with the Cauchy–Schwarz inequality gives

$$\begin{aligned} Q_t(A_t^K \cap \{|x_T| \leq r\}) &= \int_{B_r} \rho_T(x, m, v)\, \mathrm{d}x\, \mathrm{d}m\, \mathrm{d}v \\ &\leq \|\mathbf{1}_{B_r}\|_2 \|\rho_T\|_2 = \sqrt{4C_mC_v(1+T)^{3h}r}\, \|\rho_T\|_2 \\ &\leq C(1+T)^{\frac{3h}{2}} G^{T-t} 2^{\frac{K}{2}} \sqrt{r}. \end{aligned} \tag{25}$$

If we pick $r = (4C)^{-2}(1+T)^{-3h} G^{2(t-T)} 2^{-K}$, then the bound in (25) is at most $1/4$. Notice that $Q_t(A_t^K) \geq 3/4$, so $Q_t(A_t^K \cap \{|x_T| > r\}) \geq Q_t(A_t^K) - Q_t(A_t^K \cap \{|x_T| \leq r\}) \geq 1/2$. Consequently, $\mathbb{E}_\pi[x_T^2 | E_t^+] \geq r^2 Q_t(A_t^K \cap \{|x_T| > r\}) \geq r^2/2$. Therefore, for a fixed $c_1 > 0$, we have

$$\begin{aligned} \mathcal{R}_T(\pi) &\geq P_\pi(E_t^+) \mathbb{E}_\pi[x_T^2 | E_t^+] \geq 2^{-(T-t+2)} r^2 = \frac{(1+T)^{-6h} 2^{-2K}}{4(4C)^4}\, 2^{-(T-t)} G^{-4(T-t)} \\ &\geq c_1 (1+T)^{-6h} 2^{-16S_t - D(T-t)}. \end{aligned} \tag{26}$$

Here the last inequality uses $K \leq 8S_t + 1$ and $D = 1 + 4\log G$. Taking logarithms in (26) gives $S_t \geq (1/16)\left(\log(c_1/\mathcal{R}_T(\pi)) - 6h\log(1+T) - D(T-t)\right)$. Each of the three widths is at least one, so $S_t \geq 3$. Choose a sufficiently small fixed $\zeta > 0$ so that $\log(\zeta/\mathcal{R}_T(\pi)) \leq \log(c_1/\mathcal{R}_T(\pi)) - 48$. We write $\ell_T = \log(\zeta/\mathcal{R}_T(\pi)) - 6h\log(1+T)$.

$$S_t \geq \max\left\{ 3,\ 3 + \frac{1}{16}\big(\ell_T - D(T-t)\big) \right\} = 3 + \frac{1}{16}[\ell_T - D(T-t)]_+. \tag{27}$$

Summing (27) over $t = 1, \ldots, T$ gives $\mathcal{B}_T(\pi) = \sum_{t=1}^T (S_t + \mathbb{E}_\pi b_t^g) \geq 4T + (1/16)\sum_{k=0}^{T-1} [\ell_T - Dk]_+$, which proves (22) for every horizon T.

Now assume $\mathcal{R}_T(\pi) \le \varepsilon$ and $T \ge \lceil \log(1/\varepsilon)/D \rceil$. We write $L = \log(1/\varepsilon)$.

Case 1: $T \le L^2$**.** For sufficiently small ε, we have $\ell_T \ge L + \log\zeta - 6h\log(1+L^2) \ge L/2$. Since $T \ge \lceil L/D \rceil$, we obtain

$$\mathcal{B}_T(\pi) \ge 4T + \frac{1}{16}\sum_{k=0}^{T-1}\left[\frac{L}{2} - Dk\right]_+ \ge 4T + \frac{1}{16}\int_0^{L/(2D)}\left(\frac{L}{2} - Du\right)\mathrm{d}u = 4T + \frac{L^2}{128D}.$$

Case 2: $T > L^2$**.** The baseline cost gives $\mathcal{B}_T(\pi) \ge 4T \ge 2(T + L^2)$.

Combining the two cases yields (23) with a fixed $c > 0$. The choice of n_ε in the proof of Proposition 5 satisfies $n_\varepsilon \ge \lceil \log(1/\varepsilon)/D \rceil$ because $\chi = -\log q \le 1 < D$. Hence (23) gives the lower bound in Theorem 1 for every $T \ge n_\varepsilon$. □

5 Finite-accuracy illustrations

We simulate the scalar oracle in (4) with $(\beta_1, \beta_2) = (0.9, 0.99)$, $\eta = 1/19$, $\sigma = 1$, $x_0 = 1/4$, and unit write-error scales. These parameters satisfy (7). Candidate schedules are selected on independent pilot and confirmation samples and evaluated once on a fresh 20,000-trajectory holdout.

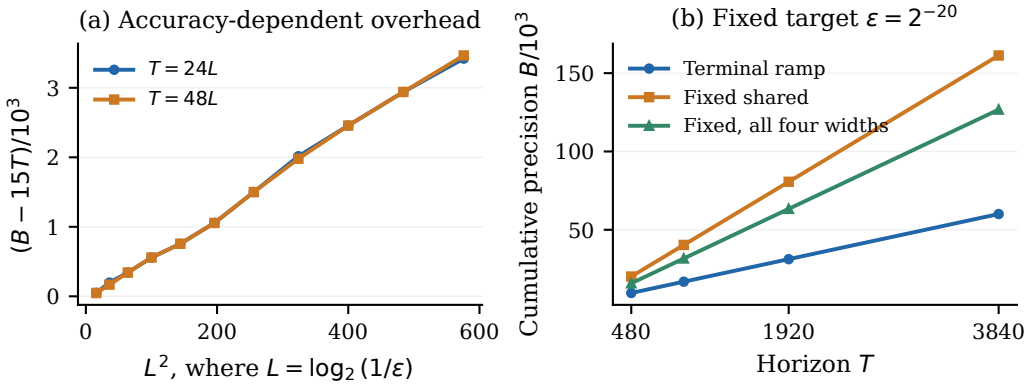

Figure 2: **Finite-accuracy costs in the scalar dither oracle.** (a) Excess precision above the coarse $15T$ cost versus L^2, where $L = \log_2(1/\epsilon)$. (b) Cost versus horizon at $\epsilon = 2^{-20}$. All plotted schedules satisfy $\widehat{\mathcal{R}}_T + 2\mathrm{SE} \le \epsilon$; the four-channel fixed baseline is a finite search rather than the infimum in Corollary 3.

Figure 2(a) shows an approximately linear L^2 excess cost over the tested range, while Figure 2(b) shows a growing cost gap between the selected terminal ramp and searched fixed-width baselines at longer horizons. The four-channel baseline is a finite search rather than the infimum in Corollary 3. After recalibrating the other three writes, all tested constant second-moment widths reach the targets, and ramping all four widths is not cheaper than the six-bit constant-v schedule at the strictest target. Full results are in Appendix E.

6 Conclusion

Our results characterize the optimal cumulative precision of dithered Adam under the stated step-size and horizon conditions: a deterministic terminal ramp attains the optimal order among all causal policies and improves on fixed widths at longer horizons. Optimizing the horizon preserves the squared-logarithmic cost under a stronger step-size condition, while finite-time cancellation can reduce the order without it. The same complexity extends, under explicit conditions, to scalar strongly convex smooth and separable objectives, highlighting the role of when precision is allocated.

AI USE STATEMENT

We used OpenAI GPT-5.6 during the research and writing of this paper. It assisted the authors in developing and checking the maximal-averaging approach to the lower bound and the arguments extending the results to scalar strongly convex smooth objectives and higher-dimensional separable objectives.

GPT-5.6 was also used to help organize and edit parts of the manuscript, improve the presentation of proofs, check notation and consistency, assist with literature search and synthesis, and prepare the conceptual figure and LaTeX. The authors independently verified all final mathematical claims, constructions, constants, and proofs and take responsibility for the full content of the paper.

REPRODUCIBILITY STATEMENT

Section 2 specifies the numerical oracle, causal policy class, and cumulative precision cost. The proofs are in Sections 3–4 and Appendices A–D. Appendix A.1 derives the state-growth bounds used in the lower bound. Appendix A.4 proves the one-step cancellation example. Section 5 reports scalar simulations, and Appendix E specifies their protocol. Appendix E.6 describes the accompanying code and saved results. The conceptual figures are schematic. No neural-network training or hardware measurements are used.

REFERENCES

Jayadev Acharya, Clement Louis Canonne, Prathamesh Mayekar, and Himanshu Tyagi. Information-constrained optimization: can adaptive processing of gradients help? In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=h7aSBWbX7S4.

Emrah Akyol and Kenneth Rose. On constrained randomized quantization. *IEEE Transactions on Signal Processing*, 61(13):3291–3302, 2013. doi: 10.1109/TSP.2013.2261296.

Congliang Chen, Li Shen, Haozhi Huang, and Wei Liu. Quantized Adam with error feedback. *ACM Trans. Intell. Syst. Technol.*, 12(5), September 2021. ISSN 2157-6904. doi: 10.1145/3470890. URL https://doi.org/10.1145/3470890.

Congliang Chen, Li Shen, Wei Liu, and Zhi-Quan Luo. Efficient-Adam: Communication-efficient distributed Adam. *IEEE Transactions on Signal Processing*, 71:3257–3266, 2023. doi: 10.1109/TSP.2023.3309461.

Matteo Croci, Massimiliano Fasi, Nicholas J. Higham, Theo Mary, and Mantas Mikaitis. Stochastic rounding: implementation, error analysis and applications. *Royal Society Open Science*, 9(3): 211631, 03 2022. ISSN 2054-5703. doi: 10.1098/rsos.211631. URL https://doi.org/10.1098/rsos.211631.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=shpkpVXzo3h.

Li Ding, Wen Fei, Yuyang Huang, Shuangrui Ding, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. AMPA: Adaptive mixed precision allocation for low-bit integer training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10960–10977. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ding24b.html.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1737–1746, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/gupta15.html.

G. H. Hardy and J. E. Littlewood. A maximal theorem with function-theoretic applications. *Acta Mathematica*, 54:81–116, 1930. doi: 10.1007/BF02547518. URL https://doi.org/10.1007/BF02547518.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Bingrui Li, Jianfei Chen, and Jun Zhu. Memory efficient optimizers with 4-bit states. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=nN8TnHB5nw.

Prathamesh Mayekar and Himanshu Tyagi. Limits on gradient compression for stochastic optimization. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2658–2663, 2020. doi: 10.1109/ISIT44484.2020.9174075.

Nicolò Michelusi, Gesualdo Scutari, and Chang-Shen Lee. Finite-bit quantization for distributed algorithms with linear convergence. *IEEE Transactions on Information Theory*, 68(11):7254–7280, 2022. doi: 10.1109/TIT.2022.3176253.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

Alexander Nagel, Nestor Riviere, and Stephen Wainger. A maximal function associated to the curve (t, t^2). *Proceedings of the National Academy of Sciences*, 73(5):1416–1417, 1976. doi: 10.1073/pnas.73.5.1416. URL https://www.pnas.org/doi/abs/10.1073/pnas.73.5.1416.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4402–4410. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/ozkara25b.html.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wwP1SCACee.

Kristi Topollai and Anna Choromanska. Understanding quantization of optimizer states in LLM pre-training: Dynamics of state staleness and effectiveness of state resets. *arXiv preprint arXiv:2603.16731*, 2026. URL https://arxiv.org/abs/2603.16731.

Yan Wang. Beyond dense Adam states: Adaptive log-space quantization for memory-efficient optimizers. arXiv:2608.22322v3, 2026. URL https://arxiv.org/abs/2608.22322v3.

Cong Xu, Wenbin Liang, Mo Yu, Anan Liu, Ke-Yue Zhang, Shunli Wang, Lizhuang Ma, Jianyong Wang, Jun Wang, and Wei Zhang. Pushing the limits of low-bit optimizers: A focus on EMA dynamics. *arXiv preprint arXiv:2505.00347*, 2025. URL https://arxiv.org/abs/2505.00347.

Guangfeng Yan, Shao-Lun Huang, Tian Lan, and Linqi Song. DQ-SGD: Dynamic quantization in SGD for communication-efficient distributed learning. In *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pp. 136–144, 2021. doi: 10.1109/MASS52906.2021.00025.

A LOWER BOUNDS AND OPTIMIZED HORIZONS

In this section, we prove the state-growth and density estimates for Proposition 8 and Corollary 2. We then give an example showing why the corollary requires an additional step-size condition.

A.1 PROOF OF LEMMA 6

Proof. Throughout this proof, C may change from line to line but depends only on the fixed parameters and initialization. Every write has radius $r_s 2^{-b_t^s} \le r_s/2$. Hence $|\xi_t^s| \le r_s/2$ almost surely for every causal policy. Let $w_t = v_t + r_v/(2\gamma_2)$. Since $[u]_+ \ge u$ and $\xi_t^v \ge -r_v/2$, the second-moment recursion yields

$$w_t \ge \beta_2 w_{t-1} + \gamma_2 g_t^2, \qquad \gamma_2 \beta_2^{t-k} g_k^2 \le w_t, \qquad 1 \le k \le t. \tag{28}$$

Combining (28) with $\widehat{v}_t \ge v_t$ gives

$$\frac{|g_k|}{\sigma + \sqrt{\widehat{v}_t}} \le \frac{\beta_2^{-(t-k)/2}}{\sqrt{\gamma_2}} \frac{\sqrt{v_t} + \sqrt{r_v/(2\gamma_2)}}{\sigma + \sqrt{v_t}} \le C\beta_2^{-(t-k)/2}, \qquad 1 \le k \le t. \tag{29}$$

We define $\theta = 1/h \in (0, 1]$ and

$$H_t = 1 + \frac{r_g}{2} + \max_{0 \le k \le t} |x_k|.$$

For $k \le t$, $g_k = x_{k-1} + \xi_k^g$, so $|g_k| \le |x_{k-1}| + r_g/2 \le H_{t-1}$. Since $\sigma + \sqrt{\widehat{v}_t} \ge \sigma$, this also gives $|g_k|/(\sigma + \sqrt{\widehat{v}_t}) \le H_{t-1}/\sigma$. Combining this bound with (29) and $\min\{a, b\} \le a^{1-\theta} b^\theta$ gives

$$\frac{|g_k|}{\sigma + \sqrt{\widehat{v}_t}} \le C H_{t-1}^{1-\theta} \beta_2^{-\theta(t-k)/2}, \qquad 1 \le k \le t. \tag{30}$$

The choice of h gives $\beta_2^\theta \ge \beta_1$, so $\beta_1 \beta_2^{-\theta/2} \le \sqrt{\beta_1} < 1$. The bias-corrected first moment expands as

$$\widehat{m}_t = \frac{\gamma_1}{1 - \beta_1^t} \sum_{k=1}^t \beta_1^{t-k} g_k + \frac{1}{1 - \beta_1^t} \sum_{k=1}^t \beta_1^{t-k} \xi_k^m.$$

Using (30) and $|\xi_k^m| \le r_m/2$ yields

$$\frac{|\widehat{m}_t|}{\sigma + \sqrt{\widehat{v}_t}} \le C H_{t-1}^{1-\theta} \sum_{k=1}^t (\beta_1 \beta_2^{-\theta/2})^{t-k} + \frac{r_m}{2\gamma_1 \sigma} \le C H_{t-1}^{1-\theta}. \tag{31}$$

By (31), the parameter update gives $|x_t - x_{t-1}| \le C H_{t-1}^{1-\theta}$. By the definition of H_t,

$$H_t = \max\{H_{t-1}, 1 + r_g/2 + |x_t|\} \le H_{t-1} + C H_{t-1}^{1-\theta}.$$

Since $0 < \theta \le 1$, we have

$$\begin{aligned} H_t^\theta &\le \left(H_{t-1} + C H_{t-1}^{1-\theta}\right)^\theta \\ &\le H_{t-1}^\theta + \theta H_{t-1}^{\theta-1} C H_{t-1}^{1-\theta} \le H_{t-1}^\theta + C. \end{aligned}$$

Summing over $j = 1, \dots, t$ gives

$$H_t^\theta - H_0^\theta = \sum_{j=1}^t (H_j^\theta - H_{j-1}^\theta) \le Ct,$$

$$H_t \le (H_0^\theta + Ct)^{1/\theta} \le C(1 + t)^h.$$

Since $[a + \xi_t^v]_+ \le a + r_v/2$ for $a \ge 0$, the moment recursions give, for $t \ge 1$,

$$|m_t| \le H_{t-1} + \frac{r_m}{2\gamma_1},$$

$$0 \le v_t \le \gamma_2 \sum_{k=1}^t \beta_2^{t-k} g_k^2 + \frac{r_v}{2\gamma_2} \le H_{t-1}^2 + \frac{r_v}{2\gamma_2}.$$

Since $H_{t-1} \le C(1 + t)^h$ and $(1 + t)^h \ge 1$, the first bound gives $|m_t| \le (C + r_m/(2\gamma_1))(1 + t)^h$. Together with the bound on H_t and the initial conditions (1), these estimates prove (18). □

A.2 Proof of Lemma 7

Proof. Step 1: Initial density. Under Q_t, $\xi_t^v > 0$, while $v_{t-1} \geq 0$. Thus the projection in the v update is inactive and does not place mass at $v_t = 0$. Conditioning successively on the m, v, and x writes gives a density for the state at time t. Their widths may depend on earlier writes, but each is fixed before its own fresh error. The m and x errors have conditional densities $1/(2\delta_t^m)$ and $1/(2\delta_t^x)$, while the positive v error has density $1/\delta_t^v$. Their product is

$$\frac{1}{2\delta_t^m}\frac{1}{\delta_t^v}\frac{1}{2\delta_t^x} = \frac{2^{b_t^m+b_t^v+b_t^x}}{4r_x r_m r_v}.$$

On A_t^K, averaging over the earlier history therefore gives $\|\rho_t\|_\infty \leq 2^K/(4r_x r_m r_v)$. Its mass is at most one, hence

$$\|\rho_t\|_2^2 \leq \|\rho_t\|_\infty \|\rho_t\|_1 \leq \frac{2^K}{4r_x r_m r_v}. \tag{32}$$

Step 2: One-iteration estimate. Fix $t \leq j < T$. For a write of type s at iteration $j+1$, A_t^K is determined before the fresh error is drawn. Conditional on the history before this write, the error remains uniform on its prescribed interval under Q_t, but its radius $\delta_{j+1}^s = r_s 2^{-b_{j+1}^s}$ may vary across histories. To bound all possible radii at once, let $z = (x, m, v)$ be the input for the state map of this write and $F_u(z)$ its output when the error is u. For example, a write of m gives $F_u(x, m, v) = (x, m+u, v)$. The allowed radii are $\mathcal{R} = \{r_s 2^{-b} : b \geq 1\}$. For a test function φ, the maximal conditional average is

$$\mathcal{M}\varphi(z) = \sup_{\delta \in \mathcal{R}} \mathbb{E}_{u\sim \mathrm{Unif}[-\delta,\delta]} |\varphi(F_u(z))|.$$

To control the density after each write, we need $\|\mathcal{M}\varphi\|_2 \leq C\|\varphi\|_2$. We write Z and Z^+ for the states before and after this write, and ϱ for the density of $Q_t(A_t^K \cap \{Z \in \cdot\})$. For every bounded compactly supported test function φ, conditioning on the full history $\mathcal{F}_{j+1,s}$ gives

$$\begin{aligned}
\left|\mathbb{E}_{Q_t}[\mathbf{1}_{A_t^K}\varphi(Z^+)]\right| &= \left|\mathbb{E}_{Q_t}\left[\mathbf{1}_{A_t^K}\mathbb{E}_{Q_t}[\varphi(F_{\xi_{j+1}^s}(Z))|\mathcal{F}_{j+1,s}]\right]\right| \\
&\leq \mathbb{E}_{Q_t}[\mathbf{1}_{A_t^K}\mathcal{M}\varphi(Z)] \\
&= \int \varrho(z)\mathcal{M}\varphi(z)\,\mathrm{d}z \leq C\|\varrho\|_2\|\varphi\|_2.
\end{aligned}$$

Here A_t^K, Z, and the selected radius are determined by $\mathcal{F}_{j+1,s}$. The supremum in $\mathcal{M}$ bounds the average for that radius even when its choice depends on information not contained in Z. By L^2 duality, the restricted law after the write has a density ϱ^+ satisfying

$$\|\varrho^+\|_2 \leq C\|\varrho\|_2.$$

For the write of v, we use the corresponding one-sided maximal average. This calculation conditions on the full history and requires no Markov assumption on the state process.

At iteration $j+1$, the write of the gradient gives $g = x + u$. The map D updates both moments using x, and E_u then adds the effect of u:

$$\begin{aligned}
D(x, m, v) &= (x, \beta_1 m + \gamma_1 x, \beta_2 v + \gamma_2 x^2), \\
E_u(x, m, v) &= (x, m + \gamma_1 u, v + \gamma_2(2xu + u^2)), \\
E_u(D(x, m, v)) &= (x, \beta_1 m + \gamma_1(x+u), \beta_2 v + \gamma_2(x+u)^2).
\end{aligned} \tag{33}$$

The deterministic map D is invertible and has $\det D' = \beta_1\beta_2$. If its input has density ϱ, its output density satisfies

$$\varrho_D(z) = \frac{\varrho(D^{-1}z)}{\beta_1\beta_2}, \qquad \|\varrho_D\|_2 = (\beta_1\beta_2)^{-1/2}\|\varrho\|_2.$$

For E_u, the coordinates

$$u_1 = \frac{m}{\gamma_1}, \qquad u_2 = \frac{v}{\gamma_2} - \frac{2xm}{\gamma_1} \quad \text{give} \quad E_u : (x, u_1, u_2) \longmapsto (x, u_1 + u, u_2 + u^2). \tag{34}$$

At each fixed x, (34) matches the dyadic parabola average bounded in Appendix B, specifically (40). The same L^2 bound holds for E_u in the original coordinates.

The subsequent writes of m and v add errors along their coordinate lines, so the line bounds in Appendix B apply. Under Q_t, the error in the write of v is uniform on $[0, \delta^v_{j+1}]$, for which Appendix B gives a one-sided line bound. Once the moments are fixed, the deterministic parameter update $x \mapsto x - s_{j+1}\widehat{m}_{j+1}$ is a translation. Its Jacobian determinant is one for $v > 0$, where the conditioned state law is supported, so this map preserves the L^2 norm of the density. The final write adds ξ^x_{j+1} and uses the line bound in the x direction.

Step 3: Density propagation. We apply the preceding density estimate successively to the gradient error and the writes of m, v, x, together with the two deterministic maps. Let C_{P} and C_{L} be the L^2 bounds for the parabola and centered line maximal operators in Appendix B. The one-sided line bound is at most $2C_{\mathrm{L}}$. Hence the restricted state law at iteration $j+1$ has a density ρ_{j+1} satisfying

$$\|\rho_{j+1}\|_2 \le \frac{2C_{\mathrm{P}}C_{\mathrm{L}}^3}{\sqrt{\beta_1\beta_2}}\|\rho_j\|_2, \qquad t \le j < T,$$

where the factor depends only on β_1, β_2. We choose G larger than two and the displayed factor and obtain

$$\|\rho_{j+1}\|_2 \le G\|\rho_j\|_2, \qquad t \le j < T. \tag{35}$$

Starting from (32), this gives

$$\|\rho_j\|_2 \le G^{j-t}\|\rho_t\|_2 \le \frac{G^{j-t}2^{\frac{K}{2}}}{\sqrt{4r_x r_m r_v}}, \qquad t \le j \le T.$$

□

A.3 Proof of Corollary 2

Proof. We assume the step-size condition in (9). We define

$$\theta_t = \frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}, \qquad \omega_t = (1-\theta_t)\left(\xi^g_t + \frac{\xi^m_t}{\gamma_1}\right).$$

With $z_t = (x_t, 2x_t - \widehat{m}_t)^\top$, the recursion becomes

$$\begin{aligned} z_t &= B_t(s_t)z_{t-1} + \begin{pmatrix}1\\2\end{pmatrix}\xi^x_t - \begin{pmatrix}s_t\\1+2s_t\end{pmatrix}\omega_t, \\ B_t(s) &= \begin{pmatrix}1-s(1+\theta_t) & s\theta_t\\ 1-\theta_t-2s(1+\theta_t) & \theta_t(1+2s)\end{pmatrix}. \end{aligned} \tag{36}$$

For every $s \in [0, \eta/\sigma]$, the matrix has nonnegative entries because

$$1-\theta_t-2s(1+\theta_t) \ge 1-\beta_1-2\left(\frac{\eta}{\sigma}\right)(1+\beta_1) > 0.$$

Its upper-left entry is at least $\lambda = 1-(\eta/\sigma)(1+\beta_1) > 1/2$.

For any policy $\pi \in \Pi_T$ with $x_0 > 0$, we consider the event $\xi^x_j \ge 0$, $\xi^g_j \le 0$ and $\xi^m_j \le 0$ for every $j \le T$. This event has probability 8^{-T}. Thus, componentwise,

$$\begin{pmatrix}1\\2\end{pmatrix}\xi^x_t - \begin{pmatrix}s_t\\1+2s_t\end{pmatrix}\omega_t \ge 0.$$

Since $z_0 = (x_0, 2x_0)^\top \ge 0$, the nonnegative matrix and additive term in (36) give $z_t \ge 0$ by induction. Its first coordinate then satisfies $x_t \ge \lambda x_{t-1}$, so $x_T \ge \lambda^T x_0$. For $x_0 < 0$, we instead take $\xi^x_j \le 0$, $\xi^g_j \ge 0$ and $\xi^m_j \ge 0$ for every $j \le T$. The additive term in (36) is then nonpositive. Since $-z_0 = (-x_0, -2x_0)^\top \ge 0$, the same induction applied to $-z_t$ gives $-x_T \ge \lambda^T(-x_0)$. Hence

$$\mathcal{R}_T(\pi) \ge x_0^2\left(\frac{\lambda^2}{8}\right)^T, \qquad \log\left(\frac{8}{\lambda^2}\right) < 5 < D. \tag{37}$$

Here D is defined in (21).

If $\mathcal{R}_T(\pi) \le \varepsilon$, taking logarithms in (37) gives, for sufficiently small ε,

$$T \ge \left\lceil \frac{\log(x_0^2/\varepsilon)}{\log(8/\lambda^2)} \right\rceil \ge \left\lceil \frac{\log(1/\varepsilon)}{D} \right\rceil, \tag{38}$$

where the last inequality uses $\log(8/\lambda^2) < D$ and fixed $x_0 \neq 0$. Thus, if $\mathcal{R}_T(\pi) \le \varepsilon$, (38) verifies the horizon condition in Proposition 8, and its lower bound gives $\mathcal{B}_T(\pi) \ge c \log^2(1/\varepsilon)$. Conversely, Proposition 5 gives $\mathcal{C}(\varepsilon, n_\varepsilon) \le C \log^2(1/\varepsilon)$. Hence

$$c \log^2(1/\varepsilon) \le \inf_{T \ge 1} \mathcal{C}(\varepsilon, T) \le C \log^2(1/\varepsilon).$$

□

A.4 One-step cancellation

We give an example in which one iteration reaches the target accuracy with cost of order $\log(1/\varepsilon)$, showing that the additional step-size condition in Corollary 2 cannot be omitted in general. Fix $\eta > \sigma$ and choose $x_0 = \eta - \sigma > 0$. With all four write errors equal to zero, the first step has

$$\widehat{m}_1 = x_0, \qquad \widehat{v}_1 = x_0^2, \qquad x_1 = x_0 \left(1 - \frac{\eta}{\sigma + x_0}\right) = 0.$$

Because $\widehat{v}_1 = x_0^2 > 0$ at zero error, it remains positive when the four errors are sufficiently small. The first-step map is then locally Lipschitz, so $|x_1| \le C \max_s |\xi_1^s|$ for a fixed C.

We choose the deterministic widths $b_1^s = \lceil (1/2) \log(1/\varepsilon) \rceil + K$ for all $s \in \{x, g, m, v\}$, where the fixed integer K is large enough that $C \max_s r_s 2^{-K} \le 1$. By (2), $|\xi_1^s| \le r_s 2^{-K} \sqrt{\varepsilon}$ almost surely. Thus, for sufficiently small ε, the local bound above gives $|x_1| \le \sqrt{\varepsilon}$ and $\mathcal{R}_1(\pi) \le \varepsilon$. The four widths cost $\mathcal{B}_1(\pi) = O(\log(1/\varepsilon))$.

Conversely, for any $T \ge 1$ and $\pi \in \Pi_T$, the final parameter error has conditional mean zero and variance $r_x^2 2^{-2b_T^x}/3$ by (2). Since $x_T = x_{T-1} - s_T \widehat{m}_T + \xi_T^x$,

$$\mathbb{E}_\pi[x_T^2 | \mathcal{F}_{T,x}] = (x_{T-1} - s_T \widehat{m}_T)^2 + \frac{r_x^2}{3} 2^{-2b_T^x} \ge \frac{r_x^2}{3} 2^{-2b_T^x}.$$

If $\mathcal{R}_T(\pi) \le \varepsilon$, taking expectations and applying Jensen's inequality gives

$$\varepsilon \ge \mathcal{R}_T(\pi) \ge \frac{r_x^2}{3} 2^{-2\mathbb{E}_\pi b_T^x}, \qquad \mathcal{B}_T(\pi) \ge \mathbb{E}_\pi b_T^x \ge \frac{1}{2} \log \frac{r_x^2}{3\varepsilon}.$$

Thus $\mathcal{R}_T(\pi) \le \varepsilon$ requires expected width of order at least $\log(1/\varepsilon)$ for the final parameter write, regardless of T. Combining this lower bound with the one-step construction gives, for these fixed parameters and initialization,

$$\inf_{T \ge 1} \mathcal{C}(\varepsilon, T) = \Theta\left(\log\left(\frac{1}{\varepsilon}\right)\right).$$

If $0 < \beta_1 < 1/3$, we can choose $1 < \eta/\sigma < 2(1 - \beta_1)/(1 + \beta_1)$, so this example also satisfies the step-size condition in (7).

B Maximal averaging estimates

In this section, we apply classical maximal-function bounds to the line and parabola averages in Appendix A.2. These bounds hold uniformly over all allowed write widths.

B.1 Line averages

For $\varphi \in L^2(\mathbb{R}^3)$, we define the maximal average along the ith coordinate direction by

$$\mathcal{M}_i \varphi(z) = \sup_{\delta > 0} \frac{1}{2\delta} \int_{-\delta}^{\delta} |\varphi(z + u e_i)| \, \mathrm{d}u, \qquad z \in \mathbb{R}^3, \quad i \in \{1, 2, 3\},$$

where e_i is the ith coordinate vector. The one-dimensional Hardy–Littlewood maximal theorem (Hardy & Littlewood, 1930), applied with the other coordinates fixed, and Fubini's theorem give

$$\|\mathcal{M}_i\varphi\|_2 \le C\|\varphi\|_2, \qquad i \in \{1,2,3\}. \tag{39}$$

A uniform average over $[0,\delta]$ or $[-\delta,0]$ is bounded by twice the centered average over $[-\delta,\delta]$, so the same type of bound holds for either one-sided maximal average. Restricting the radii to $\{r_s 2^{-b} : b \ge 1\}$ only decreases the supremum. These estimates apply to the writes of m, v, x in Appendix A.2, including the positive half-interval for the error in the write of v under Q_t.

B.2 Parabolic averages

The gradient error moves both moment coordinates along the parabola (u, u^2) in (34). For $h \in L^2(\mathbb{R}^2)$, we define

$$\mathcal{P}h(u_1,u_2) = \sup_{\delta>0} \frac{1}{2\delta} \int_{-\delta}^{\delta} |h(u_1+u, u_2+u^2)|\,\mathrm{d}u.$$

The parabolic maximal theorem of Nagel et al. (1976) gives

$$\|\mathcal{P}h\|_{L^2(\mathbb{R}^2)} \le C\|h\|_{L^2(\mathbb{R}^2)}. \tag{40}$$

Since the supremum includes every positive radius, this estimate also holds when δ is restricted to $r_g 2^{-b}$, uniformly in $r_g > 0$. In (34), x stays fixed while (u_1, u_2) changes by (u, u^2). Applying (40) at each fixed x and integrating in x therefore gives the corresponding $L^2(\mathbb{R}^3)$ bound. The linear change of moment coordinates in (34) preserves this operator bound, so it also holds for E_u in the original state coordinates (x, m, v).

Thus the maximal averages for all four write errors in Appendix A.2 satisfy $\|\mathcal{M}\varphi\|_2 \le C\|\varphi\|_2$, uniformly over the allowed widths. Together with the deterministic moment and parameter updates treated there, these estimates give the one-iteration density bound (35), with a constant depending only on β_1, β_2.

C Schedule analysis and fixed write widths

In this section, we prove the invariant bounds in Lemma 4 and the fixed-width comparison in Corollary 3.

C.1 Proof of Lemma 4

Proof. With $\widehat{m}_0 = 0$, the corrected first moment satisfies

$$\widehat{m}_t = \theta_t \widehat{m}_{t-1} + (1-\theta_t)x_{t-1} + \omega_t, \quad \theta_t = \frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}, \quad \omega_t = (1-\theta_t)\left(\xi_t^g + \frac{\xi_t^m}{\gamma_1}\right). \tag{41}$$

Thus $0 \le \theta_t \le \beta_1$ and $|\omega_t| \le \delta_t^g + \delta_t^m/\gamma_1$. In coordinates $(x, \widehat{m} - x)$, the update is given by

$$\begin{pmatrix} x_t \\ \widehat{m}_t - x_t \end{pmatrix} = A(\theta_t, s_t)\begin{pmatrix} x_{t-1} \\ \widehat{m}_{t-1} - x_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix}\xi_t^x + \begin{pmatrix} -s_t \\ 1+s_t \end{pmatrix}\omega_t,$$

where

$$A(\theta, s) = \begin{pmatrix} 1-s & -s\theta \\ s & \theta(1+s) \end{pmatrix}.$$

The norm N_t uses the weight matrix $P = \mathrm{diag}(1, \beta_1)$. For $0 \le \theta \le \beta_1$ and $\eta/(\sigma+\sqrt{V}) \le s \le \eta/\sigma$, direct expansion of $Q = P - A^\top P A$ gives

$$Q_{11} = s[2-(1+\beta_1)s] > 0, \qquad \det Q \ge \beta_1 s(1+\beta_1)[2(1-\beta_1) - (1+\beta_1)s] > 0$$

by (7). Hence Q is positive definite on this compact parameter set, so a common $q \in [1/2, 1)$ satisfies $A^\top P A \preceq q^2 P$. The two error vectors have P norms at most 2 and $H = 1 + 2\eta/\sigma$, respectively, which gives (12) whenever s_t lies in the stated interval.

We choose a fixed $C > 0$ so that the additive term in (12) is at most Cd when all three radii are at most d, and then choose $\bar{\delta} > 0$ such that

$$\left(1 + \frac{1}{\gamma_1}\right)\bar{\delta} \le R, \qquad C\bar{\delta} \le (1-q)R.$$

The support bounds $|\xi_t^s| \le \delta_t^s$ hold simultaneously almost surely. On this event, we prove the state bounds by induction. At $t = 0$, $\widehat{m}_0 = v_0 = 0$ and $N_0 = \sqrt{1+\beta_1}|x_0| \le R$. If the bounds hold at $t-1$, then $|g_t| \le R + \bar{\delta} \le 2R$ and $g_t^2 \le V/2$. Consequently,

$$0 \le v_t \le \beta_2 V(1-\beta_2^{t-1}) + \frac{\gamma_2 V}{2} + \frac{\gamma_2 V}{2} = V(1-\beta_2^t),$$

so

$$\frac{\eta}{\sigma + \sqrt{V}} \le s_t \le \frac{\eta}{\sigma}. \tag{42}$$

The contraction therefore gives $N_t \le qR + C\bar{\delta} \le R$, while (41) gives

$$|\widehat{m}_t| \le 2\theta_t R + (1-\theta_t)\left(R + \bar{\delta} + \frac{\bar{\delta}}{\gamma_1}\right) \le 2R.$$

This closes the induction. Since $|x_t| \le N_t$ and $|m_t| \le |\widehat{m}_t|$, the states belong to $\mathcal{D}$. □

C.2 Proof of Corollary 3

Proof. For any causal policy of finite expected cost, conditional on the full history before the final write of x, the error ξ_T^x has mean zero and variance $(\delta_T^x)^2/3$. Taking expectations and applying Jensen's inequality gives

$$\mathcal{R}_T(\pi) \ge \frac{r_x^2}{3}\mathbb{E}_\pi 2^{-2b_T^x} \ge \frac{r_x^2}{3}2^{-2\mathbb{E}_\pi b_T^x}. \tag{43}$$

For $\pi \in \Pi_T^{\text{fixed}}$ with $\mathcal{R}_T(\pi) \le \varepsilon$, its parameter width b^x is paid at every iteration. Hence

$$\mathcal{B}_T(\pi) \ge T\mathbb{E}_\pi b^x \ge \frac{T}{2}\log\frac{r_x^2}{3\varepsilon}.$$

This also holds when the expected cost is infinite, and proves the lower bound for $\mathcal{C}_{\text{fixed}}(\varepsilon, T)$.

For the upper bound, the write of v uses a fixed positive integer width b^v, independent of ε and T, satisfying $r_v 2^{-b^v} \le \gamma_2 V/2$. To keep the accumulated error below $\sqrt{\varepsilon}/2$, we choose constant widths

$$b^s = \left\lceil \frac{1}{2}\log\frac{1}{\varepsilon} \right\rceil + K, \qquad s \in \{x, g, m\}.$$

We choose the fixed integer $K > 0$ large enough that, for all sufficiently small ε, the radii $\delta_t^s = r_s 2^{-b^s}$ satisfy

$$\delta_t^s \le \bar{\delta}, \qquad s \in \{x,g,m\}, \qquad 2\delta_t^x + H\left(\delta_t^g + \frac{\delta_t^m}{\gamma_1}\right) \le \frac{(1-q)\sqrt{\varepsilon}}{2}.$$

Lemma 4 therefore applies. For $T \ge n_\varepsilon$, (12) and (16) give

$$|x_T| \le N_T \le q^T N_0 + \frac{(1-q)\sqrt{\varepsilon}}{2}\sum_{k=0}^{T-1} q^k \le R\varepsilon + \frac{\sqrt{\varepsilon}}{2} \le \sqrt{\varepsilon}$$

almost surely, which implies $\mathcal{R}_T(\pi) \le \varepsilon$. Each of the three additive widths is $(1/2)\log(1/\varepsilon) + O(1)$, so $\mathcal{B}_T(\pi) = T(b^x + b^g + b^m + b^v) = O(T\log(1/\varepsilon))$. Together with the lower bound, this proves $\mathcal{C}_{\text{fixed}}(\varepsilon, T) = \Theta(T\log(1/\varepsilon))$. Comparison with (8) gives the ratio in Corollary 3. □

D SCALAR STRONGLY CONVEX OBJECTIVES AND SEPARABLE SUMS

In this section, we extend the precision complexity bounds in Theorem 1 to scalar strongly convex smooth objectives under explicit step-size and initialization conditions. We then derive the dimension dependence for separable objectives.

For fixed $0 < \mu \le L$, we assume that $f \in C^1(\mathbb{R})$ satisfies the following conditions:

1. f attains its minimum at zero:
$$f(x) \ge f(0), \qquad \forall x \in \mathbb{R}.$$
2. f is μ-strongly convex:
$$f(y) \ge f(x) + f'(x)(y-x) + \frac{\mu}{2}(y-x)^2, \qquad \forall x, y \in \mathbb{R}.$$
3. f' is L-Lipschitz:
$$|f'(x) - f'(y)| \le L|x-y|, \qquad \forall x, y \in \mathbb{R}.$$

We write $F(x) = f(x) - f(0)$ and replace the exact gradient in (4) by $f'(x_{t-1})$, so the stored gradient is $g_t = f'(x_{t-1}) + \xi_t^g$. Both moments use this stored gradient. The error model, optimizer parameters, and causal policy class are those of Section 2. All fixed parameters and the initialization are independent of ε and T. We denote the precision complexity for this recursion by
$$\mathcal{C}_f(\varepsilon, T) = \inf_{\substack{\pi \in \Pi_T \\ \mathbb{E}_\pi x_T^2 \le \varepsilon}} \mathcal{B}_T(\pi),$$
where $\mathcal{B}_T(\pi)$ is defined in (5). We do not assume that the iterates remain in a prescribed bounded region.

D.1 PRECISION COMPLEXITY FOR SCALAR OBJECTIVES

For the upper bound, we impose the step-size condition
$$0 < \frac{L\eta}{\sigma} \le \frac{\gamma_1}{32}, \qquad 0 < \beta_1, \beta_2 < 1. \tag{44}$$
We choose a fixed $0 < V_f \le \sigma^2$ and define
$$c_0 = \frac{5\eta}{\sigma\gamma_1}, \qquad q_f = \sqrt{1 - \frac{\mu\eta}{4\sigma}}, \qquad R_f = \frac{\sqrt{\mu V_f}}{4L}.$$
We assume that the initialization satisfies
$$0 < |x_0| \le \frac{R_f}{\sqrt{\frac{L}{2} + c_0 L^2}}, \quad m_0 = v_0 = 0. \tag{45}$$
These choices allow the upper-bound construction to keep $\widehat{v}_t \le V_f$ and hence $s_t \in [\eta/(2\sigma), \eta/\sigma]$.

Theorem 9 (Precision complexity at prescribed horizons)**.** *Assume that the step-size and initialization conditions in (44)–(45) hold. For every sufficiently small $\varepsilon > 0$, there exists a threshold $n_{\varepsilon,f} = \Theta(\log(1/\varepsilon))$ such that, for every deterministic horizon $T \ge n_{\varepsilon,f}$,*
$$c\left[T + \log^2\left(\frac{1}{\varepsilon}\right)\right] \le \mathcal{C}_f(\varepsilon, T) \le C\left[T + \log^2\left(\frac{1}{\varepsilon}\right)\right],$$
where $c, C > 0$ are independent of ε and T and depend only on the fixed parameters and initialization. The upper bound is attained by a deterministic policy, and the lower bound applies to every causal policy.

The same conditions also give
$$\inf_{T \ge 1} \mathcal{C}_f(\varepsilon, T) = \Theta\left(\log^2\left(\frac{1}{\varepsilon}\right)\right).$$
Both conclusions hold as well when the target is $\mathbb{E}_\pi F(x_T) \le \varepsilon$ or $\mathbb{E}_\pi |f'(x_T)|^2 \le \varepsilon$, because
$$\frac{\mu}{2}x^2 \le F(x) \le \frac{L}{2}x^2, \qquad \mu^2 x^2 \le |f'(x)|^2 \le L^2 x^2, \qquad \forall x \in \mathbb{R}. \tag{46}$$
The threshold $n_{\varepsilon,f}$ is chosen for each target in (53). We first prove the lower bound, including the restriction on horizons needed when T can be chosen. We then establish a contraction estimate and construct a deterministic precision schedule as in Section 3.

D.2 LOWER BOUND

We first consider an arbitrary deterministic horizon T and policy $\pi \in \Pi_T$. Since $|f'(x)| \le L|x|$, the proof of Lemma 6 applies with $H_t = 1+r_g/2+L\max_{0\le k\le t}|x_k|$. Estimates (28) and (30) then give $H_t \le H_{t-1} + CLH_{t-1}^{1-1/h}$, and hence $H_t \le C(1+t)^h$, with the same exponent h. Consequently, $|m_t| \le C_m(1+t)^h$ and $v_t \le C_v(1+t)^{2h}$, where the constants may also depend on L.

For the density estimate, the moment update with gradient $f'(x)+u$ is $E_u \circ D$, where

$$D(x,m,v) = (x, \beta_1 m + \gamma_1 f'(x), \beta_2 v + \gamma_2 f'(x)^2),$$
$$E_u(x,m,v) = (x, m+\gamma_1 u, v + \gamma_2(2f'(x)u + u^2)).$$

At each fixed x, D has the same linear part in (m,v) as in Appendix A.2. The coordinates $u_1 = m/\gamma_1$ and $u_2 = v/\gamma_2 - 2f'(x)m/\gamma_1$ give

$$E_u : (x,u_1,u_2) \longmapsto (x, u_1+u, u_2+u^2).$$

Thus the same parabola bound applies without differentiating f'. With the same conditioning on positive errors in the writes of v, the remaining state maps are unchanged. Together with the state bounds, this gives (22) with the same constants G, D, h and a fixed $0 < \zeta \le x_0^2$. The summation in Proposition 8 therefore yields

$$\mathcal{B}_T(\pi) \ge c\left[T + \log^2\left(\frac{1}{\varepsilon}\right)\right]$$

whenever $\mathbb{E}_\pi x_T^2 \le \varepsilon$ and $T \ge \lceil \log(1/\varepsilon)/D \rceil$, for sufficiently small ε. This prescribed-horizon lower bound requires neither (44) nor the upper bound on $|x_0|$ in (45).

To extend the lower bound to all deterministic horizons, we now use (44). We define $y_t = 2x_t - \widehat{m}_t/L$, with $\widehat{m}_0 := 0$, and $\nu(x) = f'(x)/(Lx)$ for $x \ne 0$. Strong convexity and smoothness give $\nu(x) \in [\mu/L, 1]$. We set $\nu(0) = \mu/L$. The corrected moment recursion is $\widehat{m}_t = \theta_t \widehat{m}_{t-1} + (1-\theta_t)f'(x_{t-1}) + \omega_t$, where

$$\theta_t = \frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}, \qquad \omega_t = (1-\theta_t)\xi_t^g + \frac{\xi_t^m}{1-\beta_1^t}.$$

With $C_t = 2\theta_t + (1-\theta_t)\nu(x_{t-1})$, the update becomes

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 1 - Ls_tC_t & Ls_t\theta_t \\ 2 - (1+2Ls_t)C_t & \theta_t(1+2Ls_t) \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix} \xi_t^x - \begin{pmatrix} Ls_t \\ 1+2Ls_t \end{pmatrix} \frac{\omega_t}{L}. \tag{47}$$

Since $0 \le C_t \le 1+\theta_t$, condition (44) makes the matrix nonnegative for every $0 \le s_t \le \eta/\sigma$. In particular,

$$2 - (1+2Ls_t)C_t \ge \gamma_1 - 2\frac{L\eta}{\sigma}(1+\beta_1) > 0.$$

Its first diagonal entry is at least $\lambda_f = 1 - (L\eta/\sigma)(1+\beta_1) > 1/2$. For $x_0 > 0$, the event $\xi_t^x \ge 0$, $\xi_t^g \le 0$ and $\xi_t^m \le 0$ for all $1 \le t \le T$ has probability 8^{-T}. On this event, $\omega_t \le 0$, so the additive vector in (47) is nonnegative. Starting from $(x_0, y_0) = (x_0, 2x_0) \ge 0$, we obtain $x_t, y_t \ge 0$ and $x_t \ge \lambda_f x_{t-1}$ by induction. Hence $x_T \ge \lambda_f^T x_0$. For $x_0 < 0$, we reverse the three signs and apply the same argument to $(-x_t, -y_t)$. In either case,

$$\mathbb{E}_\pi x_T^2 \ge x_0^2 \left(\frac{\lambda_f^2}{8}\right)^T, \qquad \log\left(\frac{8}{\lambda_f^2}\right) < 5 < D.$$

If $\mathbb{E}_\pi x_T^2 \le \varepsilon$, then, for sufficiently small ε,

$$T \ge \left\lceil \frac{\log(x_0^2/\varepsilon)}{\log(8/\lambda_f^2)} \right\rceil \ge \left\lceil \frac{\log(1/\varepsilon)}{D} \right\rceil.$$

Thus every policy with $\mathbb{E}_\pi x_T^2 \le \varepsilon$ meets the horizon condition for the preceding lower bound. Taking the infimum over policies and deterministic horizons gives $\inf_{T\ge1} \mathcal{C}_f(\varepsilon, T) \ge c\log^2(1/\varepsilon)$. The comparisons in (46) give the same lower bounds for the other two targets after rescaling ε by fixed constants.

D.3 Contraction without write errors

For the upper bound, we control the objective and the error in the first moment through

$$\mathcal{V}(x, \widehat{m}) = F(x) + c_0[\widehat{m} - f'(x)]^2.$$

The next lemma gives a contraction factor uniform over the effective steps allowed by the second-moment bound.

Lemma 10. *Assume that the step-size condition in (44) holds. For every $\theta \in [0, \beta_1]$ and $s \in [\eta/(2\sigma), \eta/\sigma]$, the noiseless update*

$$\widehat{m}^+ = \theta\widehat{m} + (1-\theta)f'(x), \qquad x^+ = x - s\widehat{m}^+$$

satisfies

$$\mathcal{V}(x^+, \widehat{m}^+) \le q_f^2 \mathcal{V}(x, \widehat{m}), \qquad \forall (x, \widehat{m}) \in \mathbb{R}^2. \tag{48}$$

Proof. We write $g = f'(x)$, $e = \widehat{m} - g$ and $e^+ = \widehat{m}^+ - f'(x^+)$. Then $\widehat{m}^+ = g + \theta e$. Smoothness and the bounds $s^2 \le (\eta/\sigma)s$ and $|\theta s g e| \le (s/4)g^2 + s\theta^2 e^2$ give

$$\begin{aligned} F(x^+) &\le F(x) - sg(g+\theta e) + \frac{Ls^2}{2}(g+\theta e)^2 \\ &\le F(x) - s\left(\frac{3}{4} - \frac{L\eta}{\sigma}\right) g^2 + s\left(1 + \frac{L\eta}{\sigma}\right)\theta^2 e^2. \end{aligned}$$

Since $L\eta/\sigma \le 1/32$ and $s \ge \eta/(2\sigma)$, we obtain

$$F(x^+) \le F(x) - \frac{\eta}{4\sigma} g^2 + \frac{5\eta}{4\sigma}\beta_1^2 e^2. \tag{49}$$

For the tracking error, Lipschitz continuity gives

$$|e^+| \le \theta|e| + Ls|g + \theta e| \le \beta_1\left(1 + \frac{L\eta}{\sigma}\right)|e| + \frac{L\eta}{\sigma}|g|.$$

Condition (44) implies $\beta_1(1 + L\eta/\sigma) \le 1 - 3\gamma_1/4$. Young's inequality with parameter $\gamma_1/4$ then gives

$$|e^+|^2 \le \left(1 - \frac{\gamma_1}{2}\right) e^2 + \frac{5L^2\eta^2}{\gamma_1\sigma^2} g^2. \tag{50}$$

Multiplying (50) by $c_0 = 5\eta/(\sigma\gamma_1)$ and adding (49), we obtain

$$\begin{aligned} \mathcal{V}(x^+, \widehat{m}^+) &\le \mathcal{V}(x, \widehat{m}) + \left(-\frac{\eta}{4\sigma} + \frac{25L^2\eta^3}{\gamma_1^2\sigma^3}\right) g^2 + \left(-\frac{\gamma_1 c_0}{2} + \frac{5\eta\beta_1^2}{4\sigma}\right) e^2 \\ &\le \mathcal{V}(x, \widehat{m}) - \frac{\eta}{8\sigma} g^2 - \frac{\gamma_1 c_0}{4} e^2 \\ &\le \left(1 - \frac{\mu\eta}{4\sigma}\right) \mathcal{V}(x, \widehat{m}) = q_f^2 \mathcal{V}(x, \widehat{m}). \end{aligned}$$

The second inequality follows from (44). The last uses $g^2 \ge 2\mu F(x)$ and $\mu\eta/\sigma \le \gamma_1$. □

D.4 Perturbed contraction and invariant bounds

We now include the write errors. The second-moment bound keeps s_t in the interval required by Lemma 10, and the perturbed contraction keeps the state in the region used to prove that bound. We write $\mathcal{V}_t = \mathcal{V}(x_t, \widehat{m}_t)$, with $\widehat{m}_0 = 0$.

Lemma 11. *Assume that the step-size and initialization conditions in (44)–(45) hold. There exist constants $K_f, \bar{\delta}_f > 0$, depending only on $\mu, L, \eta, \sigma, \beta_1, V_f$, such that any causal policy satisfying*

$$\delta_t^v \le \frac{\gamma_2 V_f}{2}, \qquad \delta_t^s \le \bar{\delta}_f, \qquad s \in \{x, g, m\}, \quad 1 \le t \le T,$$

obeys, almost surely,

$$\sqrt{\mathcal{V}_t} \le R_f, \qquad 0 \le v_t \le V_f(1 - \beta_2^t), \qquad 0 \le t \le T,$$

and

$$\sqrt{\mathcal{V}_t} \le q_f\sqrt{\mathcal{V}_{t-1}} + K_f \max_{s\in\{x,g,m\}} \delta_t^s, \qquad 1 \le t \le T. \tag{51}$$

Proof. We first assume $s_t \in [\eta/(2\sigma), \eta/\sigma]$ and compare the actual update with the noiseless update $(x^+, \widehat{m}^+)$ from the same state, using the same realized s_t and θ_t. Then

$$\widehat{m}_t - \widehat{m}^+ = \omega_t, \qquad x_t - x^+ = \xi_t^x - s_t\omega_t,$$

so the difference between the moment errors is at most $(1 + Ls_t)|\omega_t| + L|\xi_t^x|$.

Also, $|f'|^2 \le 2LF$ makes $\sqrt{F}$ a $\sqrt{L/2}$-Lipschitz function, with the bound extended across zero by continuity. Applying the triangle inequality to $(\sqrt{F}, \sqrt{c_0}[\widehat{m} - f'(x)])$ and using $|\omega_t| \le |\xi_t^g| + |\xi_t^m|/\gamma_1$ gives a fixed $K_f > 0$ such that

$$\sqrt{\mathcal{V}_t} \le \sqrt{\mathcal{V}(x^+, \widehat{m}^+)} + K_f \max_{s\in\{x,g,m\}} |\xi_t^s| \le q_f\sqrt{\mathcal{V}_{t-1}} + K_f \max_{s\in\{x,g,m\}} |\xi_t^s|.$$

The last inequality uses Lemma 10, which holds uniformly in s_t. We choose $\bar{\delta}_f > 0$ small enough that

$$\bar{\delta}_f \le \sqrt{\frac{V_f}{8}}, \qquad K_f\bar{\delta}_f \le (1 - q_f)R_f.$$

The error bounds $|\xi_t^s| \le \delta_t^s$ hold simultaneously almost surely. On these paths, we prove the state bounds by induction. At $t = 0$, assumption (45) gives

$$\mathcal{V}_0 = F(x_0) + c_0|f'(x_0)|^2 \le \left(\frac{L}{2} + c_0L^2\right) x_0^2 \le R_f^2, \qquad v_0 = 0.$$

Assuming the bounds at $t - 1$, strong convexity and the choice of R_f give

$$|f'(x_{t-1})| \le L\sqrt{\frac{2}{\mu}}\sqrt{\mathcal{V}_{t-1}} \le L\sqrt{\frac{2}{\mu}}R_f = \sqrt{\frac{V_f}{8}}.$$

Since $|\xi_t^g| \le \bar{\delta}_f \le \sqrt{V_f/8}$, it follows that $|g_t| \le \sqrt{V_f/2}$. Hence

$$0 \le v_t \le \beta_2 v_{t-1} + \gamma_2 g_t^2 + \delta_t^v \le \beta_2 V_f(1 - \beta_2^{t-1}) + \gamma_2 V_f = V_f(1 - \beta_2^t).$$

Thus $\widehat{v}_t \le V_f \le \sigma^2$, which places s_t in $[\eta/(2\sigma), \eta/\sigma]$. The perturbation estimate above now proves (51) and gives

$$\sqrt{\mathcal{V}_t} \le q_f R_f + K_f\bar{\delta}_f \le R_f.$$

This closes the induction. □

D.5 Upper bound

As in Section 3, we keep the second-moment width fixed and increase the other three widths near the end of the run. We first choose positive integers $b_{0,f}, b_{v,f}$, independently of ε and T, such that

$$r_v 2^{-b_{v,f}} \le \frac{\gamma_2 V_f}{2}, \qquad r_s 2^{-b_{0,f}} \le \bar{\delta}_f, \qquad s \in \{x, g, m\}. \tag{52}$$

These widths satisfy the radius conditions in Lemma 11.

We define $\chi_f = -\log q_f > 0$. To reach each target, we choose $n_{\varepsilon,f}$ as the smallest integer $n \ge 2$ such that

$$q_f^n(R_f + nK_f\bar{\delta}_f) \le \begin{cases} \sqrt{\frac{\mu\varepsilon}{2}}, & \text{parameter MSE,} \\ \sqrt{\varepsilon}, & \text{excess objective,} \\ \sqrt{\frac{\varepsilon}{2L}}, & \text{squared gradient.} \end{cases} \tag{53}$$

For a prescribed $T \ge n_{\varepsilon,f}$, we use the deterministic policy π given by the schedule

$$b_t^s = b_{0,f} + \lceil \chi_f \max\{0, t - (T - n_{\varepsilon,f})\} \rceil, \qquad b_t^v = b_{v,f}, \qquad s \in \{x, g, m\}, \quad 1 \le t \le T.$$

Lemma 11 gives $\sqrt{\mathcal{V}_{T-n_{\varepsilon,f}}} \le R_f$ almost surely. At time $t = T - n_{\varepsilon,f} + j$, the radii satisfy

$$\delta^s_{T-n_{\varepsilon,f}+j} = r_s 2^{-b_{0,f}-\lceil \chi_f j \rceil} \le \bar{\delta}_f q_f^j, \qquad 1 \le j \le n_{\varepsilon,f}, \qquad \forall s \in \{x, g, m\}.$$

Substituting these radii into (51) gives

$$\sqrt{\mathcal{V}_{T-n_{\varepsilon,f}+j}} \le q_f \sqrt{\mathcal{V}_{T-n_{\varepsilon,f}+j-1}} + K_f \bar{\delta}_f q_f^j.$$

Applying this inequality over $j = 1, \dots, n_{\varepsilon,f}$ gives

$$\sqrt{\mathcal{V}_T} \le q_f^{n_{\varepsilon,f}} R_f + K_f \bar{\delta}_f \sum_{j=1}^{n_{\varepsilon,f}} q_f^{n_{\varepsilon,f}-j} q_f^j = q_f^{n_{\varepsilon,f}} (R_f + n_{\varepsilon,f} K_f \bar{\delta}_f).$$

The choice in (53) gives the corresponding target almost surely, since

$$x_T^2 \le \frac{2\mathcal{V}_T}{\mu}, \qquad F(x_T) \le \mathcal{V}_T, \qquad |f'(x_T)|^2 \le 2LF(x_T) \le 2L\mathcal{V}_T.$$

Condition (53) requires $q_f^n R_f = O(\sqrt{\varepsilon})$, so $n_{\varepsilon,f} = \Omega(\log(1/\varepsilon))$. Conversely, $n = \lceil \log(1/\varepsilon)/\chi_f \rceil$ makes its left-hand side $O(\varepsilon \log(1/\varepsilon))$, which satisfies each target for sufficiently small ε. Thus $n_{\varepsilon,f} = \Theta(\log(1/\varepsilon))$. Since the widths are deterministic, their total cost is

$$\begin{aligned}
\mathcal{B}_T(\pi) &= \sum_{t=1}^{T} (b_t^x + b_t^g + b_t^m + b_t^v) \\
&= (3b_{0,f} + b_{v,f})T + 3 \sum_{j=1}^{n_{\varepsilon,f}} \lceil \chi_f j \rceil \\
&= O(T + n_{\varepsilon,f}^2) = O\left(T + \log^2\left(\frac{1}{\varepsilon}\right)\right).
\end{aligned}$$

Together with the lower bounds, this proves the fixed-horizon claim in Theorem 9 and the corresponding bounds for the other two targets. Choosing $T = n_{\varepsilon,f}$ gives $\inf_{T \ge 1} \mathcal{C}_f(\varepsilon, T) \le C \log^2(1/\varepsilon)$. The lower bound proved above holds for every policy meeting the target at any deterministic horizon, so it also proves the claimed order when the horizon is optimized.

D.6 Dimension dependence for separable objectives

We first consider the separable quadratic $f(x) = \|x\|_2^2/2$. We denote by $\mathcal{C}_d(\varepsilon, T)$ the infimum of the expected cumulative precision cost, summed over all coordinates, among causal policies satisfying $\mathbb{E}_\pi \|x_T\|_2^2 \le \varepsilon$. Each write receives fresh dither after its width is selected, and the selection may use the full revealed history across coordinates.

Corollary 12 (Precision complexity at prescribed horizons)**.** *For $f(x) = \|x\|_2^2/2$ in d dimensions, suppose every coordinate has the same fixed initialization $x_{0,i} = x_0 \ne 0$, zero initial moments, and common optimizer parameters and error scales satisfying (1) and (7). For every sufficiently small $\varepsilon/d > 0$, the threshold $n_{\varepsilon/d} = \Theta(\log(d/\varepsilon))$ in Theorem 1 has the following property. For every deterministic horizon $T \ge n_{\varepsilon/d}$,*

$$c\left[dT + d\log^2\left(\frac{d}{\varepsilon}\right)\right] \le \mathcal{C}_d(\varepsilon, T) \le C\left[dT + d\log^2\left(\frac{d}{\varepsilon}\right)\right], \tag{54}$$

where $c, C > 0$ and the smallness threshold for ε/d are independent of d, T, ε. The upper bound is attained by a deterministic policy, and the prescribed-horizon lower bound does not require (7).

If the stronger step-size condition in (9) also holds, then

$$\inf_{T \ge 1} \mathcal{C}_d(\varepsilon, T) = \Theta\left(d\log^2\left(\frac{d}{\varepsilon}\right)\right).$$

Proof. We fix a causal policy π. For coordinate i, we write $\mathcal{R}_{T,i}(\pi) = \mathbb{E}_\pi x_{T,i}^2$ and denote its expected cost by $\mathcal{B}_{T,i}(\pi)$, so that $\mathcal{B}_T(\pi) = \sum_{i=1}^d \mathcal{B}_{T,i}(\pi)$. Given the full history before a write, its fresh error still has the prescribed conditional uniform law. Thus the scalar density argument applies

to each coordinate even when its widths depend on past writes in other coordinates. Conditioning on positive second-moment errors only in coordinate i gives the same scalar constants, so (22) yields

$$\mathcal{B}_{T,i}(\pi) \geq 4T + \frac{1}{16}\sum_{k=0}^{T-1}\left[\log\frac{\zeta}{\mathcal{R}_{T,i}(\pi)} - 6h\log(1+T) - Dk\right]_+, \qquad \forall i \in \{1,\dots,d\}.$$

For each fixed k, T, $u \mapsto [\log(\zeta/u) - 6h\log(1+T) - Dk]_+$ is convex on $(0,\infty)$, being the maximum of a convex function and zero. Summing over coordinates and applying Jensen's inequality gives

$$\mathcal{B}_T(\pi) \geq 4dT + \frac{d}{16}\sum_{k=0}^{T-1}\left[\log\frac{\zeta d}{\mathbb{E}_\pi\|x_T\|_2^2} - 6h\log(1+T) - Dk\right]_+. \tag{55}$$

This is d times (22), with scalar risk $\mathbb{E}_\pi\|x_T\|_2^2/d$. If $\mathbb{E}_\pi\|x_T\|_2^2 \leq \varepsilon$, the summation argument in Proposition 8 therefore gives

$$\mathcal{B}_T(\pi) \geq cd\left[T + \log^2\left(\frac{d}{\varepsilon}\right)\right], \qquad T \geq \left\lceil\frac{\log(d/\varepsilon)}{D}\right\rceil,$$

for sufficiently small ε/d, with c and the smallness threshold independent of d. Proposition 8 also verifies that $n_{\varepsilon/d}$ meets this horizon condition. For the upper bound, we run the schedule of Proposition 5 with target ε/d in every coordinate. The total risk is at most ε, and the costs add to $O(dT + d\log^2(d/\varepsilon))$ for $T \geq n_{\varepsilon/d}$. If (9) also holds, we can remove the prescribed horizon condition. Applying (37) to each coordinate gives

$$\mathbb{E}_\pi\|x_T\|_2^2 \geq dx_0^2\left(\frac{\lambda^2}{8}\right)^T.$$

If $\mathbb{E}_\pi\|x_T\|_2^2 \leq \varepsilon$, the logarithmic comparison in (38) gives $T \geq \lceil\log(d/\varepsilon)/D\rceil$ for sufficiently small ε/d. The preceding lower bound therefore applies at every such horizon. Choosing $T = n_{\varepsilon/d}$ gives the matching upper bound. □

For $f(x) = \sum_{i=1}^d f_i(x_i)$, we assume that each f_i satisfies the conditions at the start of this section with common curvature bounds μ, L. We use common optimizer parameters and error scales satisfying (44), a common choice of V_f, and zero initial moments. The initialization $x_{0,i} = x_0 \neq 0$ is independent of i, d and satisfies (45). The scalar lower bound and construction above have constants uniform over these objectives. Applying the same coordinatewise lower-bound argument and using target ε/d in each coordinate gives (54) for $T \geq n_{\varepsilon/d,f}$, as well as the corresponding bound over optimized horizons. Summing (46) over coordinates gives the same cost orders for total expected excess objective and squared gradient, with the scalar threshold chosen for the corresponding target at accuracy ε/d.

In these families, $\|x_0^{(d)}\|_2^2 = dx_0^2$. The constants in (54) are uniform in d for this fixed coordinate initialization. If the total initial norm is fixed and the coordinates shrink with d, the scalar constant ζ may depend on d, so the same uniformity does not follow.

E Numerical protocol and additional results

In this section, we describe the experiments on the constructive schedule and its cost at finite accuracies. The lower bound over every causal policy is a mathematical conclusion of Proposition 8; a finite collection of simulated policies cannot establish that quantifier.

E.1 Oracle implementation and measured quantities

We specify the numerical recursion and the risk and cost measurements used throughout the experiments. We implement the scalar quadratic recursion (4) in NumPy float64, with write order $g \to m \to v \to x$, both bias corrections, and projection only of the second moment onto $[0,\infty)$.

Each width is fixed before its own independent uniform draw. The exact gradient before its numerical write is x_{t-1}, with no stochastic-gradient noise. The objective, damping, scales and hyperparameters are identical across compared strategies. The default configuration is

$$\beta_1 = 0.9, \quad \beta_2 = 0.99, \quad \sigma = 1, \quad \eta = 1/19,$$
$$x_0 = 1/4, \quad m_0 = v_0 = 0, \quad r_x = r_g = r_m = r_v = 1.$$

Here η/σ is one half of the upper-construction step threshold. It exceeds the stronger bound (9), so the default experiments concern prescribed horizons, not optimized-horizon complexity.

For independent trajectory outcomes $Y_i = x_{T,i}^2$, we report

$$\widehat{\mathcal{R}}_T = \frac{1}{N}\sum_{i=1}^{N} Y_i, \qquad \mathrm{SE} = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \widehat{\mathcal{R}}_T)^2}{N(N-1)}}.$$

The target is mean-square error, not root-mean-square error or $f(x_T)$. The quantity $\widehat{\mathcal{R}}_T + 2\mathrm{SE}$ is a normal-approximation diagnostic, without a finite-sample or simultaneous coverage guarantee. For deterministic schedules, the summed write cost is exact and has no Monte Carlo standard error. All four writes, including the gradient and the constant second moment, are included in this sum.

The implementation does not clip x or m or enforce a fixed upper bound on v. It records observed state maxima, the frequency of the $v = 0$ projection, and exceptional paths. The diagnostic threshold is $|x| > 10$, $|m| > 10$, or $v > 100$; these empirical checks describe sampled trajectories and do not establish bounds independent of horizon. No selected holdout trajectory triggers these thresholds. Across the calibrated settings, the observed maxima are 4.546, 4.651 and 14.001 for $|x|$, $|m|$ and v, respectively; projection frequencies range from 0.0000185 to 0.1159. Per-setting diagnostics are in the supplement's `aggregate_results.csv` files (Appendix E.6). In particular, the calibrated floors below are not asserted to equal the sufficient floors in Lemma 4. The simulations isolate the dithered numerical model and do not measure physical storage, runtime savings or training performance on general objectives.

E.2 Candidate policies and independent evaluation

We calibrate terminal-ramp and fixed-width candidates on pilot paths, then evaluate the selected policies on independent holdout paths. Writing $L_\varepsilon = \log(1/\varepsilon)$, the main accuracy sweep uses $L_\varepsilon \in \{4, 6, \ldots, 24\}$ and $T \in \{24L_\varepsilon, 48L_\varepsilon\}$. For each target and horizon, the terminal-ramp candidates are

$$b_t^x = b_t^g = b_t^m = 3 + \lceil \alpha[t - (T - n + 1)]_+ \rceil, \qquad b_t^v = 6. \tag{56}$$

We search $\alpha \in \{0.025, 0.05, 0.1\}$ and $n \in \{1, T\} \cup \{2, 8, 14, \ldots\} \cap [1, T]$, discarding candidates with any width greater than 30 and deduplicating identical schedules. The ramp duration n and slope α are empirical choices; they are not the certified n_ε and χ from Section 3.

The selection procedure is as follows.

1. Evaluate all candidates on a common 512-trajectory pilot sample. A candidate qualifies when $\widehat{\mathcal{R}}_T + 2\mathrm{SE} \le 0.9\varepsilon$. For ramps, refine the grid at every integer duration within six steps of the three cheapest pilot qualifiers. If none qualifies, refine around the candidate with the smallest pilot diagnostic. Evaluate new candidates using the same pilot stream and all three slopes.
2. Re-evaluate the six cheapest pilot qualifiers on an independent 20,000-trajectory confirmation sample. If fewer than six qualify, use all of them; if none qualifies, confirm up to six candidates with the smallest pilot diagnostic. Choose the cheapest confirmation qualifier under the same 0.9ε threshold. If none qualifies, retain the candidate with the smallest confirmation diagnostic and record selection failure.
3. Evaluate the selected candidate exactly once on a fresh 20,000-trajectory holdout. Record its outcome whether it passes or fails $\widehat{\mathcal{R}}_T + 2\mathrm{SE} \le \varepsilon$. No holdout outcome changes the selected schedule or triggers a new search.

The finite grid, pilot screening and six-candidate shortlist may miss cheaper feasible schedules. Reported costs describe the selected policies, not a global optimum or a sample-based lower bound.

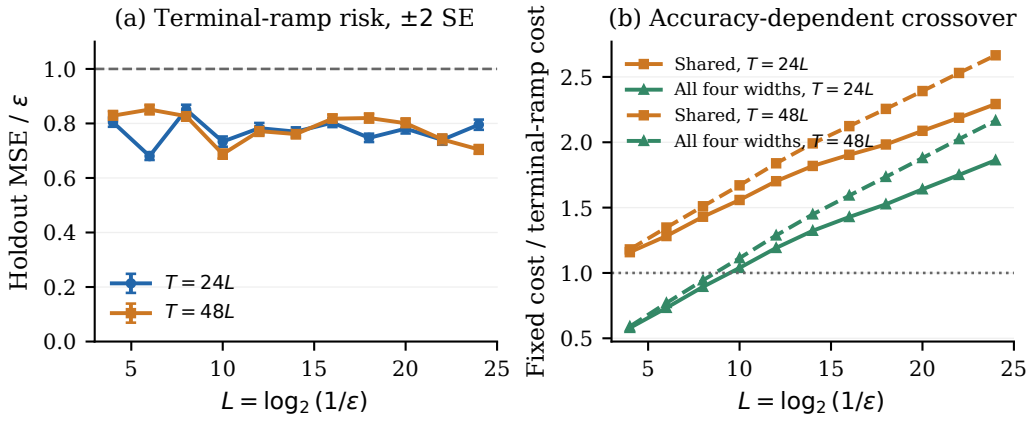

Figure 3: Holdout risk and accuracy-dependent costs. Left: mean MSE divided by its target with ±2 standard errors for the 22 selected ramps. Right: fixed-format cost divided by ramp cost at identical targets and horizons. The four-channel fixed candidate search varies all four widths. It is cheaper at $L = 4$; temporal allocation becomes cheaper as the target tightens.

The fixed-shared family has $b^x = b^g = b^m \in \{1, \dots, 24\}$ and $b^v = 6$. After selecting its width b, the fixed-channelwise family searches (b^x, b^g, b^m) independently in $\{\max(1, b-3), \dots, \min(24, b+1)\}$, also with $b^v = 6$. Each family uses the same pilot/confirmation/holdout procedure. These two archived baselines do not optimize the second-moment width. For the stronger four-channel fixed baseline, we first pilot-test all shared widths $b^x = b^g = b^m \in \{1, \dots, 24\}$ at each $b^v \in \{1, 2, 4, 6, 8, 12\}$. For each b^v, let b be the least shared width passing the pilot threshold, or the one with the smallest pilot diagnostic if none passes. We then search b^x, b^g, b^m independently in $\{\max(1, b-3), \dots, \min(24, b+1)\}$ at that b^v. After deduplication, this gives 864 candidates per target and horizon. The independent confirmation and holdout stages are exactly as above; the pilot outcomes alone determine the candidate grid. Thus all four write widths vary across the searched family, but this local finite search is not the infimum over the class in Corollary 3.

For the fixed-target horizon comparison, $\varepsilon = 2^{-20}$ and $T \in \{480, 960, 1920, 3840\}$. Dynamic results from an identical accuracy-sweep configuration are reused rather than counted as independent measurements. Across all experiments, repeated settings and shared random streams mean that rows in the aggregate table are not independent replications.

The random generator is NumPy `default_rng` (PCG64). The global seed is 1729. For each horizon and stage we create `SeedSequence([1729,T,stage])` and use the first value from `generate_state(1)` as the seed. Stages 1, 2 and 3 denote pilot, confirmation and holdout, respectively. Candidates and strategies at the same horizon use common random numbers within each stage. Trajectory visualizations use stage 4 and 1,000 paths. The original five suites used Python 3.13.15, NumPy 2.3.5, pandas 3.0.1, Matplotlib 3.11.2 and PyYAML 6.0.3. The four-channel search used Python 3.12.14 with the same NumPy, pandas and Matplotlib versions and PyYAML 6.0. Eleven simulator and search tests cover deterministic updates, write ordering, integer widths and holdout isolation.

E.3 Scaling and the fixed-format cost crossover

We compare the finite-accuracy cost of the selected ramps with that of fixed widths at the same targets and horizons.

All 22 selected ramps pass the holdout diagnostic (Figure 3). The ordinary least-squares fit to $B = aT + bL_\varepsilon^2 + c$ gives

$$a = 14.9915, \qquad b = 6.18394, \qquad c = -65.8678, \qquad R^2 = 0.9999216.$$

The coefficient of T is close to the coarse cost $3 \cdot 3 + 6 = 15$. Figure 2(a) subtracts that baseline to expose the accuracy-dependent overhead. The fit describes a finite window with searched integer

schedules; its high R^2 includes the linear horizon term and does not establish an asymptotic exponent or an optimal constant.

Table 1 reports the strictest accuracy point at the longer sweep horizon. The comparison matches the target, not the achieved MSE: discrete fixed widths overshoot the target more than the selected ramp. At $L_\varepsilon = 4$, the four-channel fixed-to-ramp cost ratios are 0.579 and 0.590 at $T = 96$ and 192. The ratio is 2.166 at $L_\varepsilon = 24$, $T = 1152$. At the separate fixed target 2^{-20}, increasing T from 480 to 3840 raises it from 1.640 to 2.110. The searched four-channel baseline is stronger than the earlier fixed-$b^v = 6$ channelwise baseline, but remains a finite-grid comparison.

Table 1: Target $\varepsilon = 2^{-24}$ and horizon $T = 1152$. MSE columns are in units of 10^{-8}; the target is 5.9605 in these units. Each mean and SE uses 20,000 independent holdout trajectories. The last two rows show the effect of allowing b^v to vary across fixed candidates.

Policy	Cost	Mean MSE	SE	Mean +2 SE
Terminal ramp	20,745	4.2004	0.0479	4.2961
Fixed shared	55,296	2.2846	0.0227	2.3300
Fixed channelwise, $b^v = 6$	50,688	3.2816	0.0331	3.3477
Fixed all four, $b^v = 1$	44,928	3.8082	0.0385	3.8851

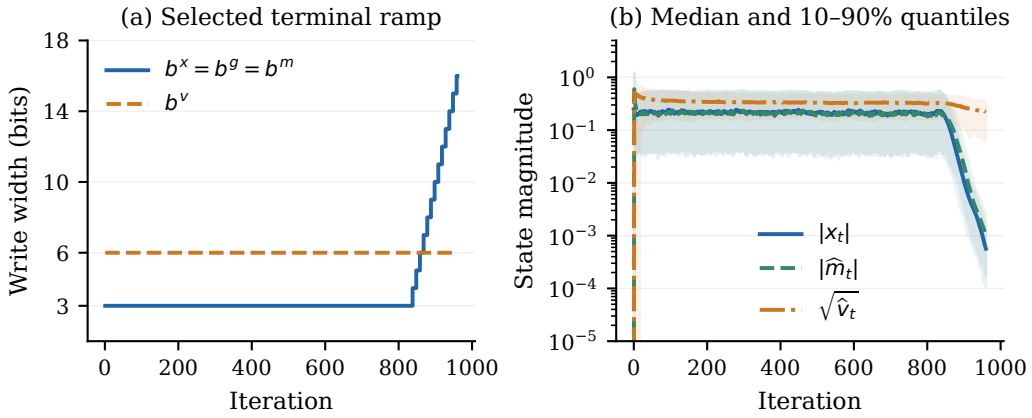

Figure 4: Selected schedule and trajectories at $\varepsilon = 2^{-20}$, $T = 960$, with $\alpha = 0.1$, $n = 124$, $b_0 = 3$ and $b^v = 6$. State curves are medians; bands are pointwise 10–90% quantiles from 1,000 independent visualization trajectories, not confidence intervals.

Figure 4 shows a representative selected ramp. Its first precision increment occurs after iteration 837; $b^x = b^g = b^m$ reaches 16 bits at the final step while b^v stays at six. The schedule costs 16,857 at this target and horizon.

E.4 Constant second-moment precision and recalibration

We test whether a constant second-moment width remains economical after recalibrating the other three widths for each target. For each $L_\varepsilon \in \{8, 16, 24\}$ at $T = 48L_\varepsilon$, we test constant $b^v \in \{1, 2, 4, 6, 8, 12\}$ and a schedule that ramps all four widths, starting the second moment at six bits. The slope and duration of the all-channel ramp are shared across all four writes, so its second-moment width is $b_t^v = 6 + \lceil \alpha[t-(T-n+1)]_+ \rceil$. The slope and duration of the three additive ramps are searched independently for each choice of second-moment precision. Thus this comparison measures the cost of reaching a common target after recalibration, rather than the isolated effect of changing b^v while holding the other writes fixed.

All tested constant widths, including one bit, pass at all three targets. At the strictest target, the cheapest tested second-moment choice is two bits, with a longer additive ramp than the six-bit policy (Table 2). Keeping six bits costs about 4.8% less than ramping all channels. These observations

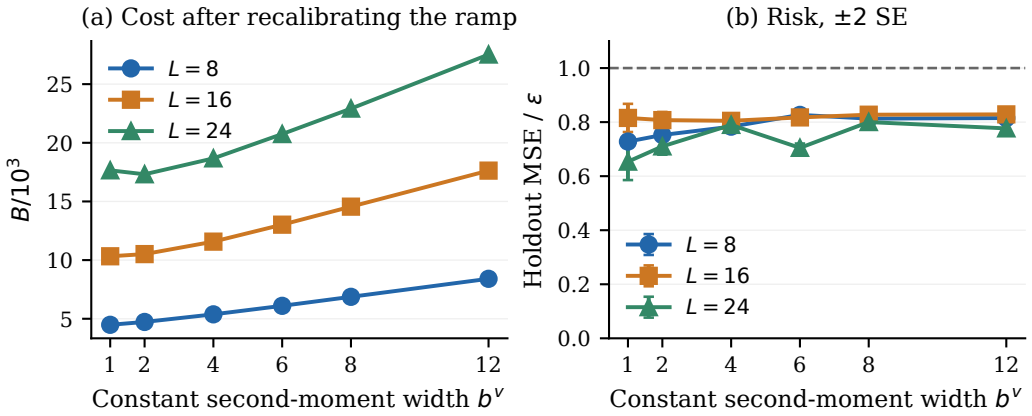

Figure 5: Constant second-moment precision after independent recalibration of the other three ramps. Each of the 18 constant-width settings passes the holdout mean-plus-two-SE diagnostic. The three all-channel ramps also pass. The nonmonotone cost curve reflects both the direct v cost and the changed additive ramp selected at each width.

Table 2: Second-moment comparison at $\varepsilon = 2^{-24}$, $T = 1152$. The last two columns are normalized by ε; all rows use 20,000 holdout trajectories. The three additive floors are three bits.

v width	α	n	Cost	Mean MSE/ε	2 SE/ε
1	0.05	277	17,652	0.654	0.068
2	0.05	240	17,316	0.710	0.031
4	0.1	153	18,672	0.790	0.022
6	0.1	148	20,745	0.705	0.016
8	0.1	145	22,914	0.800	0.019
12	0.1	145	27,522	0.777	0.018
Ramped from 6	0.1	146	21,780	0.776	0.019

support the usefulness of constant second-moment precision in this regime; they do not estimate the minimum floor required by the pathwise invariant proof.

E.5 Step, momentum and scale sensitivity

We vary the step size, moment parameters, and error scales to examine the selected schedule after recalibration. We test η/σ at 10%, 25% and 50% of the threshold $2(1-\beta_1)/(1+\beta_1)$ with the default moment parameters, $(\beta_1, \beta_2) = (0.8, 0.98)$ at 25% of this threshold, and asymmetric scales $(r_x, r_g, r_m, r_v) = (0.5, 2, 0.5, 2)$ at 50% of this threshold. All variants retain $b_0 = 3$, $b^v = 6$, $\sigma = 1$ and $x_0 = 1/4$. If ρ is the ratio of the new η/σ to its default value, we set $T = \lceil 24L_\varepsilon/\rho \rceil$, multiply the slope grid by ρ, and use ramp-duration stride $\max\{1, \mathrm{round}(6/\rho)\}$, with local refinement within that stride. This covers comparable nominal contraction times. Horizons therefore differ across variants, and their costs are not a same-horizon comparison.

At the smallest step and strictest target, the selected slope is 0.01 and the ramp duration is 1,183 of the 2,880 iterations. Its holdout diagnostic is $5.2925 \times 10^{-8} < 2^{-24}$. The longer ramp is consistent with slower attenuation of early numerical errors. All ten sensitivity settings pass after recalibration (Table 3).

E.6 Code and saved-result reproduction

The experiment project contains the simulator, tests, configurations and saved results. The five original suites are under `experiments/full_run/results/`; the additional four-channel search is implemented in `experiments/run_fixed_all_selected.py`

Table 3: Sensitivity results at $L_\varepsilon = 12$ and 24. Each mean uses 20,000 holdout trajectories. Every mean plus two SE is below its target.

Variant	L_ε	T	Cost	Mean MSE/ε	2 SE/ε
$h = 0.10$	12	1440	30,357	0.276	0.029
$h = 0.10$	24	2880	65,985	0.862	0.026
$h = 0.25$	12	576	10,638	0.811	0.021
$h = 0.25$	24	1152	25,380	0.800	0.016
$h = 0.50$	12	288	5,076	0.783	0.019
$h = 0.50$	24	576	12,060	0.795	0.019
$(\beta_1, \beta_2) = (0.8, 0.98)$	12	273	4,812	0.846	0.017
$(\beta_1, \beta_2) = (0.8, 0.98)$	24	546	11,169	0.632	0.050
Asymmetric scales	12	288	4,878	0.732	0.017
Asymmetric scales	24	576	11,622	0.717	0.016

and writes to `experiments/results/fixed_all_channel/`. The script `experiments/generate_revised_figures.py` reads these saved results and regenerates Figure 2, Figure 3, and Table 1 without simulations. The original saved trajectories and ablations generate Figures 4–5 and Tables 2–3. The archived uncalibrated heuristic is excluded from the comparisons.

F Comparison with related numerical models

In this section, we compare the numerical errors modeled, the resources counted, and the guarantees proved in related precision work. Table 4 shows how these choices differ from our cumulative-write question.

Table 4: Related precision guarantees. Each row uses its stated numerical model and resource; the last row gives a matching cumulative-cost law.

Line of work	Numerical resource or model	Type of guarantee
Low-bit optimizer states (Dettmers et al., 2022; Li et al., 2023)	Blockwise state formats; peak memory	Empirical training performance at small fixed widths
Quantized Adam (Tang et al., 2026)	Relative errors in four channels	Sufficient precision for convergence
Dynamic allocation (Yan et al., 2021; Ding et al., 2024)	Gradient communication or mixed integer precision	Error-constrained allocation; restricted tight examples in DQ-SGD
Information constraints (Mayekar & Tyagi, 2020; Acharya et al., 2021)	Information in oracle-gradient outputs	Precision and convergence lower bounds; adaptive processing allowed in Acharya et al. (2021)
This work	Fixed-scale dither; sum of all four write widths	Matching order over all causal policies, without a prescribed state domain

The choice of oracle also distinguishes this question from related optimizer work. Quantized moving averages exhibit swamping and state staleness (Xu et al., 2025; Topollai & Choromanska, 2026), while floating-point stochastic rounding has a different error law from fixed-scale uniform dither (Akyol & Rose, 2013; Croci et al., 2022; Ozkara et al., 2025). Error-feedback Adam adds residual states (Chen et al., 2021; 2023); distributed methods may adapt their encoding (Michelusi et al., 2022). The empirical study *Beyond Dense Adam States* fits log-space quantizer ranges to observed state blocks and treats state topology and state-specific precision jointly (Wang, 2026). These adaptive-range designs change the numerical oracle considered here.