ADAM UNDER BOUNDED VARIANCE: POLYLOGARITHMIC CONFIDENCE WITHOUT CLIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-probability guarantees quantify the reliability of Adam and other adaptive optimizers. Under bounded noise variance, Jin et al. (2026) show that Adam's second moment normalization already yields polylogarithmic confidence for the preconditioned gradient energy, while their ordinary squared gradient guarantee retains polynomial dependence on $1/\delta$. This motivates asking whether Adam can preserve that confidence benefit in ordinary stationarity. We show that adaptive normalization coupled with second moment memory of order $\sqrt{T}$ achieves this without clipping or stronger noise assumptions. For smooth objectives bounded below and conditionally unbiased noise with bounded variance, we combine $1-\beta_2 \asymp T^{-1/2}$ with a suitably chosen constant learning rate, any fixed first moment coefficient, and positive second moment initialization. For any fixed positive offset, $T \geq 1$, and $0 < \delta < 1$, the average squared gradient norm is $\mathcal{O}(\mathrm{polylog}(1/\delta)/\sqrt{T})$ with probability at least $1-\delta$. When the target confidence is specified before the run, adjusting the memory and learning rate sharpens the guarantee to $\mathcal{O}(\min\{1, \sqrt{\log(e/\delta)/T}\})$ under the same assumptions. Descent localized by objective height and a potential coupling accumulation with displacement control the cumulative gradient cost of large denominators and give both guarantees through a common analysis.

1 INTRODUCTION

High-probability convergence guarantees quantify the reliability of a single run of stochastic optimization. For smooth nonconvex objectives, the average squared gradient norm measures progress toward stationarity (Ghadimi & Lan, 2013). A guarantee that holds with probability at least $1-\delta$ must describe both how this measure decreases with the iteration budget and how it grows as the requested failure probability δ becomes smaller. Polylogarithmic confidence dependence makes high reliability possible at a modest additional computational cost.

For Adam (Kingma & Ba, 2015), obtaining this reliability from its own adaptive mechanism would explain how it handles rare, large stochastic gradients. High-probability analyses assume exponential noise moments for Adam (Hong & Lin, 2024b), or use clipping for gradient methods (Nguyen et al., 2023) and Adam-Norm/AdaGrad-Norm (Chezhegov et al., 2025) under bounded variance. We consider bounded conditional noise variance, without assuming moments above order two.

Jin et al. (2026) identify second moment normalization as the source of Adam's confidence advantage over SGD. Under bounded variance, they already obtain $\mathrm{polylog}(1/\delta)$ control of the preconditioned gradient energy: the cumulative squared gradient weighted by Adam's coordinatewise step sizes. Their conversion to the ordinary average squared gradient, however, incurs a polynomial confidence cost. It remains open whether Adam's normalization and memory can preserve the intermediate polylogarithmic confidence benefit in ordinary stationarity.

Motivated by this intermediate control, we ask:

Can Adam's normalization and memory achieve polylogarithmic confidence under bounded variance at the $T^{-1/2}$ rate?

We establish this guarantee through Adam's normalization and memory decay, without clipping or tuning based on confidence. Prescribing the confidence level before the run permits a sharper dependence on the square root of the logarithm.

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.

The central difficulty is to retain the confidence benefit of normalization when passing from preconditioned gradient energy to ordinary stationarity. A large stochastic gradient increases its own denominator and limits the immediate update, but the enlarged denominator also suppresses subsequent progress. Bounding its maximum over the trajectory can incur a polynomial cost in $1/\delta$, even when the preconditioned gradient energy has a polylogarithmic guarantee. Our analysis controls the cumulative gradient cost of these intervals with large denominators. A stopped descent argument localizes the trajectory. Tuning independent of confidence uses a threshold for the objective height that grows with the requested confidence; tuning based on confidence permits a constant threshold. A joint potential for the accumulator and displacement then combines second moment decay with the gradient changes caused by motion during each interval. This yields exponential control of the accumulated interval cost. Its dependence on the memory length $1/(1-\beta_2)$ determines the confidence cost in the final conversion for both tunings.

Contribution. We study Adam in Algorithm 1. For smooth objectives bounded below and conditionally unbiased stochastic gradients with bounded variance, fix any first moment coefficient $\beta_1 \in [0,1)$, positive second moment initialization, and positive offset ϵ. Choosing $1-\beta_2 \asymp T^{-1/2}$ and $\gamma \asymp T^{-1/2}$ gives effective memory of order $\sqrt{T}$ and, for every $T \geq 1$ and $0 < \delta < 1$, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{\operatorname{polylog}(1/\delta)}{\sqrt{T}}\right). \tag{1}$$

With parameters depending on δ, the convergence rate becomes $\mathcal{O}(\min\{1, \sqrt{\log(e/\delta)/T}\})$ under the same assumptions. The implicit constants are independent of T, δ. Both rules are selected before sampling and held constant during the run. Theorems B.1 and B.2 specify their full parameter dependence.

2 Preliminaries

We study the stochastic optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, where $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and only stochastic estimates of its gradient are available. Starting from a deterministic point $\mathbf{x}_0$, an iterative method queries the oracle at $\mathbf{x}_t$ to obtain $g_t \in \mathbb{R}^d$ and uses this observation to form $\mathbf{x}_{t+1}$.

All random variables are defined on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq -1}, \mathbb{P})$. For $t \geq 0$, $\mathscr{F}_{t-1}$ contains the information available immediately before query t, including any oracle state; $\mathscr{F}_{-1}$ is the initial history. The iterate $\mathbf{x}_t$ is $\mathscr{F}_{t-1}$-measurable, and the returned gradient g_t is $\mathscr{F}_t$-measurable. Expectations and conditional expectations are denoted by $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot \mid \mathscr{F}_{t-1}]$, respectively.

2.1 Adaptive moment estimation (Adam)

Adam uses exponential averages of gradients and squared gradients to construct a coordinatewise update (Kingma & Ba, 2015). We use the recursion analyzed by Jin et al. (2026): M_t and V_t are the exponential averages of gradients and squared gradients after query t, respectively. Their incoming values M_{t-1} and V_{t-1} are $\mathscr{F}_{t-1}$-measurable.

Algorithm 1 Adam

Require: $\mathbf{x}_0 \in \mathbb{R}^d$, $M_{-1} = 0$, $V_{-1} = v1$ with $v > 0$; $\beta_1 \in [0,1)$, $\beta_2 \in [0,1)$, $\epsilon > 0$, $\gamma > 0$, horizon $T \in \mathbb{N}_+$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: Query the oracle at $\mathbf{x}_t$ to obtain g_t.
3: Update the first moment: $M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1) g_t$.
4: Update the second moment: $V_t \leftarrow \beta_2 V_{t-1} + (1-\beta_2) g_t^{\odot 2}$.
5: Take a normalized momentum step: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \gamma \dfrac{M_t}{\sqrt{V_t} + \epsilon}$.
6: **end for**

In Algorithm 1, squares, square roots, and division act coordinatewise; $g_t^{\odot 2}$ denotes the vector of squared coordinates. The offset ϵ is outside the square root, and $V_{-1} = v1$ initializes every coordinate with the same positive scalar. The iterate is denoted by $\mathbf{x}_t$, and $V_{t,i}$ denotes coordinate i of V_t; the same coordinate notation applies to the other vectors. We write $\partial_i f(\mathbf{x})$ for the ith coordinate of $\nabla f(\mathbf{x})$. Unless otherwise specified, $\|\cdot\|$ denotes the Euclidean norm throughout this paper.

2.2 Technical assumptions

The objective conditions describe the descent geometry, and the oracle condition controls the noise around the true gradient.

Assumption 1 (Finite lower bound)**.** There exists a finite $f_\star \in \mathbb{R}$ such that

$$f(\mathbf{x}) \geq f_\star \qquad \text{for every } \mathbf{x} \in \mathbb{R}^d.$$

This gives a finite budget $f(\mathbf{x}_0) - f_\star$ against which accumulated descent can be compared, as in standard nonconvex stochastic optimization (Ghadimi & Lan, 2013).

Assumption 2 (Lipschitz gradient)**.** For a fixed $L > 0$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \text{for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

This standard smoothness condition controls the quadratic remainder of an update and relates gradient changes to displacement along the trajectory (Ghadimi & Lan, 2013; Jin et al., 2026).

Assumption 3 (Conditional variance)**.** For a fixed $\sigma \geq 0$, the oracle satisfies, almost surely at every iteration,

$$\mathbb{E}[g_t \mid \mathscr{F}_{t-1}] = \nabla f(\mathbf{x}_t), \qquad \mathbb{E}[\|g_t - \nabla f(\mathbf{x}_t)\|^2 \mid \mathscr{F}_{t-1}] \leq \sigma^2.$$

As in Jin et al. (2026), the moment bound concerns centered noise, allowing the stochastic gradient's second moment to grow with the true gradient. Both conditions hold conditional on the full observed history, including the oracle state. Fixed minibatch averages are covered, and moments above order two need not exist.

Stationarity and asymptotic notation. The performance criterion is the average over a single run $T^{-1} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2$. The problem parameters $d, L, f(\mathbf{x}_0) - f_\star, \sigma$, the first momentum β_1, the positive initialization v, and the positive smoothing constant ϵ are fixed. We use $\mathcal{O}$ for upper bounds up to positive constants, $\asymp$ for equality up to positive constant factors, and $\widetilde{\mathcal{O}}$ to suppress logarithmic factors.

3 Main results

The central choice is the duration of second moment memory. We explain how it preserves the confidence benefit of normalization and quantify the convergence rate.

3.1 Preserving normalization through the final conversion

Jin et al. (2026, Proposition 1) show that, under bounded noise variance, Adam's *preconditioned gradient energy*,

$$\sum_{t=0}^{T-1} \sum_{i=1}^{d} \frac{\gamma\big(\partial_i f(\mathbf{x}_t)\big)^2}{\sqrt{V_{t,i}} + \epsilon},$$

already has a $\mathrm{polylog}(1/\delta)$ bound with probability at least $1 - \delta$ under their prescribed tuning. This quantity weights each squared gradient coordinate by its adaptive step size. Their ordinary squared gradient guarantee is $\mathcal{O}(T^{-1/2}\delta^{-1/2}\,\mathrm{polylog}(1/\delta))$ (Jin et al., 2026, Theorem 3). The additional $\delta^{-1/2}$ dependence enters when removing the adaptive denominator: second moment control bounds its square in expectation, and Markov's inequality gives a factor proportional to the square root of $1/\delta$. The role of their choice $\beta_2 = 1 - 1/T$ is central to this conversion.

Memory determines the duration of a large denominator. With $\beta_2 = 1 - 1/T$, an old contribution retains the fraction $(1 - 1/T)^T \asymp 1$ over the entire horizon. A rare, large observation can therefore suppress subsequent updates for order T steps, even though normalization controls the immediate update. This persistence when memory spans the horizon obstructs polylogarithmic confidence at the $T^{-1/2}$ rate for ordinary stationarity.

Exponential forgetting suggests controlling the duration of that influence. After $t - s$ steps, a contribution to the incoming state $V_{s-1,i}$ has been multiplied by $\beta_2^{t-s} \le e^{-(t-s)(1-\beta_2)}$. Temporarily omit subsequent inputs and consider $V_{s-1,i} > K^2$, where K^2 is a deterministic accumulator threshold. The number of decay steps needed to reach K^2 is

$$\left\lceil \frac{\log(V_{s-1,i}/K^2)}{-\log \beta_2} \right\rceil \le 1 + \frac{\log(V_{s-1,i}/K^2)}{1 - \beta_2}.$$

The influence time therefore grows logarithmically with the initial accumulator size and proportionally to the memory length $1/(1 - \beta_2)$. Shorter memory reduces this logarithmic duration cost.

From forgetting to ordinary stationarity. In the actual trajectory, fresh observations replenish the accumulator and the gradient changes as the iterate moves. We partition each coordinate into intervals above and below K^2. Below the threshold, the denominator is controlled directly. Above it, Step 3 of Section 4 constructs a joint potential for the accumulator and displacement that combines exponential forgetting, smoothness, and a logarithmic denominator reserve to bound the cumulative squared gradient cost exponentially. Step 4 uses the linear dependence of entry charges on the squared entering observation, so conditional second moments suffice, and combines the interval bounds through an exponential supermartingale to obtain logarithmic confidence dependence. Step 5 substitutes the memory and choices of objective height to give the stationarity rates. The formal bounds for intervals and entry charges are Lemmas B.13 and B.14.

3.2 Convergence rate of Adam

Fix $\Delta_0 > 0$ with $f(\mathbf{x}_0) - f_\star \le \Delta_0$, and write $\Lambda_\delta = \log(128d/\delta)$. We use $1 - \beta_2 \asymp T^{-1/2}$ and $\gamma \asymp T^{-1/2}$, with proportionality constants independent of T, δ and exact coefficients in (B.6). The following proposition quantifies the conversion from preconditioned gradient energy to the ordinary squared gradient criterion.

Proposition 1 (Conversion to ordinary gradients, informal)**.** *Under Assumptions 1 to 3, fix $\beta_1 \in [0, 1)$, $v > 0$, and $0 < \epsilon \le 1$. Recall $\Lambda_\delta = \log(128d/\delta)$ defined at the beginning of Section 3.2. Use the square root parameter rule stated there, with the precise coefficients in* (B.6)*. Whenever $T \in \mathbb{N}_+$, $0 < \delta < 1$, and $\Lambda_\delta \le \sqrt{T}$, Adam satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le C\left[\frac{\Lambda_\delta^2}{\gamma T} + \frac{\Lambda_\delta^2}{(1-\beta_2)T} + \frac{\Lambda_\delta \gamma^2}{1-\beta_2}\left(1 + \log(1 + \Lambda_\delta)\right)\right] \tag{2}$$

with probability at least $1 - \delta$. The constant C is independent of T, δ.

Preconditioned gradient energy and the threshold depending on the height supply the first term in (2); the aggregate interval reward supplies the second; the denominator reserve supplies the third. Under (B.6), their respective orders are

$$\frac{\Lambda_\delta^2}{\sqrt{T}}, \qquad \frac{\Lambda_\delta^2}{\sqrt{T}}, \qquad \frac{\Lambda_\delta[1 + \log(1 + \Lambda_\delta)]}{\sqrt{T}}.$$

Since $1 + \log(1 + \Lambda_\delta) \le 2\Lambda_\delta$ for $\Lambda_\delta \ge 1$, the total rate is $\mathcal{O}(\Lambda_\delta^2/\sqrt{T})$. Proposition B.3 specifies the full parameter dependence and the preconditioned gradient energy used in the conversion.

Theorem 1 (Convergence rate with parameters independent of δ, informal)**.** *Under Assumptions 1 to 3, fix $\beta_1 \in [0, 1)$, $v > 0$, and $0 < \epsilon \le 1$. Use the square root parameter rule stated at the beginning of Section 3.2, with the precise coefficients in* (B.6)*. There is $C > 0$, independent of T, δ,*

such that, for every $T \in \mathbb{N}_+$ and $0 < \delta < 1$, Adam satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{C\Lambda_\delta^4}{\sqrt{T}}$$

with probability at least $1-\delta$. Here Λ_δ is the confidence factor defined at the beginning of Section 3.2. If $\Lambda_\delta \leq \sqrt{T}$, the same parameters give the bound $C\Lambda_\delta^2/\sqrt{T}$ with probability at least $1-\delta$.

The same parameter rule independent of confidence applies to all horizons. Theorem B.1 and Corollary B.1 in Section B.2 give the exact tuning, the dependence on smoothness and initial gap, and the extension to every positive offset. They also specify uniformity over objective and oracle classes. The proof is in Section B.8. At fixed confidence and offset the rate is $\mathcal{O}(T^{-1/2})$. For $\delta = T^{-u}$ with fixed $u > 0$, the quadratic logarithm bound applies for all sufficiently large T, giving $\mathcal{O}(T^{-1/2}(\log T)^2)$.

3.3 Convergence rate with parameters depending on δ

We next give a convergence rate when the second momentum and learning rate may depend on the prescribed failure probability δ. Choosing algorithmic parameters for a target confidence is common in high-probability convergence analyses, including stochastic mirror descent (Liu et al., 2023a, Theorem 3.1) and Adam (Li et al., 2023, Theorem 4.1). Here β_2 and γ are selected using T, δ before sampling and held constant during the run. Their dependence on δ keeps the objective height at a constant multiple of the chosen initial gap bound, so the gradient conversion gives square root logarithmic confidence dependence.

Theorem 2 (Convergence rate with parameters depending on δ, informal)**.** *Under Assumptions 1 to 3, fix $\beta_1 \in [0, 1)$, $v > 0$, and $0 < \epsilon \leq 1$. Use the deterministic parameter rule* (B.8)*. There is $U > 0$, independent of T, δ, such that, for every $T \in \mathbb{N}_+$ and $0 < \delta < 1$, Adam satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq U \min\left\{1, \sqrt{\frac{\Lambda_\delta}{T}}\right\}$$

with probability at least $1-\delta$. Here $\Lambda_\delta = \log(128d/\delta)$ is the confidence factor defined at the beginning of Section 3.2.

At fixed problem parameters and offset, the rule has $1-\beta_2 \asymp \sqrt{\Lambda_\delta/T}$ and $\gamma \asymp (T\Lambda_\delta)^{-1/2}$ when $\Lambda_\delta \leq T$, giving memory of order $\sqrt{T/\Lambda_\delta}$. For $\Lambda_\delta > T$, $1-\beta_2 \asymp 1$ and $\gamma \asymp T^{-1}$, yielding a constant pathwise bound. The exact choices are in (B.8), with the same fixed factors as (B.6). Theorem B.2 states the full dependence on smoothness, the positive gap bound, and every positive offset. Its proof combines Lemma B.10 with the same conversion estimate used for Theorem 1.

3.4 Comparison of confidence dependence

Comparison at the same stationarity measure. We compare the guarantees under their stated deterministic tunings. For $T \geq 10$, Theorem 3 of Jin et al. (2026) uses $1-\beta_2 = 1/T$ and γ a sufficiently small fixed multiple of $T^{-1/2}$. With probability at least $1-\delta$, it gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{C\log^6(C/\delta)}{\sqrt{\delta T}}.$$

For fixed positive offsets, the two guarantees in this comparison use tuning independent of confidence and have order $T^{-1/2}$ at fixed confidence. Theorem 1 has polylogarithmic confidence dependence, whereas the guarantee with memory of order T includes $\delta^{-1/2}$. At $\delta = T^{-u}$ for fixed $u > 0$ and sufficiently large T, their respective orders are $CT^{-1/2}(\log T)^2$ and $CT^{(u-1)/2}(\log T)^6$.

4 PROOF SKETCH

Both guarantees follow from stopped descent and a conversion from preconditioned gradient energy to ordinary squared gradients under Assumptions 1 to 3. Fix $T \in \mathbb{N}_+$, $\epsilon > 0$, and $0 < \delta < 1$. Recall the positive initial gap bound $f(\mathbf{x}_0) - f_\star \le \Delta_0$ and the confidence factor $\Lambda_\delta = \log(128d/\delta)$, both defined at the beginning of Section 3.2. We use the prescribed parameter choices in (B.6) and (B.8), whose fixed factors are specified in (B.2), (B.3), and (B.4). At fixed problem parameters and offset, write $\asymp$ for matching $\mathcal{O}$ and Ω bounds. The *rule independent of confidence* has

$$1 - \beta_2 \asymp T^{-1/2}, \qquad \gamma \asymp T^{-1/2}, \tag{3}$$

so its effective memory is of order $\sqrt{T}$. The *rule depending on* δ has

$$1 - \beta_2 \asymp \min\left\{1, \sqrt{\frac{\Lambda_\delta}{T}}\right\}, \qquad \gamma \asymp \frac{1}{\sqrt{T \min\{\Lambda_\delta, T\}}}. \tag{4}$$

Its memory is of order $\sqrt{T/\Lambda_\delta}$ when $\Lambda_\delta \le T$; otherwise it is constant and $\gamma \asymp T^{-1}$. Throughout, C denotes constants independent of T, δ, H.

To account for momentum, use the incoming moment states M_{t-1}, V_{t-1} from Algorithm 1 and define

$$\mathbf{y}_t = \mathbf{x}_t - \frac{\gamma\beta_1}{1-\beta_1}\frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon}.$$

The auxiliary trajectory removes the direct momentum lag. Its update is

$$\mathbf{y}_{t+1} - \mathbf{y}_t = -\gamma\frac{g_t}{\sqrt{V_t} + \epsilon} + \frac{\gamma\beta_1 M_{t-1}}{1-\beta_1}\left(\frac{1}{\sqrt{V_{t-1}} + \epsilon} - \frac{1}{\sqrt{V_t} + \epsilon}\right),$$

where multiplication and division act coordinatewise. The second term is the correction caused by the changing denominator. By (B.16), the offset $\|\mathbf{x}_t - \mathbf{y}_t\|$ is $O(\gamma/\sqrt{1-\beta_2})$: it is $O(T^{-1/4})$ for the rule independent of confidence and $O(T^{-1/4}\Lambda_\delta^{-3/4})$ for the rule depending on δ when $\Lambda_\delta \le T$. The corresponding definition and update decomposition appear in (B.14) and (B.18).

Step 1 (Stopped descent with a predictable signal). As in the proof of Lemma B.3, smoothness gives

$$f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t) \le \langle \nabla f(\mathbf{y}_t), \mathbf{y}_{t+1} - \mathbf{y}_t\rangle + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2.$$

We substitute the auxiliary update above and control its conditional mean and fluctuations. For a deterministic cutoff $H \ge 16\Delta_0$ for the objective gap, stop at $\tau_H = \inf\{t \ge 0 : f(\mathbf{x}_t) - f_\star \ge H\}$. Before this time, (B.13) gives $\|\nabla f(\mathbf{x}_t)\|^2 \le 2LH$ and the conditional second moment is at most $S_H = 2LH + \sigma^2$. For tuning independent of confidence, we take $H = 16\Delta_0\Lambda_\delta$ when $\Lambda_\delta \le \sqrt{T}$; for tuning based on δ, we take $H = 16\Delta_0$ when $\Lambda_\delta \le T$. Step 5 handles the remaining confidence ranges. Thus H and the local second moment bound S_H grow linearly with Λ_δ in the first range and are constant in the second. The negative term in the stopped descent estimate controls the predictable preconditioned gradient energy, with the fixed learning rate factor separated out,

$$W_n = \sum_{t=0}^{n-1} I_t \sum_{i=1}^{d} \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon}, \qquad I_t = 1_{\{t < \tau_H\}}.$$

In these localization ranges, the target budget $W_T \le H/\gamma$ has order $\Lambda_\delta\sqrt{T}$ for (3) and $\sqrt{T\Lambda_\delta}$ for (4), at fixed problem parameters and offset. These are the stopped energy and indicator of an active state in (B.15) and (B.12). Its denominator is measurable before the current query, whereas the actual denominator uses the updated V_t in Algorithm 1. Replacing the current random denominator by this predictable one introduces a normalization bias. The auxiliary update also contributes the denominator correction and the quadratic smoothness remainder. After conditioning on the history before query t, the useful descent term is $-\gamma(W_{t+1} - W_t)/8$; the remaining terms are controlled by the energies of the normalized updates in Step 2. Lemmas B.6 and B.8 give the precise bounds. The conditional mean of the stopped objective increment is

$$\mu_t^f = \mathbb{E}\left[f(\mathbf{y}_{(t+1)\wedge\tau_H}) - f(\mathbf{y}_{t\wedge\tau_H}) \,\middle|\, \mathscr{F}_{t-1}\right].$$

Subtracting this mean, defined also in (B.27), leaves a martingale difference whose cumulative fluctuation is controlled in the proof of Lemma B.9.

Step 2 (Normalized energy and localization). Both the descent errors and the martingale's conditional variance are controlled by the energies of the normalized samples and momentum

$$A_T = \sum_{t=0}^{T-1} I_t \left\| \frac{g_t}{\sqrt{V_t} + \epsilon} \right\|^2, \qquad P_T = \sum_{t=0}^{T-1} I_t \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\|^2.$$

These sums measure the accumulated squared normalized updates. They are the instances for horizon T of (B.15).

Lemma 1 (Normalized energy, informal)**.** *Under either parameter rule above, for the stopped trajectory at height $H \geq 16\Delta_0$, with probability at least $1 - \delta/8$,*

$$A_T + P_T + \sum_{t=0}^{T-1} I_t \mathbb{E}\left[\left\| \frac{g_t}{\sqrt{V_t} + \epsilon} \right\|^2 \middle| \mathscr{F}_{t-1} \right] \leq C \left[T + \frac{\Lambda_\delta + \log(H/\Delta_0)}{1 - \beta_2} \right].$$

The formal statement is Lemma B.5.

Lemma 1 follows by telescoping logarithms of the second moment accumulator and controlling the predictable energy by its bounded realized charges (Lemma B.2). Thus a large sample contributes logarithmically to the cumulative budget. Lemma B.1 then controls the stopped objective fluctuation. Use the height and confidence range specified in Step 1 for each rule. In both ranges the energy budget is $\mathcal{O}(T)$. The prescribed tunings make the cumulative descent errors small enough that, simultaneously for $n \leq T$, $f(\mathbf{y}_{n \wedge \tau_H}) - f_\star + (\gamma/8) W_n \leq H/8$, as in (B.31) and (B.33). The auxiliary offset is at most $\sqrt{\Delta_0 / L}$. If a first exit occurred, smoothness would give $\sqrt{f(\mathbf{x}_{\tau_H}) - f_\star} \leq \sqrt{H/8} + \sqrt{\Delta_0/2} < \sqrt{H}$, contradicting its definition. Hence the trajectory stays below H, and nonnegativity of $f - f_\star$ gives $W_T \leq H/\gamma$. The resulting event satisfies

$$\mathbb{P}\{\tau_H > T, \quad W_T \leq H/\gamma\} \geq 1 - 3\delta/16.$$

Both statements follow from the common budget for the stopped height in Lemma B.9. Proposition B.1 proves the first; Lemma B.10 proves the second by substituting (4) at $H = 16\Delta_0$. This is the point where the two tuning rules enter the argument differently.

Step 3 (Exponential control on intervals with high accumulators). The remaining task is to recover ordinary gradients from W_T. Classify each coordinate by whether its incoming state $V_{t-1,i}$ from Algorithm 1 exceeds the deterministic threshold

$$K^2 \asymp v + (H/\Delta_0) S_H, \qquad K^2 \geq \max\{v, 16 S_H\}.$$

The fixed multiplier is specified in (B.48). At fixed problem parameters, K has order $\sqrt{L\Delta_0}\Lambda_\delta$ for $H = 16\Delta_0\Lambda_\delta$ and $\sqrt{L\Delta_0}$ for $H = 16\Delta_0$. Thus low states have a controlled denominator, while high states require a bound on their total duration and gradient cost. Among states $t < \tau_H \wedge T$, a high interval $[s, e)$ begins after a low state and ends at the next low state or $\tau_H \wedge T$. Within this interval, smoothness relates the gradient to its entry value and the displacement from $\mathbf{y}_s$. The proof couples the capped displacement with the denominator reserve:

$$\phi_t = \min\{L^2 \left\| \mathbf{y}_t - \mathbf{y}_s \right\|^2, LH\}, \qquad \Psi_t = \sum_{j=1}^{d} \log\left(1 + \frac{S_H}{V_{t-1,j} + \epsilon^2}\right).$$

The motion potential ϕ_t tracks displacement on the scale of squared gradients, capped at LH. The cap has order Λ_δ for the first height choice and order one for the second. The reserve Ψ_t increases as accumulators decay, with $0 \leq \Psi_t \leq d \log(1 + S_H/\epsilon^2)$. Its upper bound is $O(\log(1+\Lambda_\delta))$ for the first choice and $O(1)$ for the second, at fixed problem parameters and offset. Lemmas B.11 and B.12 give the corresponding drift estimates. The negative term involving normalized energy in the reserve drift pays for displacement. Combining this drift with accumulator decay gives a conditional exponential bound for the interval reward $\sum_{t=s}^{e-1} [(\partial_i f(\mathbf{x}_s))^2 + \phi_t]$, with exponent coefficient $(1 - \beta_2)/(12LH)$. Its reciprocal has order $\Lambda_\delta \sqrt{T}$ under (3) and $\sqrt{T/\Lambda_\delta}$ under (4), in the respective localization ranges. The bound depends on the entry gradients and reserve Ψ_s, multiplied by the normalized accumulator excess $V_{s-1,i}/(\beta_2 K^2) - 1$. Lemma B.13 gives this conditional bound; it controls the whole interval through its entry state without requiring a uniform bound on the accumulator.

Step 4 (Entry charges and ordinary squared gradients). An interval beginning at s follows a low state, so the accumulator update gives

$$0 \le \frac{V_{s-1,i}}{\beta_2 K^2} - 1 \le \frac{(1-\beta_2) g_{s-1,i}^2}{\beta_2 K^2}.$$

Smoothness bounds each squared entry gradient by twice its value before the draw plus $CL^2\gamma^2/(1-\beta_2)$. These quantities, evaluated before the draw, are measurable before the entry decision. Taking conditional expectation, with the entry indicator included, then bounds the squared sample by its conditional second moment. This produces a predictable charge for entering the interval; see (B.46) and its derivation.

For each coordinate, a draw in a low state and any ensuing high interval form a stopping time block. At exponential scale $(1-\beta_2)/(12LH)$ from Step 3, deducting the predictable entry charge in the exponent makes each block's exponential reward have conditional mean at most one. Products over successive blocks therefore form a nonnegative exponential supermartingale starting at one, without requiring block independence. Pad terminated blocks with zero rewards and charges. For $0 < \rho < 1$, Markov's inequality at block index T, followed by a union bound over coordinates, gives a penalty $\log(d/\rho)$ in the exponent per coordinate, with total failure probability at most ρ. Thus composition preserves logarithmic confidence dependence; the full argument is in the proof of Lemma B.14.

The total charge depending on the gradients is proportional to $HS_H/(\Delta_0 K^2)$ times the stopped squared gradient sum itself. The threshold in Step 3 makes its coefficient less than $1/2$; moving this term to the left closes the bound. Low states are controlled directly by W_T, as in the proof of Lemma B.15.

Lemma 2 (Ordinary gradient conversion, informal)**.** *Under either parameter rule above, for every $H \ge 16\Delta_0$ and auxiliary failure probability $0 < \rho < 1$, with probability at least $1-\rho$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} I_t \left\|\nabla f(\mathbf{x}_t)\right\|^2 \le CH\sqrt{\frac{L}{\Delta_0}}\frac{W_T}{T} + \frac{CLH}{(1-\beta_2)T}\log\frac{d}{\rho} + \frac{CL^2H\gamma^2}{(1-\beta_2)\Delta_0}\left[1+\log\left(\frac{H}{\Delta_0}\right)\right].$$

The formal result and estimate of the entry charges are Lemmas B.14 and B.15.

The three terms in Lemma 2 account for preconditioned gradient energy, the confidence cost of interval rewards, and the reserve. The term for the interval reward is proportional to $H\log(d/\rho)/((1-\beta_2)T)$: both the localization height and the memory length determine its confidence dependence.

Step 5 (The two convergence guarantees). In the range $\Lambda_\delta \le \sqrt{T}$, apply Lemma 2 with $H = 16\Delta_0\Lambda_\delta$ and $\rho = \delta/2$. Intersecting with Step 2 costs at most $3\delta/16 + \delta/2 < \delta$. On this event, $I_t = 1$ for $t < T$ and $W_T \le H/\gamma$. Substituting the rule independent of confidence (3) and bounding the reserve logarithm gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \le CL\Delta_0\frac{\Lambda_\delta^2}{\sqrt{T}}.$$

For $\Lambda_\delta > \sqrt{T}$, the pathwise step bound gives $T^{-1}\sum_{t<T}\left\|\nabla f(\mathbf{x}_t)\right\|^2 = \mathcal{O}(L\Delta_0 T^{3/2})$, which is covered by $\mathcal{O}(L\Delta_0\Lambda_\delta^4/\sqrt{T})$. This proves the guarantee with parameters independent of confidence; see Proposition B.2 and Section B.8.

For the rule depending on δ (4) with $\Lambda_\delta \le T$, use $H = 16\Delta_0$ and the second localization in Step 2. The same conversion lemma and failure budget apply. Since $S_H = 32L\Delta_0 + \sigma^2$ and $W_T \le 16\Delta_0/\gamma$, its three terms have respective orders

$$\sqrt{\frac{\Lambda_\delta}{T}}, \qquad \sqrt{\frac{\Lambda_\delta}{T}}, \qquad \frac{1}{\sqrt{T}\Lambda_\delta^{3/2}}$$

at fixed problem parameters and offset. They give the square root logarithmic rate. For $\Lambda_\delta > T$, use the saturated branch of (4). The same filter bound limits the total displacement to $\sqrt{\Delta_0/L}$, so smoothness gives the deterministic bound $8L\Delta_0$ on the average squared gradient. Section B.9 supplies the explicit coefficients and combines both ranges. Figure 1 summarizes the argument using the lemmas and steps above.

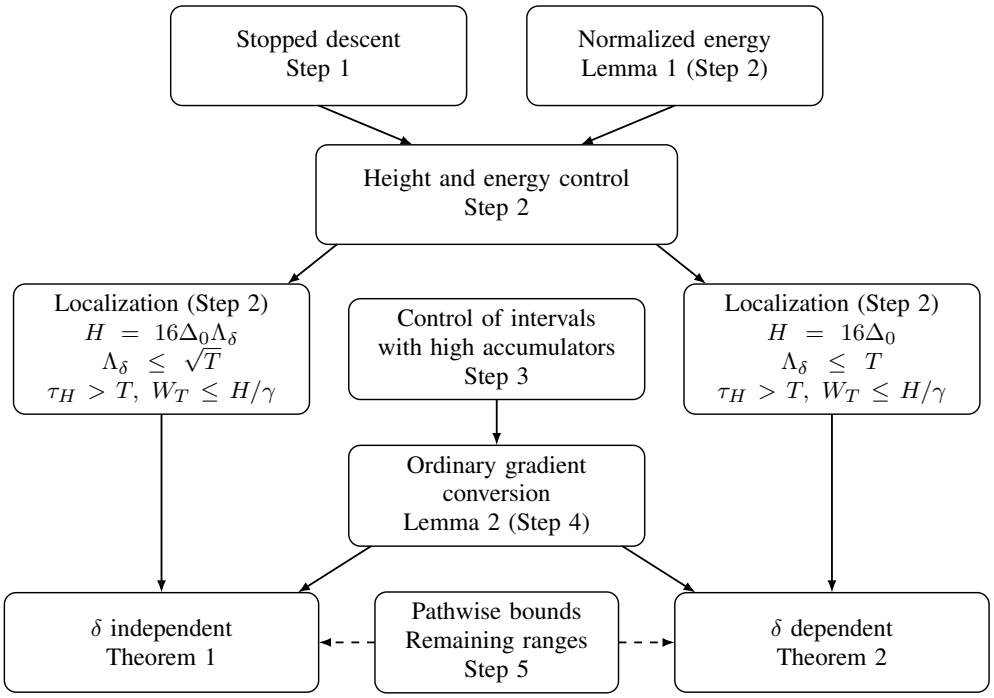


Figure 1: Proof outline for Theorems 1 and 2 using the informal lemmas and steps in Section 4. Stopped descent and normalized energy yield two localization heights, which combine with the same ordinary gradient conversion. Dashed arrows show the pathwise completion in Step 5. Figure 2 in Appendix B.1 gives the full dependencies among the formal results.

5 Related work

Adam and RMSProp analyses identify convergent parameter choices and establish stationarity guarantees (Zou et al., 2019; Défossez et al., 2022; Wang et al., 2023a; Jin et al., 2025). Closest to our setting, Jin et al. (2026) analyze Adam under bounded conditional variance with memory spanning the horizon: preconditioned gradient energy has polylogarithmic confidence dependence, while their ordinary gradient bound retains a polynomial factor. High-probability guarantees under heavy tails also use clipped normalized momentum (Cutkosky & Mehta, 2021), minibatch normalization (Hübler et al., 2025), or clipped adaptive methods (Chezhegov et al., 2025). Here, shorter memory and control of the cumulative gradient cost of intervals with large denominators preserve polylogarithmic confidence in Adam's ordinary stationarity. Appendix A discusses the broader literature.

6 Conclusion

For smooth objectives bounded below and conditionally unbiased noise with bounded variance, Adam preserves its confidence benefit in ordinary squared gradient stationarity without clipping or higher noise moments. Square root memory with parameters independent of confidence gives $\mathcal{O}(\mathrm{polylog}(1/\delta)/\sqrt{T})$; prescribing δ permits shorter memory and the sharper rate $\mathcal{O}(\min\{1, \sqrt{\log(e/\delta)/T}\})$. Both guarantees cover every $T \geq 1$ and $0 < \delta < 1$, with explicit dependence on smoothness, the initial gap, and any positive offset. The tuning determines the localization height in a common stopped descent argument; a potential combining displacement and the reserve, together with conditional entry charges, then controls the cumulative gradient cost of intervals with large denominators. Memory decay thus preserves the confidence benefit of normalization through the final gradient conversion.

REPRODUCIBILITY STATEMENT

The appendix supplies the full proofs, including stopping time bookkeeping, constant selection, and general smoothness and initial gap dependence.

AI USE STATEMENT

Generative AI tools assisted with language polishing, LaTeX typesetting. All mathematical derivations and verfication were perfored by the human authors.

REFERENCES

Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1147–1171. PMLR, 2023.

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

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, 2020.

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

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

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 313–355. PMLR, 2022.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 89–160. PMLR, 2023.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

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

Yusu Hong and Junhong Lin. Revisiting convergence of AdaGrad with relaxed assumptions. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1727–1750. PMLR, 2024a.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10827–10877, 2024b. doi: 10.52202/079017-0346.

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

Ruinan Jin, Xiaoyu Wang, and Baoxiang Wang. Stability and convergence analysis of AdaGrad for non-convex optimization via novel stopping time-based techniques, 2024. arXiv:2409.05023v3.

Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of Adam: Bridging the gap with SGD. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27979–28030. PMLR, 2025.

Ruinan Jin, Yingbin Liang, and Shaofeng Zou. Why Adam can beat SGD: Second-moment normalization yields sharper tails, 2026. arXiv:2603.03099v8.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations*, 2022.

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

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52166–52196, 2023. doi: 10.52202/075280-2272.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12931–12963. PMLR, 2022.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 983–992. PMLR, 2019.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum, 2020. arXiv:2007.14294.

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

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Lê Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21884–21914. PMLR, 2023a.

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

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with Sub-Weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24191–24222, 2023. doi: 10.52202/075280-1052.

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

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

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648. PMLR, 2023.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72438–72474, 2024. doi: 10.52202/079017-2309.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39006–39032, 2023a. doi: 10.52202/075280-1695.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 161–190. PMLR, 2023b.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

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

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

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance. *Transactions on Machine Learning Research*, 2025.

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

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*, 2024.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135, 2019.

CONTENTS

1 Introduction **1**

2 Preliminaries **2**
2.1 Adaptive moment estimation (Adam) 2
2.2 Technical assumptions 3

3 Main results **3**
3.1 Preserving normalization through the final conversion 3
3.2 Convergence rate of Adam 4
3.3 Convergence rate with parameters depending on δ 5
3.4 Comparison of confidence dependence 5

4 Proof sketch **6**

5 Related work **9**

6 Conclusion **9**

Reproducibility statement **10**

AI use statement **10**

References **10**

A Extended related work **14**
A.1 Convergence of Adam and RMSProp 14
A.2 High-probability guarantees and heavy-tailed noise 14
A.3 Adaptive stepsizes and the gradient conversion 14

B Proofs of Lemmas and Theorems **15**
B.1 Proof framework 15
B.2 Formal statements 16
B.3 Local control of the Adam trajectory 17
B.3.1 Concentration and smoothness 17
B.3.2 Actual momentum and normalized energy 18
B.3.3 Conditional normalization and increment bounds 21
B.3.4 Stopped descent and the full probability budget 24
B.3.5 Height control with tuning independent of confidence 27
B.3.6 Height control with tuning dependent on confidence 28
B.4 Joint control of high accumulators and interval motion 30
B.5 Entry charges at stopping times and conversion to squared gradients 35
B.6 A stationarity guarantee with parameters independent of confidence 38
B.7 Quantitative conversion to ordinary gradients 40
B.8 Completion of the stationarity and conversion guarantees 40
B.9 Convergence rate with parameters depending on δ 41

A Extended related work

A.1 Convergence of Adam and RMSProp

Adam convergence studies cover failure modes (Reddi et al., 2018), sufficient conditions on memory and step size for Adam, related adaptive methods, and RMSProp (Chen et al., 2019; Zou et al., 2019; Shi et al., 2021), minibatching (Zaheer et al., 2018), and random reshuffling for Adam (Zhang et al., 2022). Bounds for finite horizons cover Adam and AdaGrad under bounded gradients (Défossez et al., 2022), Adam under bounded variance (Wang et al., 2023a), and generalized smoothness with bounded or sub-Gaussian noise (Li et al., 2023). Under ABC second moment conditions, Jin et al. (2025) establish bounds for finite horizons and asymptotic convergence of the last iterate.

Under coordinatewise generalized smoothness and affine noise variance, Zhang et al. (2025) analyze RMSProp and Adam with the regularizer inside the square root. Zhou et al. (2024) study AMSGrad, corrected RMSProp, and AdaGrad, including high-probability results under bounded gradients and sub-Gaussian noise. ADOPT changes the ordering of normalization and momentum and excludes the current gradient from the denominator to obtain convergence across second moment parameter choices (Taniguchi et al., 2024). Using a vector formulation with online-to-nonconvex conversion and discounted regret, Pang (2026) analyze Adam under heavy-tailed noise for a relaxed stationarity criterion.

A.2 High-probability guarantees and heavy-tailed noise

High-probability analyses track confidence rates determined by noise for stochastic gradient methods (Li & Orabona, 2020; Kavis et al., 2022; Liu et al., 2023a). Nonconvex SGD guarantees cover sub-Weibull noise (Li & Liu, 2022; Madden et al., 2024); Adam guarantees use assumptions on exponential moments (Hong & Lin, 2024b) or bounded conditional variance (Jin et al., 2026).

Clipping provides another route for convex and nonconvex stochastic optimization under heavy tails (Gorbunov et al., 2020; Nguyen et al., 2023). Sadiev et al. (2023) give high-probability guarantees for clipped stochastic methods under bounded centered noise moments of order between one and two. In expectation, Zhang et al. (2020) establish clipping rates and matching lower bounds under heavy-tailed noise. Clipped normalized momentum admits high-probability guarantees under moment bounds on the stochastic gradient (Cutkosky & Mehta, 2021); Liu et al. (2023b) extend the analysis to moment bounds on centered noise.

Cutkosky & Mehta (2020) show that momentum enables normalized SGD to attain rates in expectation for nonconvex objectives without large batches. Normalization has also been studied under centered noise moment bounds (Hübler et al., 2025). Liu & Zhou (2025) obtain convergence rates in expectation for normalized momentum without clipping under generalized smoothness and finite noise moments. Adaptive variants include clipped Adam-Norm and AdaGrad-Norm, together with delayed coordinatewise extensions (Chezhegov et al., 2025).

A.3 Adaptive stepsizes and the gradient conversion

Nonconvex AdaGrad analyses address adaptation to unknown parameters (Li & Orabona, 2019; Ward et al., 2020; Faw et al., 2022). Potential arguments control dependence between gradients and adaptive denominators (Wang et al., 2023b); stopping time arguments handle relaxed smoothness and moment conditions (Faw et al., 2023; Jin et al., 2024). High-probability adaptive guarantees are also available for AdaGrad stepsizes under affine bounded or sub-Gaussian noise (Attia & Koren, 2023). Hong & Lin (2024a) analyze AdaGrad with momentum under noise envelopes depending on the objective gap and gradient norm, including extensions to generalized smoothness.

In our analysis, descent localizes the objective height, and second moment memory limits intervals with large accumulators. With concentration tools related to Tropp (2011), a joint potential for displacement and the reserve combines bounded normalized increments and conditional entry charges to convert preconditioned gradient energy into bounds on the average squared gradient along the trajectory via exponential martingales for block rewards. Under Assumption 3, Adam in Algorithm 1 achieves polylogarithmic confidence with tuning independent of confidence and memory of order $\sqrt{T}$; prescribed confidence permits shorter memory and square root logarithmic dependence.

B Proofs of Lemmas and Theorems

B.1 Proof framework

The dependencies in Figure 2 connect the local trajectory estimates to the two stationarity guarantees. The parameter scale and shared coefficients are fixed in (B.2), (B.3), (B.4), and (B.5). The stopped height budget in Lemma B.9 supplies two localization conclusions. For parameters independent of δ, Proposition B.1 uses $H = 16\Delta_0\Lambda_\delta$ when $\Lambda_\delta \leq \sqrt{T}$. For parameters depending on δ, Lemma B.10 uses $H = 16\Delta_0$ when $\Lambda_\delta \leq T$. Both conclusions bound the preconditioned gradient energy and keep the iterates below the chosen objective height. The interval estimates in Lemmas B.13 and B.14 feed into Lemma B.15, which recovers ordinary squared gradients from that energy. Outside these localization ranges, the trajectory bound (B.16) completes the argument pathwise. The quantitative conversion proposition in Section B.7 follows these estimates and is proved there. The proofs in Sections B.8 and B.9 assemble these estimates with their stated coefficients and failure probabilities.

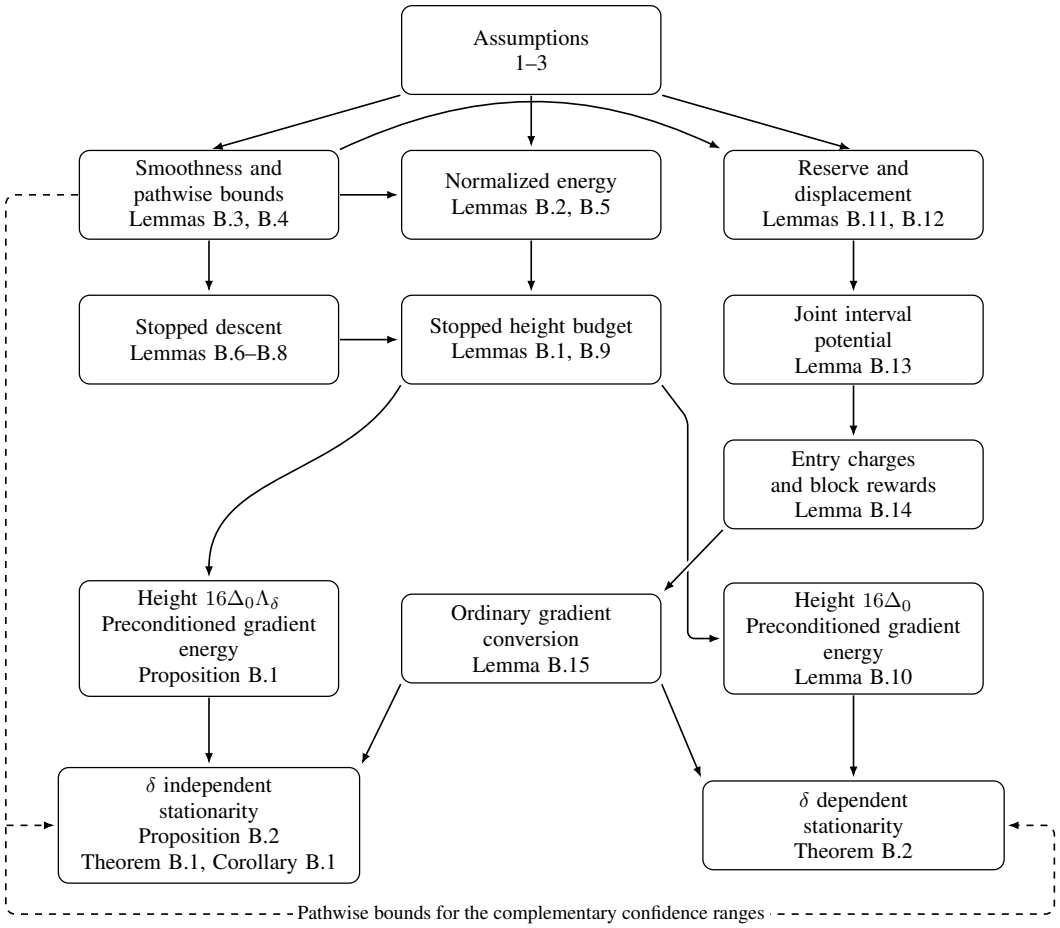


Figure 2: Formal proof dependencies for Theorems B.1 and B.2. The common stopped height budget yields localization under two tuning rules. Both use the same interval analysis and ordinary gradient conversion. Solid arrows show proof dependencies; dashed arrows indicate pathwise bounds for the complementary confidence ranges. The outline in the main text is given in Figure 1.

B.2 FORMAL STATEMENTS

Fix any positive upper bound Δ_0 on the initial objective gap and define the confidence factor by

$$0 < \Delta_0, \qquad f(\mathbf{x}_0) - f_\star \le \Delta_0, \qquad \Lambda_\delta = \log \frac{128d}{\delta}. \tag{B.1}$$

The lower bound $f_\star$ may be any finite lower bound for f. A positive gap bound also covers $f(\mathbf{x}_0) = f_\star$. The coefficients below depend on the displayed problem scales and are independent of ϵ, T, δ.

For $d \ge 1$, $L > 0$, $\sigma \ge 0$, $\beta_1 \in [0, 1)$, and $v > 0$, define the parameter scale

$$K_\star = \frac{d\big(1 + (\sigma^2 + v)/(L\Delta_0) + L\Delta_0/v\big)}{1 - \beta_1} \ge 1. \tag{B.2}$$

The two parameter rules use the same fixed coefficients, with $K_\star$ from (B.2):

$$c_{\beta_2} = \frac{1 - \beta_1^2}{4}, \tag{B.3}$$

$$c_\gamma = \frac{d^5}{2^{24} K_\star^6}, \tag{B.4}$$

$$C = \frac{2^{50} K_\star^9}{d^8}. \tag{B.5}$$

Theorem B.1 (Convergence rate with parameters independent of δ)**.** *Under Assumptions 1 to 3, fix $d \ge 1$, $\beta_1 \in [0, 1)$, $v > 0$, $L > 0$, and $\sigma \ge 0$. Use the initial gap scale Δ_0 and confidence factor Λ_δ from* (B.1)*, and the coefficients c_{β_2}, c_γ, C from* (B.3)*,* (B.4)*, and* (B.5)*. For every $\epsilon > 0$, $T \in \mathbb{N}_+$, and $0 < \delta < 1$, choose*

$$1 - \beta_2 = \frac{c_{\beta_2}}{\sqrt{T}}, \qquad \gamma = \frac{c_\gamma \min\{\epsilon, \sqrt{L\Delta_0}\}}{L\sqrt{T}}. \tag{B.6}$$

Then $\beta_1^2 < \beta_2 < 1$, and Algorithm 1 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le CL\Delta_0 \left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right) \frac{\Lambda_\delta^4}{\sqrt{T}} \tag{B.7}$$

with probability at least $1 - \delta$. If $\Lambda_\delta \le \sqrt{T}$, the power Λ_δ^4 in (B.7) *can be replaced by Λ_δ^2. The coefficients in* (B.6) *are independent of δ and constant during a run; the same rule applies to every horizon. For a class with common positive smoothness and gap bounds and a common variance bound, those bounds may be used in the definitions; the resulting tuning and guarantee are uniform over that class.*

Corollary B.1 (Linear dependence on the inverse offset)**.** *Under Theorem B.1, restrict $0 < \epsilon \le 1$ and use* (B.6)*. For every $T \in \mathbb{N}_+$ and $0 < \delta < 1$, Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le CL\Delta_0 \left(1 + \frac{1}{\sqrt{L\Delta_0}} + \sqrt{L\Delta_0}\right) \frac{\Lambda_\delta^4}{\epsilon\sqrt{T}}$$

with probability at least $1 - \delta$, where Δ_0 is the initial gap scale in (B.1)*. If $\Lambda_\delta \le \sqrt{T}$, the power Λ_δ^4 can be replaced by Λ_δ^2. In particular, for $L = \Delta_0 = 1$, the bound for every horizon is $C\Lambda_\delta^4/(\epsilon\sqrt{T})$, with C independent of ϵ, T, δ.*

Theorem B.2 (Convergence rate with parameters depending on δ)**.** *Under Assumptions 1 to 3, fix $d \ge 1$, $\beta_1 \in [0, 1)$, $v > 0$, $L > 0$, and $\sigma \ge 0$. Use any positive initial gap bound Δ_0 and the confidence factor Λ_δ from* (B.1)*. Let c_{β_2}, c_γ, C be the common coefficients defined in* (B.3)*,* (B.4)*, and* (B.5)*. For every $\epsilon > 0$, $T \in \mathbb{N}_+$, and $0 < \delta < 1$, choose*

$$1 - \beta_2 = c_{\beta_2} \min\left\{1, \sqrt{\frac{\Lambda_\delta}{T}}\right\}, \qquad \gamma = \frac{c_\gamma \min\{\epsilon, \sqrt{L\Delta_0}\}}{L\sqrt{T \min\{\Lambda_\delta, T\}}}. \tag{B.8}$$

Then $\beta_1^2 < \beta_2 < 1$, and Algorithm 1 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le CL\Delta_0 \left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right) \min\left\{1, \sqrt{\frac{\Lambda_\delta}{T}}\right\} \tag{B.9}$$

with probability at least $1 - \delta$. Both parameters are selected using T, δ before sampling and held constant throughout the run. Their coefficients and C are independent of ϵ, T, δ. For common positive smoothness and initial gap bounds and a common variance bound, the tuning and guarantee hold uniformly over the corresponding objective and oracle class. The proof is in Section B.9.

B.3 Local control of the Adam trajectory

Throughout the proof of the upper bound, $f : \mathbb{R}^d \to \mathbb{R}$ has an L-Lipschitz gradient, $L > 0$, and satisfies $f \geq f_\star$. The deterministic initial point obeys $f(\mathbf{x}_0) - f_\star \leq \Delta_0$, where Δ_0 is defined in (B.1). Fix $\beta_1 \in [0, 1)$, $v > 0$, $\sigma \geq 0$, and $\epsilon > 0$. The filtration $(\mathscr{F}_{t-1})_{t\geq 0}$ contains the algorithm and oracle states before each draw. All stopping times are indexed by the query number and taken with respect to this filtration. For such a stopping time S, $\mathscr{F}_{S-1}$ denotes its stopped σ-field. Write $g_t = \nabla f(\mathbf{x}_t) + \xi_t$. The oracle assumptions are

$$\mathbb{E}[\xi_t \mid \mathscr{F}_{t-1}] = 0, \qquad \mathbb{E}[\|\xi_t\|^2 \mid \mathscr{F}_{t-1}] \leq \sigma^2, \qquad g_t \text{ is } \mathscr{F}_t\text{-measurable.} \tag{B.10}$$

The iterates and moment states follow Algorithm 1: V_{t-1} is the incoming second moment state, and V_t incorporates the current sample g_t. Constants denoted by c, C in intermediate estimates may vary between occurrences and depend only on $d, \beta_1, \sigma/\sqrt{L\Delta_0}, v/(L\Delta_0)$, unless stated otherwise. They are uniform in H, ϵ, T, δ, the objective, and the conditional oracle law. The parameter scale $K_\star$ is defined in (B.2); its definition and $v/(L\Delta_0) + L\Delta_0/v \geq 2$ give $K_\star/d \geq 3/(1-\beta_1) \geq 3$. For the intermediate estimates assume

$$0 < 1 - \beta_2 \leq \frac{1-\beta_1^2}{2}. \tag{B.11}$$

In particular, $\beta_2 \geq (1+\beta_1^2)/2 \geq 1/2$.

B.3.1 Concentration and smoothness

We use the following martingale inequality for bounded increments; see also Tropp (2011).

Lemma B.1 (Maximal martingale bound)**.** *Let Z_t be $\mathscr{F}_t$-measurable with $\mathbb{E}[Z_t \mid \mathscr{F}_{t-1}] = 0$ and $|Z_t| \leq D_0$ almost surely, where $D_0 \geq 0$ is deterministic. Suppose a deterministic $B_0 \geq 0$ satisfies*

$$\sum_{t=0}^{T-1} \mathbb{E}[Z_t^2 \mid \mathscr{F}_{t-1}] \leq B_0 \quad \textit{almost surely}.$$

For $0 < \delta < 1$,

$$\mathbb{P}\left\{ \max_{0\leq n\leq T} \left| \sum_{t=0}^{n-1} Z_t \right| > \sqrt{2B_0 \log(2/\delta)} + \frac{2D_0}{3}\log(2/\delta) \right\} \leq \delta.$$

Proof. If $D_0 = 0$ or $B_0 = 0$, all increments vanish almost surely. Otherwise, for $0 \leq \lambda < 3/D_0$, expansion of the exponential, conditional centering, and $k! \geq 2 \cdot 3^{k-2}$ for $k \geq 2$ give

$$\begin{aligned}\mathbb{E}[e^{\lambda Z_t} \mid \mathscr{F}_{t-1}] &\leq 1 + \frac{\lambda^2 \mathbb{E}[Z_t^2 \mid \mathscr{F}_{t-1}]}{2} \sum_{k=2}^{\infty} \left(\frac{\lambda D_0}{3}\right)^{k-2} \\ &\leq \exp\left(\frac{\lambda^2 \mathbb{E}[Z_t^2 \mid \mathscr{F}_{t-1}]}{2(1-\lambda D_0/3)}\right).\end{aligned}$$

The exponential of the cumulative increments minus this predictable compensator is a nonnegative supermartingale starting at one. Stop it at the first partial sum at least $u > 0$, or at T. The deterministic variance budget and $\lambda = u/(B_0 + D_0 u/3)$ yield

$$\mathbb{P}\left\{ \max_{0\leq n\leq T} \sum_{t=0}^{n-1} Z_t \geq u \right\} \leq \exp\left(-\frac{u^2}{2(B_0 + D_0 u/3)}\right).$$

For $u = \sqrt{2B_0\log(2/\delta)} + (2D_0/3)\log(2/\delta)$, the right side is at most $\delta/2$. Apply the same argument to $-Z_t$ and take a union bound. □

Lemma B.2 (Predictable energy from bounded charges)**.** *Suppose X_t is $\mathscr{F}_t$-measurable and $0 \leq X_t \leq D_0$ for a deterministic $D_0 > 0$. Put $r_t = \mathbb{E}[X_t \mid \mathscr{F}_{t-1}]$. For $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sum_{t=0}^{T-1} r_t \leq \frac{1}{1-e^{-1}} \left(\sum_{t=0}^{T-1} X_t + D_0 \log\frac{1}{\delta} \right).$$

Proof. The chord bound $e^{-x} \le 1 - (1-e^{-1})x$ for $x \in [0,1]$ gives

$$\mathbb{E}[e^{-X_t/D_0} \mid \mathscr{F}_{t-1}] \le 1 - \frac{(1-e^{-1})r_t}{D_0} \le \exp\left(-\frac{(1-e^{-1})r_t}{D_0}\right).$$

Thus $\exp\big(D_0^{-1}\sum_{t=0}^{n-1}((1-e^{-1})r_t - X_t)\big)$ is a nonnegative supermartingale starting at one. Markov's inequality at time T proves the assertion. □

Lemma B.3 (Self-bounding objective)**.** *For every* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|\nabla f(\mathbf{x})\|^2 \le 2L(f(\mathbf{x}) - f_\star), \qquad |\sqrt{f(\mathbf{x}) - f_\star} - \sqrt{f(\mathbf{y}) - f_\star}| \le \sqrt{\frac{L}{2}}\,\|\mathbf{x} - \mathbf{y}\|.$$

Proof. Smoothness gives

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

Set $\mathbf{y} = \mathbf{x} - \nabla f(\mathbf{x})/L$ and use $f(\mathbf{y}) \ge f_\star$. For $r > 0$, the first inequality implies $\big\|\nabla\sqrt{f - f_\star + r}\big\| \le \sqrt{L/2}$. Integrating this gradient on a segment and letting $r \downarrow 0$ proves the second inequality. □

Fix a deterministic height $H \ge 16\Delta_0$ for the objective gap, with Δ_0 from (B.1). Define the stopping time, its predictable indicator, and the conditional second moment scale by

$$\begin{gathered}\tau_H = \inf\{t \ge 0 : f(\mathbf{x}_t) - f_\star \ge H\}, \qquad I_t = 1_{\{t < \tau_H\}}, \\ S_H = 2LH + \sigma^2 \le CLH.\end{gathered} \tag{B.12}$$

Here $\inf \varnothing = \infty$. On the active event $\{t < \tau_H\}$, Lemma B.3 and (B.10) imply

$$\begin{aligned}\|\nabla f(\mathbf{x}_t)\|^2 &\le 2LH, \\ \mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] &= \|\nabla f(\mathbf{x}_t)\|^2 + \mathbb{E}[\|\xi_t\|^2 \mid \mathscr{F}_{t-1}] \le S_H.\end{aligned} \tag{B.13}$$

Using the incoming moment states of Algorithm 1, define the auxiliary trajectory

$$\mathbf{y}_t = \mathbf{x}_t - \frac{\gamma\beta_1}{1-\beta_1}\frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon}. \tag{B.14}$$

For integers $n \ge 0$, set

$$\begin{aligned}W_n &= \sum_{t=0}^{n-1} I_t \sum_{i=1}^{d} \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon}, \\ P_n &= \sum_{t=0}^{n-1} I_t \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\|^2, \qquad A_n = \sum_{t=0}^{n-1} I_t \left\| \frac{g_t}{\sqrt{V_t} + \epsilon} \right\|^2.\end{aligned} \tag{B.15}$$

The quantities W_n, P_n, and A_n are, respectively, the preconditioned gradient energy, normalized momentum energy, and normalized sample energy over the active prefix $0 \le t < n \wedge \tau_H$. The denominator in W_n is predictable and is used only for analysis.

B.3.2 Actual momentum and normalized energy

Lemma B.4 (Reachable filters and auxiliary update)**.** *Under* (B.11)*, the moment states of Algorithm 1 satisfy, along every realized trajectory,*

$$\begin{gathered}\frac{M_{t-1,i}^2}{V_{t-1,i}} \le \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)} \le \frac{C}{1-\beta_2}, \\ \|\mathbf{x}_t - \mathbf{y}_t\| \le C\gamma \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\|, \\ \|\mathbf{x}_t - \mathbf{y}_t\| + \|\mathbf{x}_{t+1} - \mathbf{x}_t\| \le \frac{C\gamma}{\sqrt{1-\beta_2}}.\end{gathered} \tag{B.16}$$

For every integer $n \geq 0$,

$$\sum_{t=0}^{n-1} \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\|^2 \leq C \sum_{t=0}^{n-1} \left\| \frac{g_t}{\sqrt{V_t} + \epsilon} \right\|^2. \tag{B.17}$$

The auxiliary update is exactly

$$\mathbf{y}_{t+1} - \mathbf{y}_t = -\gamma \frac{g_t}{\sqrt{V_t} + \epsilon} + F_t + J_t, \tag{B.18}$$

where F_t is predictable and

$$F_{t,i} = \frac{\gamma \beta_1 M_{t-1,i}}{1-\beta_1} \left(\frac{1}{\left(\sqrt{V_{t-1,i}} + \epsilon\right)} - \frac{1}{\left(\sqrt{\beta_2 V_{t-1,i}} + \epsilon\right)} \right),$$
$$J_{t,i} = \frac{\gamma \beta_1 M_{t-1,i}}{1-\beta_1} \left(\frac{1}{\left(\sqrt{\beta_2 V_{t-1,i}} + \epsilon\right)} - \frac{1}{\left(\sqrt{V_{t,i}} + \epsilon\right)} \right).$$

They obey the pathwise bounds

$$\begin{aligned} |F_{t,i}| &\leq C\gamma(1-\beta_2) \frac{|M_{t-1,i}|}{\sqrt{V_{t-1,i}} + \epsilon}, \\ |J_{t,i}| &\leq C\gamma\sqrt{1-\beta_2}\, \frac{g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2}, \\ J_{t,i}^2 &\leq C\gamma^2 \frac{g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2}, \\ \frac{(1-\beta_2) g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2} &\leq 1, \qquad \|\mathbf{y}_{t+1} - \mathbf{y}_t\| \leq \frac{C\gamma}{\sqrt{1-\beta_2}}. \end{aligned} \tag{B.19}$$

Proof. Unrolling the moment updates in Algorithm 1 gives, for $t \geq 1$,

$$M_{t-1,i} = (1-\beta_1) \sum_{j=0}^{t-1} \beta_1^j g_{t-1-j,i}, \qquad V_{t-1,i} = \beta_2^t v + (1-\beta_2) \sum_{j=0}^{t-1} \beta_2^j g_{t-1-j,i}^2.$$

Cauchy–Schwarz therefore implies

$$M_{t-1,i}^2 \leq (1-\beta_1)^2 \left(\sum_{j=0}^{t-1} \frac{\beta_1^{2j}}{(1-\beta_2)\beta_2^j} \right) \left((1-\beta_2) \sum_{j=0}^{t-1} \beta_2^j g_{t-1-j,i}^2 \right).$$

Since $\beta_2 \geq (1+\beta_1^2)/2 > \beta_1^2$, the geometric sum proves the first bound in (B.16), which also holds at $t = 0$. The remaining bounds follow from the parameter update in Algorithm 1 and the definition of the auxiliary trajectory (B.14).

Because $\sqrt{V_{t,i}} + \epsilon \geq \sqrt{\beta_2}\,(\sqrt{V_{t-1,i}} + \epsilon)$,

$$\frac{M_{t,i}}{\sqrt{V_{t,i}} + \epsilon} = \frac{\beta_1(\sqrt{V_{t-1,i}} + \epsilon)}{\sqrt{V_{t,i}} + \epsilon} \frac{M_{t-1,i}}{\sqrt{V_{t-1,i}} + \epsilon} + (1-\beta_1) \frac{g_{t,i}}{\sqrt{V_{t,i}} + \epsilon},$$
$$0 \leq \frac{\beta_1(\sqrt{V_{t-1,i}} + \epsilon)}{\sqrt{V_{t,i}} + \epsilon} \leq \rho_{\beta_1} := \beta_1 \sqrt{\frac{2}{1+\beta_1^2}} < 1.$$

For $\beta_1 > 0$, iteration from $\frac{M_{-1}}{\sqrt{V_{-1}}+\epsilon} = 0$ and Cauchy–Schwarz give

$$\left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\|^2 \leq \frac{(1-\beta_1)^2}{1-\rho_{\beta_1}} \sum_{s=0}^{t-1} \rho_{\beta_1}^{t-1-s} \left\| \frac{g_s}{\sqrt{V_s} + \epsilon} \right\|^2.$$

Summing proves (B.17). If $\beta_1 = 0$, the same bound follows directly from the first moment update in Algorithm 1. Substitution in the definition of the auxiliary trajectory (B.14) gives

$$\mathbf{y}_{t+1} - \mathbf{y}_t = -\gamma \frac{g_t}{\sqrt{V_t} + \epsilon} + \frac{\gamma \beta_1 M_{t-1}}{1 - \beta_1} \left(\frac{1}{\left(\sqrt{V_{t-1}} + \epsilon\right)} - \frac{1}{\left(\sqrt{V_t} + \epsilon\right)} \right).$$

Splitting the reciprocal difference at $\sqrt{\beta_2 V_{t-1}} + \epsilon$ proves (B.18). For the predictable part,

$$|F_{t,i}| \le \frac{\gamma \beta_1}{1 - \beta_1} \left| \frac{M_{t-1,i}}{\sqrt{V_{t-1,i}} + \epsilon} \right| \frac{(1 - \sqrt{\beta_2})\sqrt{V_{t-1,i}}}{\left(\sqrt{\beta_2 V_{t-1,i}} + \epsilon\right)} \le C\gamma(1 - \beta_2) \left| \frac{M_{t-1,i}}{\sqrt{V_{t-1,i}} + \epsilon} \right|.$$

To bound $J_{t,i}$, put $s = \sqrt{\beta_2 V_{t-1,i} + (1 - \beta_2) g_{t,i}^2}$. Both s and $\sqrt{\beta_2 V_{t-1,i}}$ are positive because $v > 0$ and $\beta_2 > 0$. Rationalization yields

$$J_{t,i} = \frac{\gamma \beta_1 M_{t-1,i}}{1 - \beta_1} \frac{(1 - \beta_2) g_{t,i}^2}{(\sqrt{\beta_2 V_{t-1,i}} + \epsilon)(s + \epsilon)(s + \sqrt{\beta_2 V_{t-1,i}})}.$$

By (B.16), $\sqrt{1 - \beta_2}\, |\beta_1 M_{t-1,i}/(1 - \beta_1)| \le C\sqrt{\beta_2 V_{t-1,i}}$. Also

$$\frac{\sqrt{\beta_2 V_{t-1,i}}(s + \epsilon)}{(s + \sqrt{\beta_2 V_{t-1,i}})(\sqrt{\beta_2 V_{t-1,i}} + \epsilon)} \le 1,$$

since the denominator minus the numerator is $\beta_2 V_{t-1,i} + \epsilon s \ge 0$. Consequently,

$$|J_{t,i}| \le C\gamma \sqrt{1 - \beta_2}\, \frac{g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2}.$$

The bound

$$\frac{(1 - \beta_2) g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2} \le 1$$

then gives

$$J_{t,i}^2 \le C\gamma^2 \frac{g_{t,i}^2}{(\sqrt{V_{t,i}} + \epsilon)^2}.$$

For the remaining pathwise bounds, the first estimate in (B.16) and $1 - \beta_1^2/\beta_2 \ge (1 - \beta_1^2)/2$ give

$$\frac{M_{t-1,i}^2}{V_{t-1,i}} \le \frac{2(1 - \beta_1)}{(1 - \beta_2)(1 + \beta_1)}, \qquad \|\mathbf{x}_{t+1} - \mathbf{x}_t\| \le \frac{\sqrt{2d}\,\gamma}{\sqrt{1 - \beta_2}}.$$

Using (B.14) and $K_\star \ge d/(1 - \beta_1)$ from (B.2), we also obtain

$$\|\mathbf{x}_t - \mathbf{y}_t\|^2 \le \frac{2d\beta_1^2\gamma^2}{(1 - \beta_1^2)(1 - \beta_2)} \le \frac{2K_\star\gamma^2}{1 - \beta_2}.$$

The identity $\mathbf{y}_{t+1} - \mathbf{y}_t = (\mathbf{x}_{t+1} - \mathbf{x}_t) + (\mathbf{x}_t - \mathbf{y}_t) - (\mathbf{x}_{t+1} - \mathbf{y}_{t+1})$ therefore gives $\|\mathbf{y}_{t+1} - \mathbf{y}_t\| \le 5\sqrt{K_\star}\gamma/\sqrt{1 - \beta_2}$, proving the last inequality in (B.19). □

Lemma B.5 (Normalized energy and its compensator)**.** *For $T \in \mathbb{N}_+$ and $0 < \delta < 1$, put*

$$\ell_\delta = 1 + \log \frac{16}{\delta}, \qquad B_{H,\delta} = T + \frac{\ell_\delta + \log(H/\Delta_0)}{1 - \beta_2}. \tag{B.20}$$

For the energies A_T, P_T in (B.15)*, there is an event of probability at least $1 - \delta/8$ on which*

$$\begin{aligned} A_T + P_T + \sum_{t=0}^{T-1} I_t \mathbb{E}\left[\left\| \frac{g_t}{\sqrt{V_t} + \epsilon} \right\|^2 \middle| \mathscr{F}_{t-1} \right] &\le C B_{H,\delta}, \\ \sum_{t=0}^{T-1} I_t \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}} + \epsilon} \right\| &\le C\sqrt{T B_{H,\delta}}. \end{aligned} \tag{B.21}$$

The same upper bounds control every prefix of these nonnegative sums.

Proof. For the second moment state V_t in Algorithm 1, the inequality $1 - r \le -\log r$ for $r > 0$ gives, at every active time,

$$\frac{(1-\beta_2)g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2} \le \frac{(1-\beta_2)g_{t,i}^2}{V_{t,i}} = 1 - \frac{\beta_2 V_{t-1,i}}{V_{t,i}} \le \log \frac{V_{t,i}}{\beta_2 V_{t-1,i}}.$$

Sum through the active prefix. Unrolling the second moment update in Algorithm 1 and multiplying its final value by $\beta_2^{T-(T\wedge\tau_H)}$ yields

$$(1-\beta_2)A_T \le -dT\log\beta_2 + \sum_{i=1}^{d} \log \frac{\beta_2^T v + (1-\beta_2)\sum_{s=0}^{T-1}\beta_2^{T-1-s} I_s g_{s,i}^2}{v}.$$

By (B.13), with the moment scale S_H from (B.12), $\mathbb{E}[I_t g_{t,i}^2 \mid \mathscr{F}_{t-1}] \le S_H$, hence $\mathbb{E}[\beta_2^T v + (1-\beta_2)\sum_{s=0}^{T-1}\beta_2^{T-1-s} I_s g_{s,i}^2] \le \max\{v, S_H\}$. Markov's inequality and concavity of the logarithm imply, except with probability $\delta/16$,

$$\sum_{i=1}^{d} \log \frac{\beta_2^T v + (1-\beta_2)\sum_{s=0}^{T-1}\beta_2^{T-1-s} I_s g_{s,i}^2}{v} \le d\log\frac{16\max\{v, S_H\}}{\delta v}.$$

For the height $H \ge 16\Delta_0$ in (B.12), $S_H \le (2 + \sigma^2/(16L\Delta_0))LH$, so $\log(\max\{v, S_H\}/v) \le C + \log(H/\Delta_0)$. Together with $-\log\beta_2 \le 2(1-\beta_2)$, this proves $A_T \le CB_{H,\delta}$ with a constant independent of H. On every path, the active indices form an initial segment; (B.17) applied to this segment gives $P_T \le CA_T$.

Apply Lemma B.2 with failure probability $\delta/16$ to $X_t = I_t \left\| \frac{g_t}{\sqrt{V_t}+\epsilon} \right\|^2 \le d/(1-\beta_2)$ and $r_t = I_t \mathbb{E}\left[\left\| \frac{g_t}{\sqrt{V_t}+\epsilon} \right\|^2 \middle| \mathscr{F}_{t-1} \right]$. Except with another probability $\delta/16$,

$$\sum_{t=0}^{T-1} I_t \mathbb{E}\left[\left\| \frac{g_t}{\sqrt{V_t}+\epsilon} \right\|^2 \middle| \mathscr{F}_{t-1} \right] \le \frac{1}{1-e^{-1}} \left(A_T + \frac{d}{1-\beta_2}\log\frac{16}{\delta} \right).$$

Intersect the two events. Cauchy–Schwarz gives $\sum_{t=0}^{T-1} I_t \left\| \frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon} \right\| \le \sqrt{TP_T}$, proving (B.21). All summands are nonnegative, so their terminal bounds also control all prefixes. Quantitatively, $S_H/(LH) \le 2 + \sigma^2/(16L\Delta_0)$ and $1 - \rho_{\beta_1} \ge (1-\beta_1)/4$ give a coefficient in the filter energy bound at most 16. The coefficient bounding A_T by $B_{H,\delta}$ is at most

$$d\left[2 + \log\max\left\{1, \frac{2L\Delta_0 + \sigma^2/16}{v}\right\}\right] \le 4\frac{K_\star^2}{d}.$$

Indeed, (B.2) gives $L\Delta_0/v \le K_\star/d$, $\sigma^2/(L\Delta_0) \le K_\star/d$, and $K_\star/d \ge 3$; use $\log\max\{1,x\} \le x$ for $x > 0$. The filter bound $P_T \le 16A_T$ and the charge bound with $(1-e^{-1})^{-1} < 2$ bound the first sum in (B.21) by $19A_T + 2dB_{H,\delta} \le 2^7(K_\star^2/d)B_{H,\delta}$. Thus one may take $C = 2^7K_\star^2/d$ there. Cauchy–Schwarz also gives $\sum_{t<T} I_t \left\| M_{t-1}/(\sqrt{V_{t-1}}+\epsilon) \right\| \le \sqrt{2^7K_\star^2/d}\sqrt{TB_{H,\delta}}$. □

B.3.3 Conditional normalization and increment bounds

Lemma B.6 (Signal and denominator energy)**.** *At every active state and coordinate, with S_H as defined in* (B.12) *and V_t the updated second moment in Algorithm 1,*

$$\begin{aligned} &\left| \mathbb{E}\left[\frac{g_{t,i}}{\sqrt{V_{t,i}}+\epsilon} \middle| \mathscr{F}_{t-1} \right] - \frac{\partial_i f(\mathbf{x}_t)}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \right| \\ &\quad \le \frac{\sqrt{1-\beta_2}\,\sigma}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2} \middle| \mathscr{F}_{t-1} \right] \right)^{1/2}. \end{aligned} \tag{B.22}$$

Consequently,

$$\begin{aligned}
&\partial_i f(\mathbf{x}_t)\mathbb{E}\left[\frac{g_{t,i}}{\sqrt{V_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad\geq \frac{(\partial_i f(\mathbf{x}_t))^2}{2\left(\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon\right)} \\
&\qquad -\frac{(1-\beta_2)\sigma^2}{2\left(\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon\right)}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right], \qquad \text{(B.23)} \\
&\left|\mathbb{E}\left[\frac{g_{t,i}}{\sqrt{V_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right|^2 \\
&\quad\leq \frac{2(\partial_i f(\mathbf{x}_t))^2}{(\sqrt{\beta_2 V_{t-1,i}}+\epsilon)^2}+\frac{2(1-\beta_2)\sigma^2}{(\sqrt{\beta_2 V_{t-1,i}}+\epsilon)^2}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned}$$

The signal denominator and normalized energy also satisfy

$$\begin{aligned}
&\left(\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon\right)\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad\leq \frac{2S_H}{\sqrt{\beta_2 V_{t-1,i}}+\epsilon}\leq\frac{2S_H}{\epsilon}.
\end{aligned} \qquad \text{(B.24)}$$

Proof. Fix an active state before the draw and one coordinate. The denominator is $\sqrt{1-\beta_2}$-Lipschitz in its scalar gradient argument. Conditional unbiasedness gives

$$\begin{aligned}
&\mathbb{E}\left[\frac{g_{t,i}}{\sqrt{V_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]-\frac{\partial_i f(\mathbf{x}_t)}{\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon} \\
&\quad=\mathbb{E}\left[g_{t,i}\left(\frac{1}{\sqrt{V_{t,i}}+\epsilon}-\frac{1}{\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon}\right)\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned}$$

The Lipschitz bound and conditional Cauchy–Schwarz bound the absolute value of this expression by

$$\begin{aligned}
&\frac{\sqrt{1-\beta_2}}{\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon}\mathbb{E}\left[\frac{|g_{t,i}|\,|g_{t,i}-\partial_i f(\mathbf{x}_t)|}{\sqrt{V_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad\leq\frac{\sqrt{1-\beta_2}\,\sigma}{\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon}\left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^{1/2}.
\end{aligned}$$

This proves (B.22). Young's inequality and

$$\sqrt{\beta_2 V_{t-1,i}+(1-\beta_2)(\partial_i f(\mathbf{x}_t))^2}+\epsilon\geq\sqrt{\beta_2 V_{t-1,i}}+\epsilon$$

give (B.23). By (B.13) and (B.19),

$$\begin{aligned}
\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]&\leq\frac{S_H}{(\sqrt{\beta_2 V_{t-1,i}}+\epsilon)^2}, \\
(1-\beta_2)\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]&\leq 1.
\end{aligned}$$

Therefore

$$\begin{aligned}
\sqrt{1-\beta_2}\,\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]&\leq\left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^{1/2} \\
&\leq\frac{\sqrt{S_H}}{\sqrt{\beta_2 V_{t-1,i}}+\epsilon}.
\end{aligned}$$

Since

$$\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon \le \sqrt{\beta_2 V_{t-1,i}} + \epsilon + \sqrt{1-\beta_2}\,|\partial_i f(\mathbf{x}_t)|$$

and $|\partial_i f(\mathbf{x}_t)| \le \sqrt{S_H}$, it follows that

$$\begin{aligned}
&\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right) \mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\le \left(\sqrt{\beta_2 V_{t-1,i}} + \epsilon + \sqrt{1-\beta_2}\,|\partial_i f(\mathbf{x}_t)|\right) \mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\le \frac{S_H}{\sqrt{\beta_2 V_{t-1,i}}+\epsilon} + \frac{|\partial_i f(\mathbf{x}_t)|\sqrt{S_H}}{\sqrt{\beta_2 V_{t-1,i}}+\epsilon} \le \frac{2S_H}{\sqrt{\beta_2 V_{t-1,i}}+\epsilon}.
\end{aligned}$$

The inequality $\sqrt{\beta_2 V_{t-1,i}} + \epsilon \ge \epsilon$ completes (B.24). □

Lemma B.7 (Auxiliary increment moments)**.** *For the second moment sequence V_t of Algorithm 1, at every active state,*

$$\begin{aligned}
|\mathbb{E}[J_{t,i} \mid \mathscr{F}_{t-1}]| &\le C\gamma\sqrt{1-\beta_2}\,\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right], \\
\mathbb{E}[J_{t,i}^2 \mid \mathscr{F}_{t-1}] &\le C\gamma^2\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right], \\
|\mathbb{E}[J_{t,i} \mid \mathscr{F}_{t-1}]|^2 &\le \frac{C\gamma^2(1-\beta_2)S_H}{(\sqrt{\beta_2 V_{t-1,i}}+\epsilon)^2}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned} \tag{B.25}$$

Writing $\Delta\mathbf{y}_t = \mathbf{y}_{t+1} - \mathbf{y}_t$, one has

$$\begin{aligned}
\|\mathbb{E}[\Delta\mathbf{y}_t \mid \mathscr{F}_{t-1}]\|^2 &\le \frac{C\gamma^2}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\,\|\nabla f(\mathbf{x}_t)\|^2 \\
&\quad + \frac{C\gamma^2(1-\beta_2)S_H}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad + C\gamma^2(1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2, \\
\mathbb{E}[\|\Delta\mathbf{y}_t\|^2 \mid \mathscr{F}_{t-1}] &\le C\gamma^2\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad + C\gamma^2(1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2.
\end{aligned} \tag{B.26}$$

Proof. The first two bounds of (B.25) follow directly from (B.19). For the third, square the first and use $\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \le S_H/\left(\sqrt{\beta_2 V_{t-1,i}}+\epsilon\right)^2$. By (B.18), (B.19), and the squared triangle inequality,

$$\|\Delta\mathbf{y}_t\|^2 \le C\gamma^2\left(\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 + (1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2\right).$$

Taking conditional expectations proves the second estimate of (B.26). The first follows from the same decomposition, predictability of F_t, the second inequality in (B.23), and the third in (B.25), using $\sqrt{\beta_2 V_{t-1,i}} + \epsilon \ge \epsilon \ge \min\{\epsilon, \sqrt{L\Delta_0}\}$ and $\sigma^2 \le S_H$. The coefficient of $\|\nabla f(\mathbf{x}_t)\|^2$ and the conditional second moment constant are independent of H. For the scale $K_\star$ in (B.2), the three coefficients in the bound on the conditional mean (B.26) may be taken as 6, $18K_\star/d$, and $6(K_\star/d)^2$,

respectively; the two second moment coefficients may be taken as $20K_\star/d$ and $4(K_\star/d)^2$. Indeed, the coefficient of $\gamma(1-\beta_2)$ in the F_t bound is at most $\sqrt{2}\beta_1/(1-\beta_1)$, and that of $\gamma\sqrt{1-\beta_2}$ in the J_t bound is at most $2\beta_1/\sqrt{1-\beta_1^2}$. The squared triangle inequalities with factors 3 and 2 give the stated coefficients. □

B.3.4 Stopped descent and the full probability budget

Lemma B.8 (Stopped objective drift)**.** *Assume $\gamma \le \sqrt{(1-\beta_2)\Delta_0/L}$ and use the height H from* (B.12)*. Define*

$$\mu_t^f = \mathbb{E}\left[f(\mathbf{y}_{(t+1)\wedge\tau_H}) - f(\mathbf{y}_{t\wedge\tau_H}) \,\middle|\, \mathscr{F}_{t-1}\right]. \tag{B.27}$$

Then, with the moment states from Algorithm 1,

$$\begin{aligned}\mu_t^f \le\ & -\frac{\gamma}{8} I_t \sum_{i=1}^d \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\ & + C\left(\frac{L\gamma(1-\beta_2)H}{\epsilon} + L\gamma^2\right) I_t \mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right] \\ & + C\left(\frac{L^2\gamma^3}{\epsilon} + L\gamma^2(1-\beta_2)^2\right) I_t \left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2 + C\gamma(1-\beta_2)\sqrt{LH}\, I_t \left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|.\end{aligned} \tag{B.28}$$

Proof. We estimate the conditional drift μ_t^f defined in (B.27). At an active state put $e_t = \nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}_t)$. Smoothness and (B.16) give

$$\|e_t\| \le CL\gamma \left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\| \le \frac{CL\gamma}{\sqrt{1-\beta_2}}, \qquad \|\nabla f(\mathbf{y}_t)\| \le C\sqrt{LH}.$$

The main signal satisfies, by (B.23),

$$\begin{aligned}& -\gamma\left\langle \nabla f(\mathbf{x}_t), \mathbb{E}\left[\frac{g_t}{\sqrt{V_t}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right\rangle \\ & \quad \le -\frac{\gamma}{2}\sum_{i=1}^d \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\ & \qquad + \frac{C\gamma(1-\beta_2)L\Delta_0}{\epsilon}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right].\end{aligned}$$

For the gradient discrepancy, (B.22) and weighted Young's inequality yield

$$\begin{aligned}& \gamma\left|\left\langle e_t, \mathbb{E}\left[\frac{g_t}{\sqrt{V_t}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right\rangle\right| \\ & \quad \le \frac{\gamma}{8}\sum_{i=1}^d \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\ & \qquad + C\gamma \sum_{i=1}^d \frac{e_{t,i}^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\ & \qquad + C\gamma(1-\beta_2)L\Delta_0 \sum_{i=1}^d \frac{1}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \\ & \quad \le \frac{\gamma}{8}\sum_{i=1}^d \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} + \frac{CL^2\gamma^3}{\epsilon}\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2 \\ & \qquad + \frac{C\gamma(1-\beta_2)L\Delta_0}{\epsilon}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right].\end{aligned}$$

For its noise part the two Young factors are

$$\frac{|e_{t,i}|}{\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right)^{1/2}},$$

$$\text{and} \quad \frac{\sqrt{1-\beta_2}\,\sigma}{\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right)^{1/2}} \left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^{1/2}.$$

For the momentum correction, (B.25) and the same weighted inequality give

$$\begin{aligned}
&|\langle \nabla f(\mathbf{x}_t), \mathbb{E}[J_t \mid \mathscr{F}_{t-1}]\rangle| \\
&\quad\le \frac{\gamma}{8}\sum_{i=1}^{d} \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\
&\qquad + C\gamma(1-\beta_2)\sum_{i=1}^{d}\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right)\left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^2 \\
&\quad\le \frac{\gamma}{8}\sum_{i=1}^{d} \frac{(\partial_i f(\mathbf{x}_t))^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} + \frac{CL\gamma(1-\beta_2)H}{\epsilon}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2\,\middle|\,\mathscr{F}_{t-1}\right],
\end{aligned}$$

and

$$\begin{aligned}
&|\langle e_t, \mathbb{E}[J_t \mid \mathscr{F}_{t-1}]\rangle| \\
&\quad\le C\gamma\sum_{i=1}^{d} \frac{e_{t,i}^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon} \\
&\qquad + C\gamma(1-\beta_2)\sum_{i=1}^{d}\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right)\left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^2 \\
&\quad\le \frac{CL^2\gamma^3}{\epsilon}\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2 + \frac{CL\gamma(1-\beta_2)H}{\epsilon}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned}$$

Here (B.24) gives, for each coordinate,

$$\begin{aligned}
&\left(\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)(\partial_i f(\mathbf{x}_t))^2} + \epsilon\right)\left(\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right]\right)^2 \\
&\quad\le \frac{2S_H}{\epsilon}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad\le \frac{CLH}{\epsilon}\mathbb{E}\left[\frac{g_{t,i}^2}{(\sqrt{V_{t,i}}+\epsilon)^2}\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned}$$

Predictability and (B.19) imply

$$|\langle \nabla f(\mathbf{y}_t), F_t\rangle| \le C\gamma(1-\beta_2)\sqrt{LH}\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|.$$

Finally, the conditional smoothness remainder is bounded by $L/2$ times the second moment in (B.26). The three signal coefficients leave $1/2 - 1/8 - 1/8 = 1/4$; we weaken this to $1/8$ in the conclusion.

For every integer t, the exact stopping identity is

$$f(\mathbf{y}_{(t+1)\wedge\tau_H}) - f(\mathbf{y}_{t\wedge\tau_H}) = I_t\big(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)\big). \tag{B.29}$$

Multiplying the estimate at an active state by I_t proves (B.28) for μ_t^f defined in (B.27). The identity includes the draw from $\tau_H - 1$ to the first exited state. For the first two positive groups in (B.28),

one may take $C = 2^6(K_\star/d)^2$, with $K_\star$ from (B.2). To verify this choice, use Young coefficients $1/8, 2$ for pairs involving the signal and $1/2, 1/2$ for pairs involving the discrepancy and either the noise or J_t. The main signal and discrepancy noise then contribute $\sigma^2/(LH)$ to the coefficient of $L\gamma(1-\beta_2)H/\epsilon$. The two J_t remainders contribute at most $20\beta_1^2 S_H/((1-\beta_1^2)LH)$ by (B.24) and the explicit J_t coefficient in the proof of Lemma B.7. Their sum is bounded by

$$\frac{\sigma^2}{LH} + \frac{20}{1-\beta_1}\left(2 + \frac{\sigma^2}{16L\Delta_0}\right) \le 5\left(\frac{K_\star}{d}\right)^2.$$

Here $H \ge 16\Delta_0$, $K_\star/d \ge 3/(1-\beta_1)$, and $\sigma^2/((1-\beta_1)L\Delta_0) \le K_\star/d$ suffice. The three discrepancy remainders have total coefficient $3[\beta_1/(1-\beta_1)]^2 \le 3(K_\star/d)^2$ for $L^2\gamma^3/\epsilon$; the second moment coefficients in (B.26) give $10K_\star/d$ and $2(K_\star/d)^2$ for the smoothness terms.

For the last positive term in (B.28), one may take $C = 4\sqrt{K_\star}(K_\star/d)$. Indeed, Lemma B.4, the local condition on the step size, and $H \ge 16\Delta_0$ give

$$\frac{\|\nabla f(\mathbf{y}_t)\|}{\sqrt{LH}} \le \sqrt{2} + \frac{\sqrt{2K_\star}}{4} \le 2\sqrt{K_\star}.$$

Multiply this bound by the coefficient $\sqrt{2}\beta_1/(1-\beta_1) \le \sqrt{2}K_\star/d$ in the predictable F_t bound. This proves the stated choice separately from the terms involving squared norms. □

Lemma B.9 (Stopped height budget)**.** *Assume* (B.11) *and* $0 < \gamma \le \sqrt{(1-\beta_2)\Delta_0/L}$*. For the height* $H \ge 16\Delta_0$ *in* (B.12)*,* $T \in \mathbb{N}_+$*, and* $0 < \delta < 1$*, use the confidence budget* $B_{H,\delta}$ *and logarithmic factor* ℓ_δ *from* (B.20)*, and the preconditioned gradient energy* W_n *from* (B.15)*. An event of probability at least* $1 - 3\delta/16$ *satisfies, simultaneously for* $0 \le n \le T$*,*

$$\begin{aligned} &f(\mathbf{y}_{n\wedge\tau_H}) - f_\star + \frac{\gamma}{8}W_n \\ &\quad\le \Delta_0 + C\Bigg[\left(\frac{L\gamma(1-\beta_2)H}{\epsilon} + L\gamma^2 + \frac{L^2\gamma^3}{\epsilon}\right)B_{H,\delta} \\ &\qquad\qquad + \gamma(1-\beta_2)\sqrt{LH}\,\sqrt{TB_{H,\delta}} + \gamma\sqrt{LH}\,\sqrt{B_{H,\delta}\ell_\delta} + \frac{\gamma\sqrt{LH}\,\ell_\delta}{\sqrt{1-\beta_2}}\Bigg]. \end{aligned} \tag{B.30}$$

The coefficient in (B.30) *can be chosen as* $C = 2^{13}K_\star^4/d^3$*, with* $K_\star$ *defined in* (B.2)*.*

Proof. On the event from Lemma B.5, the normalized energies A_T, P_T in (B.15) are controlled by $B_{H,\delta}$ from (B.20). For the conditional increments μ_t^f in (B.27), sum (B.28) and use (B.21). Since $1-\beta_2 \le 1$, for every $n \le T$ this gives

$$\sum_{t=0}^{n-1}\mu_t^f \le -\frac{\gamma}{8}W_n + C\left[\left(\frac{L\gamma(1-\beta_2)H}{\epsilon} + L\gamma^2 + \frac{L^2\gamma^3}{\epsilon}\right)B_{H,\delta} + \gamma(1-\beta_2)\sqrt{LH}\,\sqrt{TB_{H,\delta}}\right].$$

With $K_\star$ from (B.2), $C = 2^{13}K_\star^4/d^3$ suffices: the terms involving squared norms have coefficient $2^6(K_\star/d)^2\,2^7K_\star^2/d = 2^{13}K_\star^4/d^3$. The predictable F_t term instead uses the bound involving the square root of P_T and has coefficient $4\sqrt{K_\star}(K_\star/d)\sqrt{2^7K_\star^2/d} \le 2^{13}K_\star^4/d^3$. To control the fluctuations, center each stopped objective increment by μ_t^f from (B.27):

$$Z_t = f(\mathbf{y}_{(t+1)\wedge\tau_H}) - f(\mathbf{y}_{t\wedge\tau_H}) - \mu_t^f.$$

These variables have conditional mean zero. At every active state before the draw, the proof of Lemma B.8 gives $\|\nabla f(\mathbf{y}_t)\| \le 2\sqrt{K_\star LH}$. By Lemma B.4 and the local condition on the step size, $L\|\mathbf{y}_{t+1} - \mathbf{y}_t\|/2 \le (5/8)\sqrt{K_\star LH}$ because $H \ge 16\Delta_0$. Smoothness therefore gives

$$|f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)| \le 3\sqrt{K_\star LH}\,\|\mathbf{y}_{t+1} - \mathbf{y}_t\|.$$

Using (B.29) and (B.26), we obtain the deterministic range and predictable variance bounds below, with V_{t-1}, V_t the second moment states from Algorithm 1:

$$|Z_t| \le 30K_\star\gamma\sqrt{\frac{LH}{1-\beta_2}}, \qquad \mathbb{E}[Z_t^2 \mid \mathscr{F}_{t-1}] \le r_t,$$

$$r_t := 180K_\star\left(\frac{K_\star}{d}\right)^2 LH\gamma^2 I_t\left(\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2\middle|\mathscr{F}_{t-1}\right] + (1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2\right).$$

For the variance estimate, conditional centering gives $\mathbb{E}[Z_t^2 \mid \mathscr{F}_{t-1}] \le I_t\mathbb{E}[(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t))^2 \mid \mathscr{F}_{t-1}]$. The two second moment coefficients in (B.26) are each at most $20(K_\star/d)^2$, which proves the displayed choice of r_t, using $K_\star$ from (B.2). For $B_{H,\delta}$ in (B.20), set $B_0 = 23040(K_\star^5/d^3)LH\gamma^2 B_{H,\delta}$. The event controlling the normalized energies gives $\sum_{t=0}^{T-1} r_t \le B_0$. Define the predictable indicators

$$\chi_t = 1_{\{\sum_{j=0}^{t} r_j \le B_0\}}.$$

Because r_t is $\mathscr{F}_{t-1}$-measurable and nonnegative, the martingale differences $\chi_t Z_t$ have total conditional variance at most B_0 on every path. On the event controlling the normalized energies, $\chi_t = 1$ for all $0 \le t < T$. Apply Lemma B.1 to $\chi_t Z_t$ with failure probability $\delta/16$. Since $\log(32/\delta) \le \ell_\delta$, its conclusion and the energy event imply

$$\max_{0\le n\le T}\left|\sum_{t=0}^{n-1} Z_t\right| \le \gamma\sqrt{LH}\left(\sqrt{46080}\,\frac{K_\star^{5/2}}{d^{3/2}}\sqrt{B_{H,\delta}\ell_\delta} + \frac{20K_\star\ell_\delta}{\sqrt{1-\beta_2}}\right).$$

The energy event costs $\delta/8$ and the martingale event costs $\delta/16$, for total failure probability at most $3\delta/16$. Summing the stopped objective increments, using $\mathbf{y}_0 = \mathbf{x}_0$ and $f(\mathbf{x}_0) - f_\star \le \Delta_0$, proves (B.30). The martingale inequality applies under the original law with a predictable variance cutoff. Finally, $K_\star/d \ge 3$ from (B.2) implies that both martingale coefficients $\sqrt{46080}K_\star^{5/2}/d^{3/2}$ and $20K_\star$ are at most $2^{13}K_\star^4/d^3$, as required in (B.30). □

B.3.5 Height control with tuning independent of confidence

Proposition B.1 (Height and preconditioned gradient energy)**.** *Use the coefficients c_{β_2}, c_γ from* (B.3) *and* (B.4)*, which depend only on $d, \beta_1, \sigma/\sqrt{L\Delta_0}, v/(L\Delta_0)$. For $T \in \mathbb{N}_+$ and $\epsilon > 0$, choose*

$$1-\beta_2 = \frac{c_{\beta_2}}{\sqrt{T}}, \qquad \gamma = \frac{c_\gamma \min\{\epsilon, \sqrt{L\Delta_0}\}}{L\sqrt{T}}.$$

These parameters satisfy (B.11) *and $\gamma \le \sqrt{(1-\beta_2)\Delta_0/L}$. For every $0 < \delta < 1$ such that $\Lambda_\delta = \log(128d/\delta) \le \sqrt{T}$, take $H = 16\Delta_0\Lambda_\delta$ in* (B.12)*. With probability at least $1 - 3\delta/16$,*

$$\tau_H > T, \qquad W_T \le \frac{H}{\gamma}, \qquad f(\mathbf{y}_{n\wedge\tau_H}) - f_\star + \frac{\gamma}{8}W_n \le \frac{H}{8} \quad (0 \le n \le T). \tag{B.31}$$

The conclusion remains valid if c_γ is decreased.

Proof. Use c_{β_2}, c_γ from (B.3) and (B.4). The stated tuning gives

$$1-\beta_2 \le \frac{1-\beta_1^2}{4}, \qquad \frac{\gamma}{\sqrt{1-\beta_2}} = \frac{c_\gamma\min\{\epsilon,\sqrt{L\Delta_0}\}}{L\sqrt{c_{\beta_2}}\,T^{1/4}} \le \frac{c_\gamma}{\sqrt{c_{\beta_2}}}\sqrt{\frac{\Delta_0}{L}}.$$

With c_γ from (B.4), (B.2) and (B.3) give $c_{\beta_2} \ge d/(4K_\star)$, and therefore

$$\frac{c_\gamma^2}{c_{\beta_2}} \le 2^{-46}\frac{d^9}{K_\star^{11}} \le 1, \qquad c_\gamma \le \frac{d^3 c_{\beta_2}}{2^{18}K_\star^4}.$$

Also $c_\gamma \le 1$ by (B.4). Hence the local assumptions hold for every $T \ge 1$.

Write $\Lambda = \Lambda_\delta = \log(128d/\delta)$, using the confidence factor in (B.1). For $0 < \delta < 1$ and $d \ge 1$, one has $\Lambda \ge \log 128 > 1$. Use Lemma B.9 with $H = 16\Delta_0\Lambda$, the height in (B.12), and the same

confidence level δ. For the budget in (B.20), $\ell_\delta \le \Lambda$ and $\log(H/\Delta_0) = \log(16\Lambda) \le \Lambda$. Hence $\Lambda \le \sqrt{T}$ gives

$$B_{H,\delta} \le T + \frac{2\Lambda\sqrt{T}}{c_{\beta_2}} \le C_B T, \qquad C_B = 1 + \frac{2}{c_{\beta_2}}. \tag{B.32}$$

The coefficient C_B in (B.32) is fixed independently of T, ϵ, δ. With c_{β_2} from (B.3), the six error terms in (B.30) satisfy

$$\begin{aligned}
\frac{L\gamma(1-\beta_2)H}{\epsilon} B_{H,\delta} &\le c_\gamma c_{\beta_2} C_B H, \\
L\gamma^2 B_{H,\delta} &\le c_\gamma^2 C_B \Delta_0, \\
\frac{L^2\gamma^3}{\epsilon} B_{H,\delta} &\le c_\gamma^3 C_B \Delta_0, \\
\gamma(1-\beta_2)\sqrt{LH}\,\sqrt{T B_{H,\delta}} &\le c_\gamma c_{\beta_2} \sqrt{C_B}\sqrt{\Delta_0 H}, \\
\gamma\sqrt{LH}\,\sqrt{B_{H,\delta}\ell_\delta} &\le c_\gamma \sqrt{C_B}\sqrt{\Delta_0 H\Lambda}, \\
\frac{\gamma\sqrt{LH}\,\ell_\delta}{\sqrt{1-\beta_2}} &\le \frac{c_\gamma}{\sqrt{c_{\beta_2}}}\sqrt{\Delta_0 H\Lambda}.
\end{aligned}$$

These inequalities use $\min\{\epsilon, \sqrt{L\Delta_0}\} \le \sqrt{L\Delta_0}$, $\min\{\epsilon, \sqrt{L\Delta_0}\}/\epsilon \le 1$, $\min\{\epsilon, \sqrt{L\Delta_0}\}^3/\epsilon \le L\Delta_0$, and $\ell_\delta/T^{1/4} \le \sqrt{\Lambda}$. Because $H = 16\Delta_0\Lambda$, one has $\sqrt{\Delta_0 H} \le H/4$ and $\sqrt{\Delta_0 H\Lambda} = H/4$. Also $\Delta_0 \le H/16$. Use the parameter scale $K_\star$ from (B.2) and the fixed coefficients from (B.3) and (B.4). Since $0 < c_{\beta_2} \le 1/4$ and $C_B = 1 + 2/c_{\beta_2}$ by (B.32),

$$c_{\beta_2} C_B + \frac{C_B}{8} + \frac{(c_{\beta_2}+1)\sqrt{C_B}}{4} + \frac{1}{4\sqrt{c_{\beta_2}}} \le \frac{2}{c_{\beta_2}}.$$

The six errors in (B.30), including its coefficient $2^{13}K_\star^4/d^3$, are thus at most

$$\frac{2^{14}K_\star^4 c_\gamma H}{d^3 c_{\beta_2}} \le \frac{d}{256K_\star} H \le \frac{H}{256}.$$

Adding the initial gap $\Delta_0 \le H/16$ gives $H/8$. The same choice (B.4) ensures $\|\mathbf{x}_t - \mathbf{y}_t\| \le 2\sqrt{K_\star} c_\gamma \sqrt{\Delta_0/L}/\sqrt{c_{\beta_2}} \le 2^{-22} d^{9/2} K_\star^{-5}\sqrt{\Delta_0/L} \le \sqrt{\Delta_0/L}$ on every trajectory.

Substituting these error bounds into (B.30) gives the prefix estimate (B.31) simultaneously for all $n \le T$ on the event of Lemma B.9. If $\tau_H \le T$, it gives $f(\mathbf{y}_{\tau_H}) - f_\star \le H/8$. By Lemma B.3 and the offset bound,

$$\sqrt{f(\mathbf{x}_{\tau_H}) - f_\star} \le \sqrt{f(\mathbf{y}_{\tau_H}) - f_\star} + \sqrt{\frac{L}{2}}\,\|\mathbf{x}_{\tau_H} - \mathbf{y}_{\tau_H}\| \le \sqrt{\frac{H}{8}} + \sqrt{\frac{\Delta_0}{2}} < \sqrt{H}.$$

This contradicts the definition of τ_H. Hence $\tau_H > T$. Finally, $f \ge f_\star$ and the prefix estimate at $n = T$ yield $W_T \le H/\gamma$. Every error above decreases when c_γ decreases, proving the last assertion. □

B.3.6 Height control with tuning dependent on confidence

Lemma B.10 (Height and preconditioned gradient energy with tuning dependent on confidence)**.** *Use the objective and oracle assumptions of Section B.3, the positive initial gap bound Δ_0 from* (B.1)*, and the parameter scale $K_\star$ from* (B.2)*. Let $T \in \mathbb{N}_+$ and $0 < \delta < 1$ satisfy $\Lambda_\delta \le T$, where Λ_δ is defined in* (B.1)*. Use the corresponding branch of* (B.8)*, with c_{β_2} and c_γ from* (B.3) *and* (B.4)*, and take $H = 16\Delta_0$ in* (B.12)*. With probability at least $1 - 3\delta/16$, the stopping time and preconditioned gradient energy from* (B.12) *and* (B.15) *satisfy*

$$\tau_H > T, \qquad W_T \le \frac{16\Delta_0}{\gamma}.$$

On the same event, the auxiliary trajectory from (B.14) *satisfies*

$$f(\mathbf{y}_n) - f_\star + \frac{\gamma}{8} W_n \le 2\Delta_0 \qquad (0 \le n \le T). \tag{B.33}$$

Proof. The definitions (B.2), (B.3), and (B.4) give

$$K_\star \ge d \ge 1, \qquad \frac{d}{4K_\star} \le c_{\beta_2} \le \frac{1}{4}, \qquad c_\gamma = \frac{d^5}{2^{24}K_\star^6} \le 1.$$

Also $\Lambda_\delta \ge \log 128 > 1$ by (B.1). Since $\Lambda_\delta \le T$, the parameters in (B.8) satisfy

$$0 < 1-\beta_2 \le c_{\beta_2} \le \frac{1-\beta_1^2}{2}.$$

Moreover,

$$\frac{L\gamma^2}{(1-\beta_2)\Delta_0} = \frac{c_\gamma^2}{c_{\beta_2}} \frac{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}{L\Delta_0} \frac{1}{\sqrt{T}\Lambda_\delta^{3/2}} \le 2^{-46}\frac{d^9}{K_\star^{11}} \le 1.$$

Thus the conditions of Lemma B.9 hold.

For $H = 16\Delta_0$, the quantities in (B.20) obey $\ell_\delta \le \Lambda_\delta$ and $\log(H/\Delta_0) = \log 16 \le \Lambda_\delta$. The first branch of (B.8) therefore gives

$$B_{H,\delta} \le T + \frac{2\sqrt{T\Lambda_\delta}}{c_{\beta_2}} \le \left(1 + \frac{2}{c_{\beta_2}}\right) T \le 9\frac{K_\star}{d}T.$$

Substitute this bound and $H = 16\Delta_0$ into the six terms of (B.30). Using the coefficients from (B.3) and (B.4), each term satisfies

$$\frac{L\gamma(1-\beta_2)H}{\epsilon} B_{H,\delta} \le 144\frac{K_\star}{d}c_\gamma c_{\beta_2}\Delta_0 \le 36\frac{K_\star}{d}c_\gamma\Delta_0,$$

$$L\gamma^2 B_{H,\delta} \le \frac{9K_\star c_\gamma^2\Delta_0}{d\Lambda_\delta} \le 9\frac{K_\star}{d}c_\gamma^2\Delta_0,$$

$$\frac{L^2\gamma^3}{\epsilon} B_{H,\delta} \le \frac{9K_\star c_\gamma^3\Delta_0}{d\sqrt{T}\Lambda_\delta^{3/2}} \le 9\frac{K_\star}{d}c_\gamma^3\Delta_0,$$

$$\gamma(1-\beta_2)\sqrt{LH}\sqrt{TB_{H,\delta}} \le 12\sqrt{\frac{K_\star}{d}}c_\gamma c_{\beta_2}\Delta_0 \le 3\sqrt{\frac{K_\star}{d}}c_\gamma\Delta_0,$$

$$\gamma\sqrt{LH}\sqrt{B_{H,\delta}\ell_\delta} \le 12\sqrt{\frac{K_\star}{d}}c_\gamma\Delta_0\sqrt{\frac{\ell_\delta}{\Lambda_\delta}} \le 12\sqrt{\frac{K_\star}{d}}c_\gamma\Delta_0,$$

$$\frac{\gamma\sqrt{LH}\ell_\delta}{\sqrt{1-\beta_2}} \le \frac{4c_\gamma\Delta_0}{\sqrt{c_{\beta_2}}}\left(\frac{\Lambda_\delta}{T}\right)^{1/4} \le 8\sqrt{\frac{K_\star}{d}}c_\gamma\Delta_0.$$

These estimates use $\min\{\epsilon, \sqrt{L\Delta_0}\}/\epsilon \le 1$, $\min\{\epsilon, \sqrt{L\Delta_0}\}^2 \le L\Delta_0$, $\min\{\epsilon, \sqrt{L\Delta_0}\}^3/\epsilon \le L\Delta_0$, and $\ell_\delta \le \Lambda_\delta \le T$ from (B.20) and the lemma's hypothesis. Since $K_\star/d \ge 1$ by (B.2) and $c_\gamma \le 1$ by (B.4), their sum is at most $77(K_\star/d)c_\gamma\Delta_0$. The coefficient in (B.30), together with (B.4), yields

$$\frac{2^{13}K_\star^4}{d^3} \cdot 77\frac{K_\star}{d}c_\gamma\Delta_0 = \frac{77d\Delta_0}{2048K_\star} < \Delta_0.$$

Consequently, Lemma B.9 gives an event of probability at least $1 - 3\delta/16$ on which

$$f(\mathbf{y}_{n\wedge\tau_H}) - f_\star + \frac{\gamma}{8}W_n \le 2\Delta_0 \qquad (0 \le n \le T).$$

The explicit estimate for reachable filter states in the proof of Lemma B.4, with $K_\star$ from (B.2) and the auxiliary trajectory from (B.14), holds on every sample path and gives

$$\begin{aligned}\|\mathbf{x}_t - \mathbf{y}_t\| &\le \frac{2\sqrt{K_\star}\gamma}{\sqrt{1-\beta_2}} \\ &\le \frac{2\sqrt{K_\star}c_\gamma}{\sqrt{c_{\beta_2}}}\sqrt{\frac{\Delta_0}{L}}\frac{1}{T^{1/4}\Lambda_\delta^{3/4}} \le 2^{-22}\frac{d^{9/2}}{K_\star^5}\sqrt{\frac{\Delta_0}{L}} \le \sqrt{\frac{\Delta_0}{L}}.\end{aligned}$$

Here we used (B.8) and the coefficient bounds from (B.3) and (B.4). If $\tau_H \le T$ on the preceding event, the nonnegativity of W_n and Lemma B.3 imply

$$\sqrt{f(\mathbf{x}_{\tau_H}) - f_\star} \le \sqrt{f(\mathbf{y}_{\tau_H}) - f_\star} + \sqrt{\frac{L}{2}}\|\mathbf{x}_{\tau_H} - \mathbf{y}_{\tau_H}\| \le \sqrt{2\Delta_0} + \sqrt{\frac{\Delta_0}{2}} < 4\sqrt{\Delta_0} = \sqrt{H}.$$

This contradicts the definition in (B.12). Hence $\tau_H > T$, and the stopped prefix bound is (B.33). Finally, $f \ge f_\star$ in that bound at $n = T$ gives $W_T \le 16\Delta_0/\gamma$. □

B.4 JOINT CONTROL OF HIGH ACCUMULATORS AND INTERVAL MOTION

We use the objective, oracle, and iterates of Section B.3, with an arbitrary deterministic height $H \geq 16\Delta_0$. The second moment states V_{t-1} and V_t are the incoming and updated states, respectively, of Algorithm 1. The stopping time τ_H, indicator of active states I_t, and second moment scale $S_H = 2LH+\sigma^2$ are defined in (B.12); the stopped preconditioned gradient energy W_n is defined in (B.15). At an active state $t < \tau_H$, one has $\|\nabla f(\mathbf{x}_t)\|^2 \leq 2LH$ and $\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] \leq S_H$. All constants in this subsection depend only on $d, \beta_1, \sigma^2/(L\Delta_0), v/(L\Delta_0)$, not on $H, \epsilon, T, \delta, \beta_2, \gamma$. For their quantitative values, use $K_\star$ from (B.2); in particular, $K_\star/d \geq \max\{1, 1+\sigma^2/(L\Delta_0), v/(L\Delta_0), (1-\beta_1)^{-1}\}$. Assume

$$\begin{aligned} 0 < 1-\beta_2 \leq \frac{1-\beta_1^2}{2}, \qquad 0 < \gamma \leq \sqrt{\frac{(1-\beta_2)\Delta_0}{L}}, \\ k := \frac{L^2\gamma^2}{(1-\beta_2)^2 \min\{\epsilon, \sqrt{L\Delta_0}\}^2} \leq k_0, \qquad k_0 = 2^{-13}, \end{aligned} \tag{B.34}$$

This value of k_0 is independent of H.

Lemma B.11 (Denominator reserve)**.** *Fix $R \geq S_H$ and define*

$$\Psi_t = \sum_{j=1}^{d} \log\left(1 + \frac{R}{V_{t-1,j}+\epsilon^2}\right). \tag{B.35}$$

Then $0 \leq \Psi_t \leq d\log(1+R/\epsilon^2)$. At every active state,

$$\begin{aligned} \mathbb{E}[\Psi_{t+1} - \Psi_t \mid \mathscr{F}_{t-1}] &\leq \left(2d + \frac{S_H}{R}\right)(1-\beta_2) \\ &\quad - (1-\beta_2)\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \middle| \mathscr{F}_{t-1}\right], \\ \Psi_{t+1} - \Psi_t &\leq 2d(1-\beta_2) \qquad \textit{almost surely}. \end{aligned} \tag{B.36}$$

Proof. For one coordinate, the second moment update in Algorithm 1 gives

$$V_{t,j} + \epsilon^2 = \beta_2(V_{t-1,j}+\epsilon^2) + (1-\beta_2)(g_{t,j}^2 + \epsilon^2).$$

Since $\log z \geq 1 - z^{-1}$ and $-\log\beta_2 \leq 2(1-\beta_2)$,

$$\begin{aligned} \log(V_{t,j}+\epsilon^2) - \log(V_{t-1,j}+\epsilon^2) &= \log\beta_2 + \log\frac{V_{t,j}+\epsilon^2}{\beta_2(V_{t-1,j}+\epsilon^2)} \\ &\geq -2(1-\beta_2) + \frac{(1-\beta_2)(g_{t,j}^2+\epsilon^2)}{V_{t,j}+\epsilon^2} \\ &\geq -2(1-\beta_2) + \frac{(1-\beta_2)g_{t,j}^2}{(\sqrt{V_{t,j}}+\epsilon)^2}. \end{aligned}$$

The last inequality uses $(\sqrt{V_{t,j}}+\epsilon)^2 \geq V_{t,j}+\epsilon^2$. Also, $V_{t,j} - V_{t-1,j} = (1-\beta_2)(g_{t,j}^2 - V_{t-1,j})$, so

$$\begin{aligned} &\log(V_{t,j}+\epsilon^2+R) - \log(V_{t-1,j}+\epsilon^2+R) \\ &\leq \frac{(1-\beta_2)(g_{t,j}^2 - V_{t-1,j})}{V_{t-1,j}+\epsilon^2+R} \leq \frac{(1-\beta_2)g_{t,j}^2}{R}. \end{aligned}$$

Subtracting the two bounds, summing over coordinates, and taking conditional expectation proves the first part of (B.36). For the pathwise upper increment, $V_{t,j}+\epsilon^2 \geq \beta_2(V_{t-1,j}+\epsilon^2)$ and monotonicity give

$$\begin{aligned} &\log\left(1+\frac{R}{V_{t,j}+\epsilon^2}\right) - \log\left(1+\frac{R}{V_{t-1,j}+\epsilon^2}\right) \\ &\leq \log\left(1+\frac{R}{\beta_2(V_{t-1,j}+\epsilon^2)}\right) - \log\left(1+\frac{R}{V_{t-1,j}+\epsilon^2}\right) \\ &\leq -\log\beta_2 \leq 2(1-\beta_2). \end{aligned}$$

Summation proves the second part. □

For the remainder of the proof, take $R = S_H$. Choose a deterministic threshold K with $K^2 \geq \max\{v, 16S_H\}$, to be specified below. Classify coordinate i using the incoming accumulator $V_{t-1,i}$ from Algorithm 1: a state is low if $V_{t-1,i} \leq K^2$ and high otherwise. The scored states are $t < \tau_H \wedge T$. A high interval of positive length $[s, e)$ begins at $s = t + 1$ after a low scored state t, provided $s < \tau_H \wedge T$ and $V_{s-1,i} > K^2$. Its endpoint is the first subsequent low state or $\tau_H \wedge T$, whichever occurs first. A draw reaching τ_H or T enters no scored interval. All initial coordinates are low, and these intervals partition the scored high states of each coordinate.

Lemma B.12 (Capped motion of the full vector)**.** *Fix a high interval $[s, e)$ and keep $\nabla f(\mathbf{x}_s), \mathbf{y}_s$ as its anchors, with the auxiliary trajectory $\mathbf{y}_t$ from* (B.14)*. Use $K_\star$ from* (B.2) *and the gap height $H \geq 16\Delta_0$ from* (B.12)*. For $s \leq t \leq e$, set*

$$\phi_t = \min\{L^2 \left\|\mathbf{y}_t - \mathbf{y}_s\right\|^2, LH\}.$$

For k_0 in (B.34)*, every $s \leq t < e$ satisfies*

$$\begin{aligned}\mathbb{E}[\phi_{t+1} - \phi_t \mid \mathscr{F}_{t-1}] \leq \frac{1-\beta_2}{8}\phi_t + 576k(1-\beta_2)\left\|\nabla f(\mathbf{x}_s)\right\|^2 + 4996K_\star^3 L^2\gamma^2 \\ + \frac{596K_\star L^2\gamma^2 S_H}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right].\end{aligned} \tag{B.37}$$

Its pathwise upper increment obeys

$$\phi_{t+1} - \phi_t \leq \frac{2\sqrt{34}K_\star L\gamma}{\sqrt{1-\beta_2}}\sqrt{\phi_t} + \frac{34K_\star^2 L^2\gamma^2}{1-\beta_2}. \tag{B.38}$$

Proof. If $\phi_t = LH$, its increment is nonpositive. If $\phi_t < LH$, put $\mathbf{z} = \mathbf{y}_t - \mathbf{y}_s$ and $\Delta\mathbf{y}_t = \mathbf{y}_{t+1} - \mathbf{y}_t$. The cap and the expansion of the squared norm give

$$\phi_{t+1} - \phi_t \leq 2L^2\langle \mathbf{z}, \Delta\mathbf{y}_t\rangle + L^2\left\|\Delta\mathbf{y}_t\right\|^2.$$

The bounds in Lemmas B.4 and B.7, expressed using $K_\star$ from (B.2), give the following estimates:

$$\begin{aligned}\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2 \leq \frac{2K_\star}{1-\beta_2}, \quad \left\|\mathbf{x}_t - \mathbf{y}_t\right\|^2 \leq \frac{2K_\star\gamma^2}{1-\beta_2}, \\ \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \leq \frac{2K_\star\gamma^2}{1-\beta_2}, \qquad \left\|\Delta\mathbf{y}_t\right\|^2 \leq \frac{34K_\star^2\gamma^2}{1-\beta_2}.\end{aligned}$$

$$\begin{aligned}\left\|\mathbb{E}[\Delta\mathbf{y}_t \mid \mathscr{F}_{t-1}]\right\|^2 \leq \frac{6\gamma^2}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \\ + \frac{18K_\star\gamma^2(1-\beta_2)S_H}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right] \\ + 6K_\star^2\gamma^2(1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\left\|\Delta\mathbf{y}_t\right\|^2 \mid \mathscr{F}_{t-1}] \leq 20K_\star\gamma^2\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right] \\ + 4K_\star^2\gamma^2(1-\beta_2)^2\left\|\frac{M_{t-1}}{\sqrt{V_{t-1}}+\epsilon}\right\|^2.\end{aligned}$$

Indeed, the coordinate filter bound is $M_{t-1,i}^2/V_{t-1,i} \leq 2(1-\beta_1)^2/((1-\beta_1^2)(1-\beta_2))$. The predictable correction and the correction from the current sample in (B.18) have coefficients $\sqrt{2}\beta_1/(1-\beta_1)$ and $2\beta_1/\sqrt{1-\beta_1^2}$, respectively. The squared triangle inequality and (B.23) then give the displayed mean coefficients $6, 18K_\star, 6K_\star^2$; the second moment coefficients are bounded by $20K_\star$ and $4K_\star^2$. These estimates use $d/(1-\beta_1) \leq K_\star$ and $(1-\beta_1)^{-1} \leq K_\star$. Smoothness and the two auxiliary offsets imply

$$\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq 3\left\|\nabla f(\mathbf{x}_s)\right\|^2 + 3\phi_t + \frac{24K_\star L^2\gamma^2}{1-\beta_2}.$$

Apply Young's inequality in the form

$$2L^2\langle \mathbf{z}, \mathbb{E}[\Delta \mathbf{y}_t \mid \mathscr{F}_{t-1}]\rangle \le \frac{1-\beta_2}{32}\phi_t + \frac{32L^2}{1-\beta_2}\left\|\mathbb{E}[\Delta \mathbf{y}_t \mid \mathscr{F}_{t-1}]\right\|^2.$$

The preceding estimates give

$$\begin{aligned}\mathbb{E}[\phi_{t+1}-\phi_t \mid \mathscr{F}_{t-1}] &\le (1-\beta_2)\left(\frac{1}{32}+576k\right)\phi_t + 576k(1-\beta_2)\left\|\nabla f(\mathbf{x}_s)\right\|^2 \\ &\quad + (4608K_\star k + 388K_\star^3)L^2\gamma^2 \\ &\quad + \frac{596K_\star L^2\gamma^2 S_H}{\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\mathbb{E}\left[\left\|\frac{g_t}{\sqrt{V_t}+\epsilon}\right\|^2 \,\middle|\, \mathscr{F}_{t-1}\right].\end{aligned}$$

Here $1-\beta_2 \le 1/2$ and, from the height in (B.12), $S_H \ge 32L\Delta_0 \ge \min\{\epsilon, \sqrt{L\Delta_0}\}^2$. Also, $L^2\gamma^2/((1-\beta_2)\min\{\epsilon, \sqrt{L\Delta_0}\}^2) = k(1-\beta_2)$. Since $k \le 2^{-13}$, one has $1/32 + 576k \le 1/8$ and $4608K_\star k + 388K_\star^3 \le 4996K_\star^3$. This proves (B.37). For (B.38), apply the same expansion of the squared norm and the pathwise bound $\|\Delta \mathbf{y}_t\| \le \sqrt{34}K_\star\gamma/\sqrt{1-\beta_2}$. Both bounds also hold when $\phi_t = LH$. □

Lemma B.13 (Joint potential on a high interval)**.** *There are fixed positive constants c_A, c_2, c_3, uniform in $H \ge 16\Delta_0$ and $K^2 \ge \max\{v, 16S_H\}$, with the following property. Use $K_\star$ from* (B.2) *and k from* (B.34)*, where H is the gap height from* (B.12)*, and set $\lambda_H = 1/(12LH)$. Then*

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda_H(1-\beta_2)\sum_{t=s}^{e-1}\left[\left(\partial_i f(\mathbf{x}_s)\right)^2 + \phi_t\right]\right)\,\middle|\,\mathscr{F}_{s-1}\right] \\ &\quad\le 1 + \frac{1}{L\Delta_0}\left(\frac{V_{s-1,i}}{\beta_2 K^2}-1\right)\Big[\left(\partial_i f(\mathbf{x}_s)\right)^2 + c_2 k\left\|\nabla f(\mathbf{x}_s)\right\|^2 \\ &\qquad\qquad + c_A k(1-\beta_2)S_H(c_3+\Psi_s)\Big].\end{aligned} \tag{B.39}$$

Proof. At every high state before sampling, the second moment update in Algorithm 1 yields

$$\frac{V_{t-1,i}}{K^2} - 1 \ge 0, \qquad \frac{V_{t,i}}{\beta_2 K^2} - 1 = \frac{V_{t-1,i}}{K^2} - 1 + \frac{(1-\beta_2)g_{t,i}^2}{\beta_2 K^2} \ge 0.$$

The second expression is nonnegative even when that draw ends the interval. Therefore the expression at the next state

$$\begin{aligned}1 + \frac{1}{L\Delta_0}\left(\frac{V_{t,i}}{\beta_2 K^2}-1\right)\Big[&\left(\partial_i f(\mathbf{x}_s)\right)^2 + \phi_{t+1} + c_2 k\left\|\nabla f(\mathbf{x}_s)\right\|^2 \\ &+ c_A k(1-\beta_2)S_H(c_3+\Psi_{t+1})\Big]\end{aligned}$$

can only decrease when it is replaced by one at $t+1 = e$. Every bracket evaluated at the next state here and below uses the same anchors $\mathbf{x}_s, \mathbf{y}_s$, including on the exit draw.

Combining (B.36) and (B.37), with $K_\star$ from (B.2), choose

$$c_A = 2^{10}K_\star. \tag{B.40}$$

Since $2^{10}K_\star \ge 596K_\star$, the reserve cancels the positive term from the sample energy in (B.37). Since $R = S_H$, the remaining reserve drift is $(2d+1)c_A k(1-\beta_2)^2 S_H$. Moreover $L^2\gamma^2 \le k(1-\beta_2)^2 S_H$, so this gives

$$\begin{aligned}&\mathbb{E}\left[\phi_{t+1} - \phi_t + c_A k(1-\beta_2)S_H(\Psi_{t+1}-\Psi_t)\,|\,\mathscr{F}_{t-1}\right] \\ &\quad\le \frac{1-\beta_2}{8}\phi_t + 576k(1-\beta_2)\left\|\nabla f(\mathbf{x}_s)\right\|^2 \\ &\qquad + 8K_\star^2 c_A k(1-\beta_2)^2 S_H.\end{aligned}$$

Here $4996K_\star^3/c_A + 2d + 1 \le 8K_\star^2$. Choose

$$c_2 = 9 \cdot 2^{10}, \qquad c_3 = 2^8 K_\star^2. \tag{B.41}$$

The last two terms are at most

$$\frac{1-\beta_2}{16}\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + c_2 k \,\|\nabla f(\mathbf{x}_s)\|^2 + c_3 c_A k(1-\beta_2) S_H\Big].$$

Since $0 \le V_{t-1,i}/K^2 - 1 \le V_{t-1,i}/(\beta_2 K^2) - 1$, it follows that

$$\begin{aligned}
&\left(\frac{V_{t-1,i}}{K^2} - 1\right) \mathbb{E}\left[\phi_{t+1} - \phi_t + c_A k(1-\beta_2) S_H(\Psi_{t+1} - \Psi_t) \,|\, \mathscr{F}_{t-1}\right] \\
&\quad \le \frac{1-\beta_2}{8}\left(\frac{V_{t-1,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \,\|\nabla f(\mathbf{x}_s)\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H(c_3 + \Psi_t)\Big].
\end{aligned}$$

Write the conditional coordinate second moments as

$$S_{t,i} = \mathbb{E}[g_{t,i}^2 \mid \mathscr{F}_{t-1}], \qquad \sum_{i=1}^d S_{t,i} \le S_H \quad \text{on } \{t < \tau_H\}, \tag{B.42}$$

where the sum bound follows from (B.13). The second moment update in Algorithm 1 gives the exact product expansion

$$\begin{aligned}
&\mathbb{E}\Bigg[\left(\frac{V_{t,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_{t+1} + c_2 k \,\|\nabla f(\mathbf{x}_s)\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H(c_3 + \Psi_{t+1})\Big] \;\Big|\; \mathscr{F}_{t-1}\Bigg] \\
&\quad - \left(\frac{V_{t-1,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \,\|\nabla f(\mathbf{x}_s)\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H(c_3 + \Psi_t)\Big] \\
&= -(1-\beta_2)\frac{V_{t-1,i} - S_{t,i}}{\beta_2 K^2}\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \,\|\nabla f(\mathbf{x}_s)\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H(c_3 + \Psi_t)\Big] \\
&\quad + \left(\frac{V_{t-1,i}}{K^2} - 1\right)\mathbb{E}\left[\phi_{t+1} - \phi_t + c_A k(1-\beta_2) S_H(\Psi_{t+1} - \Psi_t) \,|\, \mathscr{F}_{t-1}\right] \\
&\quad + \frac{1-\beta_2}{\beta_2 K^2}\mathbb{E}\Big[g_{t,i}^2\big(\phi_{t+1} - \phi_t + c_A k(1-\beta_2) S_H(\Psi_{t+1} - \Psi_t)\big) \;\Big|\; \mathscr{F}_{t-1}\Big].
\end{aligned} \tag{B.43}$$

For the term from the current sample involving ϕ, the pathwise bound (B.38) implies

$$\begin{aligned}
\frac{1-\beta_2}{\beta_2 K^2}\mathbb{E}[g_{t,i}^2(\phi_{t+1} - \phi_t) \mid \mathscr{F}_{t-1}] &\le \frac{S_H}{\beta_2 K^2}\left(2\sqrt{34} K_\star L\gamma\sqrt{1-\beta_2}\sqrt{\phi_t} + 34 K_\star^2 L^2\gamma^2\right) \\
&\le \frac{1-\beta_2}{8}\phi_t + \frac{17}{2} K_\star^2 L^2 \gamma^2.
\end{aligned}$$

For the reserve, its one-sided pathwise increment gives

$$\begin{aligned}
\frac{c_A k(1-\beta_2)^2 S_H}{\beta_2 K^2}\mathbb{E}[g_{t,i}^2(\Psi_{t+1} - \Psi_t) \mid \mathscr{F}_{t-1}] &\le \frac{2d c_A k(1-\beta_2)^3 S_H^2}{\beta_2 K^2} \\
&\le \frac{d}{8} c_A k(1-\beta_2)^2 S_H.
\end{aligned}$$

These bounds use the conditional second moment of $g_{t,i}$, including when all coordinates of the current sample are dependent. Using k from (B.34) and S_H from (B.12), we have $L^2\gamma^2 = k(1 -$

$\beta_2)^2 \min\{\epsilon, \sqrt{L\Delta_0}\}^2 \le k(1-\beta_2)^2 S_H$. Thus their sum is at most

$$\begin{aligned}
&(1-\beta_2)\left[\frac{\phi_t}{8} + c_A k(1-\beta_2) S_H \left(\frac{17K_\star}{2048} + \frac{d}{8}\right)\right] \\
&\quad \le (1-\beta_2)\left[\frac{\phi_t}{8} + \frac{K_\star c_A k(1-\beta_2) S_H}{4}\right] \\
&\quad \le \frac{1-\beta_2}{4}\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H (c_3 + \Psi_t)\Big].
\end{aligned}$$

The choices are uniform for $K^2 \ge 16 S_H$, because $S_{t,i}/(\beta_2 K^2) \le 1/8$ and $1-\beta_2 \le 1/2$. Substitution in (B.43) gives contraction coefficients $7/8$ for $V_{t-1,i}/(\beta_2 K^2) - 1$ and $5/8$ for the constant term, both at least $1/2$. Divide the product bound by $L\Delta_0$. Replacing the expression at the next state by one at the endpoint then yields

$$\begin{aligned}
&\mathbb{E}\Bigg[1 + \frac{1_{\{t+1<e\}}}{L\Delta_0}\left(\frac{V_{t,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_{t+1} + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H (c_3 + \Psi_{t+1})\Big] \;\Big|\; \mathscr{F}_{t-1}\Bigg] \\
&\quad \le 1 + \frac{1}{L\Delta_0}\left[\frac{V_{t-1,i}}{\beta_2 K^2} - 1 - \frac{1-\beta_2}{2}\frac{V_{t-1,i}}{\beta_2 K^2}\right]\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t \\
&\qquad\qquad + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 + c_A k(1-\beta_2) S_H (c_3 + \Psi_t)\Big].
\end{aligned}$$

At each scored state, the reward $\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t$ lies in $[0, 3LH]$ and is at most the bracket multiplying the accumulator factor. Since $H/\Delta_0 \ge 16$ by (B.12),

$$\begin{aligned}
&\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t\Big]\left\{1 + \frac{1}{L\Delta_0}\left(\frac{V_{t-1,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t \\
&\qquad\qquad + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 + c_A k(1-\beta_2) S_H (c_3 + \Psi_t)\Big]\right\} \\
&\quad \le \frac{3H}{\Delta_0}\frac{V_{t-1,i}}{\beta_2 K^2}\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H (c_3 + \Psi_t)\Big].
\end{aligned}$$

The preceding conditional decrease is therefore at least the fraction $(1-\beta_2)[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t]/(6LH)$ of the full expression at the current state. This fraction is at most $1/4$. Since $\lambda_H = 1/(12LH)$, the elementary inequality $e^{u/2}(1-u) \le 1$ for $0 \le u \le 1/4$ and predictability of the reward show that the following expression, defined for $s \le t \le e$ and held constant thereafter, is a nonnegative supermartingale:

$$\begin{aligned}
&\exp\left(\lambda_H(1-\beta_2)\sum_{r=s}^{t-1}\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_r\Big]\right) \\
&\quad \times \left\{1 + \frac{1_{\{t<e\}}}{L\Delta_0}\left(\frac{V_{t-1,i}}{\beta_2 K^2} - 1\right)\Big[\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 \\
&\qquad\qquad + c_A k(1-\beta_2) S_H (c_3 + \Psi_t)\Big]\right\}.
\end{aligned}$$

At $t = e$, the expression in braces is one; at $t = s$, use $\phi_s = 0$. Conditional iteration therefore proves (B.39).

The endpoint is a bounded stopping time. The raw accumulator on an exit draw is still at least $\beta_2 V_{t-1,i} > \beta_2 K^2$, including a simultaneous height exit. The bracket at the next state uses the same anchors, the capped displacement, and the reserve, so no gradient bound at an exited state is needed. For each fixed H and positive ϵ, that bracket has a finite deterministic upper bound and the stopped accumulator has finite first moment. The accumulated exponential reward is at most $\exp(3LH\lambda_H(1-\beta_2)T)$. These facts justify conditional iteration at the bounded random entry and endpoint times. □

B.5 Entry charges at stopping times and conversion to squared gradients

Lemma B.14 (Uncapped block rewards)**.** *With $K_\star$ from* (B.2)*, for every $0 < \rho < 1$, with probability at least $1-\rho$,*

$$\sum_{i,\,[s,e)\ \text{high}} \sum_{t=s}^{e-1} \Big(\big(\partial_i f(\mathbf{x}_s)\big)^2 + \phi_t \Big) \le \frac{24HS_H}{\Delta_0 K^2} \Bigg[2(1+c_2 k) G_T + c_A k (1-\beta_2) T S_H \times \left(c_3 + \frac{17}{2048} + d \log\left(1 + \frac{S_H}{\epsilon^2}\right)\right)\Bigg] + \frac{12dLH}{1-\beta_2} \log\frac{d}{\rho}. \tag{B.44}$$

where

$$G_T = \sum_{t=0}^{T-1} I_t \left\|\nabla f(\mathbf{x}_t)\right\|^2. \tag{B.45}$$

Each ϕ_t in this sum uses the anchors of its own interval; c_A, c_2, c_3 are defined in (B.40) *and* (B.41)*, and k is defined in* (B.34)*. The constants are uniform in the heights $H \ge 16\Delta_0$ of* (B.12) *and $K^2 \ge \max\{v, 16S_H\}$.*

Proof. At a low scored state t, use the second moment states of Algorithm 1 and let R_t be the total reward of the high interval entered by its draw, and set $R_t = 0$ if no scored interval of positive length is entered. On a positive entry at $s = t+1$, the incoming accumulator is $V_{s-1,i} = V_{t,i}$, so

$$0 \le \frac{V_{s-1,i}}{\beta_2 K^2} - 1 \le \frac{(1-\beta_2) g_{t,i}^2}{\beta_2 K^2}.$$

The actual entry step and smoothness give

$$\big(\partial_i f(\mathbf{x}_s)\big)^2 \le 2\big(\partial_i f(\mathbf{x}_t)\big)^2 + \frac{4K_\star L^2\gamma^2}{1-\beta_2},$$
$$\left\|\nabla f(\mathbf{x}_s)\right\|^2 \le 2\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \frac{4K_\star L^2\gamma^2}{1-\beta_2}.$$

The reserve definition (B.35), together with $\phi_s = 0$, $\Psi_s \le d\log(1 + S_H/\epsilon^2)$, $k \le k_0$ from (B.34), and $L^2\gamma^2/(1-\beta_2) \le k(1-\beta_2)S_H$, gives

$$\begin{aligned}
&\big(\partial_i f(\mathbf{x}_s)\big)^2 + c_2 k \left\|\nabla f(\mathbf{x}_s)\right\|^2 + c_A k (1-\beta_2) S_H (c_3 + \Psi_s) \\
&\quad \le 2\big(\partial_i f(\mathbf{x}_t)\big)^2 + 2c_2 k \left\|\nabla f(\mathbf{x}_t)\right\|^2 \\
&\qquad + c_A k (1-\beta_2) S_H \left(c_3 + \frac{17}{2048} + d\log\left(1 + \frac{S_H}{\epsilon^2}\right)\right).
\end{aligned}$$

Indeed, the step error divided by $c_A k(1-\beta_2)S_H$ is at most $4K_\star(1 + c_2 k_0)/c_A = 17/2048$, using $c_2 k_0 = 9/8$ and the constants in (B.40) and (B.41). The right side is measurable with respect to $\mathscr{F}_{t-1}$. The event of a positive entry is measurable with respect to $\mathscr{F}_t$. On this event apply (B.39) conditional on $\mathscr{F}_t = \mathscr{F}_{s-1}$; on its complement the reward is zero. Conditioning back on $\mathscr{F}_{t-1}$ and

using the overshoot bound, $\beta_2 \geq 1/2$, and $1 + x \leq e^x$ therefore gives

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda_H(1-\beta_2)R_t) \mid \mathscr{F}_{t-1}] \\
&\leq \exp\Bigg\{ \frac{2(1-\beta_2)S_{t,i}}{L\Delta_0 K^2} \Bigg[2\big(\partial_i f(\mathbf{x}_t)\big)^2 + 2c_2 k \left\|\nabla f(\mathbf{x}_t)\right\|^2 \\
&\qquad + c_A k(1-\beta_2) S_H \left(c_3 + \frac{17}{2048} + d\log\left(1 + \frac{S_H}{\epsilon^2}\right)\right)\Bigg]\Bigg\}.
\end{aligned} \tag{B.46}$$

Here $S_{t,i}$ is the conditional moment in (B.42), and $\lambda_H = 1/(12LH)$ is defined in Lemma B.13. A draw reaching τ_H or T has zero reward and satisfies the same bound.

For each coordinate, begin a block at its initial low state. A live block starting at t_j ends at t_j+1 if its draw enters no high interval of positive length, and otherwise at the endpoint of that interval. A low scored endpoint is the next live start. Terminate at the horizon or height stop. There are at most T live blocks, with strictly increasing starts. All starts and endpoints are bounded stopping times. After termination, pad subsequent times by T and their rewards and compensators by zero. A terminal reward remains measurable when its endpoint is replaced by T, while its conditional expectation at the preceding live start is unchanged. The block filtration is $(\mathscr{F}_{t_j-1})_{j\geq 0}$, with padded times equal to T.

Let R_j be a block reward and let

$$\begin{aligned}
b_j = \frac{2(1-\beta_2)S_{t_j,i}}{L\Delta_0 K^2} \Bigg[&2\big(\partial_i f(\mathbf{x}_{t_j})\big)^2 + 2c_2 k \left\|\nabla f(\mathbf{x}_{t_j})\right\|^2 \\
&+ c_A k(1-\beta_2) S_H \left(c_3 + \frac{17}{2048} + d\log\left(1 + \frac{S_H}{\epsilon^2}\right)\right)\Bigg]
\end{aligned}$$

at live starts, with $b_j = 0$ on padded blocks. The indicator of a live start and b_j are measurable with respect to $\mathscr{F}_{t_j-1}$, whereas R_j is measurable with respect to $\mathscr{F}_{t_{j+1}-1}$. Because each t_j takes values in $\{0,\dots,T\}$, (B.46) applies at t_j by summing its conditional inequality over the events $\{t_j = t\}$. Consequently

$$\exp\left(\lambda_H(1-\beta_2)\sum_{j=0}^{m-1} R_j - \sum_{j=0}^{m-1} b_j\right)$$

is a nonnegative supermartingale in the block filtration, starting at one. Markov's inequality at block index T, with failure ρ/d, bounds the sum of its rewards by $(\sum_j b_j + \log(d/\rho))/(\lambda_H(1-\beta_2))$. Apply a union bound to these d events.

To sum the compensators, group live starts by their integer time t. Their coordinates form a subset, measurable with respect to $\mathscr{F}_{t-1}$, of $\{1,\dots,d\}$, and the subset is empty unless $I_t = 1$. By (B.42), the charges from coordinate gradients at this time satisfy

$$\sum_{i \text{ live at } t} S_{t,i}\big(\partial_i f(\mathbf{x}_t)\big)^2 \leq I_t S_H \left\|\nabla f(\mathbf{x}_t)\right\|^2.$$

The other two sums satisfy

$$\sum_{i \text{ live at } t} S_{t,i}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq I_t S_H \left\|\nabla f(\mathbf{x}_t)\right\|^2,$$

$$\sum_{i \text{ live at } t} S_{t,i} \leq I_t S_H.$$

These are pointwise inequalities; no independence of coordinates or block starts is needed. Summing over $t < T$ uses G_T from (B.45) and $\sum_{t<T} I_t \leq T$. Finally divide by $\lambda_H(1-\beta_2) = (1-\beta_2)/(12LH)$ with H from (B.12) to obtain (B.44). □

Lemma B.15 (Conversion to ordinary squared gradients)**.** *Under* (B.34)*, for every deterministic gap height* $H \geq 16\Delta_0$ *as in* (B.12)*, integer* $T \geq 1$*, and* $0 < \rho < 1$*, there is an event of probability*

at least $1-\rho$ *on which, with* $K_\star$ *from* (B.2)*,*

$$\frac{1}{T}\sum_{t=0}^{T-1} I_t \left\|\nabla f(\mathbf{x}_t)\right\|^2 \le 128\frac{K_\star}{d}\left(H\sqrt{\frac{L}{\Delta_0}}+\epsilon\right)\frac{W_T}{T} + \frac{72dLH}{(1-\beta_2)T}\log\frac{d}{\rho} + \frac{2^{17}K_\star^4 L^3 H\gamma^2}{d(1-\beta_2)\min\{\epsilon,\sqrt{L\Delta_0}\}^2}\left[1+\log\left(1+\frac{S_H}{\epsilon^2}\right)\right]. \tag{B.47}$$

The constants depend only on $d, \beta_1, \sigma^2/(L\Delta_0), v/(L\Delta_0)$*, and are independent of the stopping height* H *in* (B.12)*.*

Proof. Take

$$K^2 = v + C_K\frac{H}{\Delta_0}S_H, \qquad C_K = 2^{10}, \tag{B.48}$$

and recall that $S_H = 2LH+\sigma^2$ and $H \ge 16\Delta_0$ are the scales in (B.12). Then $K^2 \ge \max\{v, 16S_H\}$. At low states of individual coordinates, the incoming second moment $V_{t-1,i}$ from Algorithm 1 gives the bound on the predictable signal denominator

$$\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)\big(\partial_i f(\mathbf{x}_t)\big)^2}+\epsilon \le \sqrt{K^2+2(1-\beta_2)LH}+\epsilon \le 64\frac{K_\star}{d}\left(H\sqrt{\frac{L}{\Delta_0}}+\epsilon\right).$$

For the last bound, the scale (B.2) gives $K_\star/d \ge 1$, $v/(L\Delta_0) \le K_\star/d$, and $2+\sigma^2/(16L\Delta_0) \le 2K_\star/d$. Hence

$$\frac{\Delta_0[K^2+2(1-\beta_2)LH]}{LH^2} \le \frac{v}{256L\Delta_0} + 2^{10}\left(2+\frac{\sigma^2}{16L\Delta_0}\right)+\frac{1}{16} \le \left(\frac{1}{256}+2^{11}+\frac{1}{16}\right)\frac{K_\star}{d} < 2^{12}\frac{K_\star}{d}.$$

Taking square roots and using $\sqrt{K_\star/d} \le K_\star/d$ proves the denominator bound. By the definition of W_T in (B.15), all low states of individual coordinates contribute at most $64(K_\star/d)(H\sqrt{L/\Delta_0}+\epsilon)W_T$.

For a high coordinate at a time in $[s,e)$, if $\phi_t = LH$, then

$$\big(\partial_i f(\mathbf{x}_t)\big)^2 \le 2LH \le 2\left(\big(\partial_i f(\mathbf{x}_s)\big)^2+\phi_t\right).$$

If $\phi_t < LH$, smoothness and the two auxiliary offsets give

$$\big(\partial_i f(\mathbf{x}_t)\big)^2 \le 3\left(\big(\partial_i f(\mathbf{x}_s)\big)^2+\phi_t\right)+\frac{24K_\star L^2\gamma^2}{1-\beta_2}.$$

There are at most dT scored high states across all coordinates. Applying (B.44) to the stopped gradient sum G_T in (B.45) yields, on its event,

$$G_T \le 64\frac{K_\star}{d}\left(H\sqrt{\frac{L}{\Delta_0}}+\epsilon\right)W_T + \frac{144HS_H(1+c_2k)}{\Delta_0K^2}G_T + \frac{72HS_H}{\Delta_0K^2}c_A k(1-\beta_2)TS_H\left(c_3+\frac{17}{2048}+d\log\left(1+\frac{S_H}{\epsilon^2}\right)\right) + \frac{36dLH}{1-\beta_2}\log\frac{d}{\rho}+\frac{24dK_\star L^2T\gamma^2}{1-\beta_2}.$$

With C_K from (B.48), c_2 from (B.41), and k_0 from (B.34), the coefficient to absorb is at most

$$\frac{144(1+c_2k_0)}{C_K} = \frac{144(1+9/8)}{1024} = \frac{306}{1024} < \frac{1}{2}.$$

The potential estimates hold uniformly for all admissible K, so they apply to (B.48) without any further condition. Absorption multiplies every other term by less than two. After division by T, the coefficients of the preconditioned gradient energy and the confidence term are therefore $128K_\star/d$ and $72d$, respectively.

For the reserve term, the definitions (B.40), (B.41), and (B.2) imply

$$c_3 + \frac{17}{2048} + d\log\left(1+\frac{S_H}{\epsilon^2}\right) \le 257K_\star^2\left[1+\log\left(1+\frac{S_H}{\epsilon^2}\right)\right],$$
$$S_H \le \left(2+\frac{\sigma^2}{16L\Delta_0}\right)LH \le 2\frac{K_\star}{d}LH.$$

Use $HS_H/(\Delta_0 K^2) \le 1/C_K$ and $k(1-\beta_2) = L^2\gamma^2/((1-\beta_2)\min\{\epsilon, \sqrt{L\Delta_0}\}^2)$ from (B.34). The absorbed reserve term divided by T is at most

$$\frac{74016K_\star^4}{d}\frac{L^3H\gamma^2}{(1-\beta_2)\min\{\epsilon,\sqrt{L\Delta_0}\}^2}\left[1+\log\left(1+\frac{S_H}{\epsilon^2}\right)\right],$$

since $144(c_A/C_K)\cdot 257K_\star^2\cdot 2K_\star/d = 74016K_\star^4/d$. The absorbed motion term is at most $48dK_\star L^2\gamma^2/(1-\beta_2)$. Relative to $L^3H\gamma^2/((1-\beta_2)\min\{\epsilon,\sqrt{L\Delta_0}\}^2)$, its coefficient is at most $3dK_\star \le 3K_\star^4/d$, because $H \ge 16\Delta_0$ by (B.12), $\min\{\epsilon,\sqrt{L\Delta_0}\}^2 \le L\Delta_0$, and $K_\star \ge d \ge 1$. Adding these two bounds and using $74019 < 2^{17}$ proves (B.47). □

B.6 A STATIONARITY GUARANTEE WITH PARAMETERS INDEPENDENT OF CONFIDENCE

Proposition B.2 (Guarantee for squared gradients)**.** *Under Assumptions 1 to 3, use $L, f_\star, \sigma, v$ and the initial gap scale Δ_0 from* (B.1)*. Use the coefficients c_{β_2}, c_γ, C from* (B.3)*,* (B.4)*, and* (B.5)*; they depend only on $d, \beta_1, \sigma^2/(L\Delta_0), v/(L\Delta_0)$. For every $\epsilon > 0$ and integer $T \ge 1$, set*

$$\beta_2 = 1 - \frac{c_{\beta_2}}{\sqrt{T}}, \qquad \gamma = \frac{c_\gamma\min\{\epsilon,\sqrt{L\Delta_0}\}}{L\sqrt{T}}. \tag{B.49}$$

For a deterministic initial point, Algorithm 1 initialized with $M_{-1} = 0$ and $V_{-1} = v1$ has $\beta_2 \in (\beta_1^2, 1)$. For every $0 < \delta < 1$, let $\Lambda_\delta = \log(128d/\delta)$ as in (B.1)*. The same tuning satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le CL\Delta_0\frac{\sqrt{L\Delta_0}+\epsilon}{\min\{\epsilon,\sqrt{L\Delta_0}\}}\frac{\Lambda_\delta^4}{\sqrt{T}} \tag{B.50}$$

with probability at least $1-\delta$. Whenever $\Lambda_\delta \le \sqrt{T}$, it also satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le CL\Delta_0\frac{\sqrt{L\Delta_0}+\epsilon}{\min\{\epsilon,\sqrt{L\Delta_0}\}}\frac{\Lambda_\delta^2}{\sqrt{T}} \tag{B.51}$$

with probability at least $1-\delta$.

Proof. Take c_{β_2} and c_γ from (B.3) and (B.4). The local proof gives $c_\gamma \le \sqrt{c_{\beta_2}}$. With $K_\star$ from (B.2), $c_{\beta_2} \ge d/(4K_\star)$ and $c_\gamma = d^5/(2^{24}K_\star^6)$ imply

$$\left(\frac{c_\gamma}{c_{\beta_2}}\right)^2 \le \frac{2^{-44}d^8}{K_\star^{10}} \le 2^{-44} < 2^{-13} = k_0, \qquad \frac{c_\gamma^2}{c_{\beta_2}} \le \frac{2^{-46}d^9}{K_\star^{11}} \le 1.$$

For every $T \ge 1$, (B.49) satisfies the memory range and $\gamma \le \sqrt{(1-\beta_2)\Delta_0/L}$, since $\min\{\epsilon,\sqrt{L\Delta_0}\}^2 \le L\Delta_0$. Also,

$$\frac{L\gamma}{(1-\beta_2)\min\{\epsilon,\sqrt{L\Delta_0}\}} = \frac{c_\gamma}{c_{\beta_2}} \le \sqrt{k_0}.$$

Thus (B.34) holds independently of δ.

Recall $\Lambda_\delta = \log(128d/\delta)$ from (B.1). First suppose that $\Lambda_\delta \le \sqrt{T}$, and choose $H = 16\Delta_0\Lambda_\delta$ as the gap height in (B.12). The event of Proposition B.1 has failure probability at most $3\delta/16$ and

gives $\tau_H > T$ and $W_T \le H/\gamma$, with W_T from (B.15). Apply Lemma B.15 at this same height $H = 16\Delta_0\Lambda_\delta$, with $\rho = \delta/2$. The total failure probability is at most $3\delta/16 + \delta/2 = 11\delta/16 < \delta$. On the intersection, remove I_t using $\tau_H > T$. With the coefficients from (B.3) and (B.4), the parameter identities give

$$\begin{aligned}\left(H\sqrt{\frac{L}{\Delta_0}} + \epsilon\right)\frac{W_T}{T} &\le \frac{H(H\sqrt{L/\Delta_0} + \epsilon)}{\gamma T} \\ &\le \frac{256L\Delta_0}{c_\gamma}\frac{\sqrt{L\Delta_0} + \epsilon}{\min\{\epsilon, \sqrt{L\Delta_0}\}}\frac{\Lambda_\delta^2}{\sqrt{T}}, \\ \frac{L^3H\gamma^2}{(1-\beta_2)\min\{\epsilon, \sqrt{L\Delta_0}\}^2} &= \frac{c_\gamma^2 LH}{c_{\beta_2}\sqrt{T}}, \\ \frac{LH}{(1-\beta_2)T}\log\frac{2d}{\delta} &\le \frac{LH\Lambda_\delta}{c_{\beta_2}\sqrt{T}}.\end{aligned}$$

The choice of H in (B.12) gives $S_H = 2LH + \sigma^2 \le (2 + \sigma^2/(16L\Delta_0))LH$. Since $\epsilon \ge \min\{\epsilon, \sqrt{L\Delta_0}\}$ and

$$\frac{\min\{\epsilon, \sqrt{L\Delta_0}\}}{\sqrt{L\Delta_0}}\log\frac{\sqrt{L\Delta_0}}{\min\{\epsilon, \sqrt{L\Delta_0}\}} \le \frac{1}{e},$$

the reserve logarithm obeys

$$\begin{aligned}&\frac{\min\{\epsilon, \sqrt{L\Delta_0}\}}{\sqrt{L\Delta_0}}\left[1 + \log\left(1 + \frac{S_H}{\epsilon^2}\right)\right] \\ &\le 1 + \log\left(1 + \frac{S_H}{L\Delta_0}\right) + \frac{2}{e} \\ &\le \left(2 + \frac{\sigma^2}{16L\Delta_0} + \frac{1 + 2/e}{16}\right)\frac{H}{\Delta_0}.\end{aligned}$$

Substitution into (B.47), using $K_\star$ from (B.2), bounds the coefficient of

$$L\Delta_0\frac{\sqrt{L\Delta_0} + \epsilon}{\min\{\epsilon, \sqrt{L\Delta_0}\}}\frac{\Lambda_\delta^2}{\sqrt{T}}$$

by

$$\begin{aligned}&\frac{128K_\star}{d}\frac{256}{c_\gamma} + \frac{72d\cdot 16}{c_{\beta_2}} + \frac{2^{17}K_\star^4}{d}\frac{256c_\gamma^2}{c_{\beta_2}}\left(2 + \frac{\sigma^2}{16L\Delta_0} + \frac{1 + 2/e}{16}\right) \\ &\le \frac{2^{39}K_\star^7}{d^6} + 4608K_\star + \frac{3\cdot 2^{-21}d^7}{K_\star^6} < \frac{2^{40}K_\star^9}{d^8}.\end{aligned}$$

The first inequality uses the coefficients in (B.3) and (B.4), and the bound $2 + \sigma^2/(16L\Delta_0) + (1 + 2/e)/16 \le 3K_\star/d$ from (B.2). The last inequality follows from $K_\star \ge d \ge 1$ and

$$\frac{K_\star^7}{d^6} \le \frac{K_\star^9}{d^8}, \qquad K_\star \le \frac{K_\star^9}{d^8}, \qquad \frac{d^7}{K_\star^6} \le \frac{K_\star^9}{d^8}.$$

For the remaining confidence range, the pathwise step bound from Lemma B.4 gives

$$\|\mathbf{x}_t - \mathbf{x}_0\| \le \frac{\sqrt{2K_\star}\gamma t}{\sqrt{1-\beta_2}} \qquad (0 \le t \le T).$$

The bound on the square root of the gap in Lemma B.3 and $f(\mathbf{x}_0) - f_\star \le \Delta_0$ imply, almost surely,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 &\le 2L\max_{0\le t<T}\big(f(\mathbf{x}_t) - f_\star\big) \\ &\le 4L\Delta_0 + \frac{4K_\star L^2\gamma^2T^2}{1-\beta_2} \\ &= 4L\Delta_0 + \frac{4K_\star c_\gamma^2}{c_{\beta_2}}\min\{\epsilon, \sqrt{L\Delta_0}\}^2T^{3/2} \\ &\le 8K_\star L\Delta_0 T^{3/2}.\end{aligned}$$

If $\Lambda_\delta > \sqrt{T}$, this is at most

$$\frac{2^{48}K_\star^9}{d^8} L\Delta_0 \frac{\sqrt{L\Delta_0}+\epsilon}{\min\{\epsilon, \sqrt{L\Delta_0}\}} \frac{\Lambda_\delta^4}{\sqrt{T}}.$$

For $\Lambda_\delta \le \sqrt{T}$, the bound with the square of the logarithm also implies the bound with its fourth power, since $\Lambda_\delta \ge 1$. Both ranges therefore admit the coefficient $2^{48}K_\star^9/d^8$, since $8K_\star \le 2^{48}K_\star^9/d^8$. By (B.2) and (B.5), this coefficient is $C/4$. Thus (B.50) and (B.51) hold with $C/4$, and hence with the stated C. □

B.7 Quantitative conversion to ordinary gradients

Proposition B.3 (Conversion to ordinary gradients)**.** *Under Assumptions 1 to 3, use the parameter rule* (B.6) *and suppose* $\Lambda_\delta \le \sqrt{T}$*. Recall the positive initial gap bound* Δ_0 *and confidence factor* Λ_δ *from* (B.1)*, and the coefficients from* (B.3)*,* (B.4)*, and* (B.5)*. Here* $V_{t-1,i}$ *is coordinate* i *of the incoming second moment state in Algorithm 1. With probability at least* $1-\delta$*, all states* $0 \le t \le T$ *satisfy* $f(\mathbf{x}_t) - f_\star < 16\Delta_0\Lambda_\delta$*, and*

$$\sum_{t=0}^{T-1}\sum_{i=1}^{d} \frac{\big(\partial_i f(\mathbf{x}_t)\big)^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)\big(\partial_i f(\mathbf{x}_t)\big)^2} + \epsilon} \le \frac{16\Delta_0\Lambda_\delta}{\gamma}.$$

On the same event,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le{}& \frac{C(\Lambda_\delta\sqrt{L\Delta_0}+\epsilon)}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{d}\frac{\big(\partial_i f(\mathbf{x}_t)\big)^2}{\sqrt{\beta_2 V_{t-1,i} + (1-\beta_2)\big(\partial_i f(\mathbf{x}_t)\big)^2} + \epsilon} \\ &+ \frac{CL\Delta_0\Lambda_\delta^2}{(1-\beta_2)T} \\ &+ \frac{d^2 C\Lambda_\delta L^2\gamma^2}{(1-\beta_2)\min\{1,\epsilon/\sqrt{L\Delta_0}\}^2}\left[1+\log\left(1+\frac{32L\Delta_0\Lambda_\delta+\sigma^2}{\epsilon^2}\right)\right].\end{aligned} \tag{B.52}$$

The coefficient C *is independent of* ϵ, T, δ*.*

The displayed preconditioned gradient energy uses a predictable denominator; the algorithm uses $\sqrt{V_{t,i}}+\epsilon$ in its update.

Proof. Take $H = 16\Delta_0\Lambda_\delta$ in the stopping definition (B.12) and intersect the events in Proposition B.1 and Lemma B.15, assigning failure probability $\delta/2$ to the latter. Their total failure probability is at most $3\delta/16 + \delta/2 < \delta$. The first event gives, for every $0 \le t \le T$, $f(\mathbf{x}_t) - f_\star < H$ and $W_T \le H/\gamma$, with W_T defined in (B.15). Substituting this height in Lemma B.15 gives the coefficient of the preconditioned gradient energy $128(K_\star/d)(16\Lambda_\delta\sqrt{L\Delta_0}+\epsilon)$ and a confidence term at most $1152dL\Delta_0\Lambda_\delta^2/((1-\beta_2)T)$. For the reserve term, $S_H = 32L\Delta_0\Lambda_\delta + \sigma^2$ by (B.12), and

$$\frac{L^3 H\gamma^2}{(1-\beta_2)\min\{\epsilon,\sqrt{L\Delta_0}\}^2} = \frac{16\Lambda_\delta L^2\gamma^2}{(1-\beta_2)\min\{1,\epsilon/\sqrt{L\Delta_0}\}^2}.$$

The definitions (B.2) and (B.5) give $K_\star \ge d \ge 1$ and

$$2^{11}\frac{K_\star}{d} \le C, \qquad 1152d \le C, \qquad 2^{21}\frac{K_\star^4}{d} \le d^2 C.$$

These inequalities bound the coefficients of the preconditioned gradient energy, confidence term, and reserve, respectively, and prove (B.52) and all assertions of Proposition B.3. □

B.8 Completion of the stationarity and conversion guarantees

Proof of Theorem B.1 and Corollary B.1. Use the initial gap scale Δ_0 and confidence factor Λ_δ from (B.1). The parameter rule (B.6) is the one analyzed in Proposition B.2, with c_{β_2}, c_γ defined in (B.3) and (B.4). Since

$$\frac{\sqrt{L\Delta_0}+\epsilon}{\min\{\epsilon,\sqrt{L\Delta_0}\}} \le 1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon},$$

the estimate for every horizon in Proposition B.2 gives (B.7) with C from (B.5). Its estimate with the square of the logarithm gives the corresponding assertion when $\Lambda_\delta \le \sqrt{T}$. The memory range in Proposition B.1 implies $\beta_1^2 < \beta_2 < 1$.

For Corollary B.1, use $0 < \epsilon \le 1$ and

$$1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon} \le \frac{1 + 1/\sqrt{L\Delta_0} + \sqrt{L\Delta_0}}{\epsilon}.$$

Apply this inequality to the estimates with the fourth and second powers of the logarithm. If $L = \Delta_0 = 1$, the sharper coefficient $C/4$ established in the proof of Proposition B.2 and $(1+\epsilon)/\min\{\epsilon, 1\} \le 2/\epsilon$ give the stated specialization with the same C.

The formal bounds also imply Theorem 1. For fixed $L, \Delta_0 > 0$ and $0 < \epsilon \le 1$, $\min\{\epsilon, \sqrt{L\Delta_0}\}/\epsilon$ lies between $\min\{1, \sqrt{L\Delta_0}\}$ and one. Thus the rule (B.6) has the horizon orders in (3) for each fixed positive offset. In the range $\Lambda_\delta \le \sqrt{T}$, use the bound for conversion to ordinary gradients in Proposition B.3 and

$$\log\left(1 + \frac{32L\Delta_0\Lambda_\delta + \sigma^2}{\epsilon^2}\right) \le \log\max\{1, 32L\Delta_0 + \sigma^2\} + \log\left(1 + \frac{\Lambda_\delta}{\epsilon^2}\right)$$

to obtain Proposition 1, with a constant independent of T, δ.

Every estimate above uses only a positive smoothness upper bound, a positive upper bound on the initial gap, and a conditional variance upper bound. Choosing common values of these bounds for an objective and oracle class therefore gives common parameters and a uniform guarantee. □

B.9 Convergence rate with parameters depending on δ

Proof of Theorem B.2. Use $K_\star$ from (B.2) and the coefficients c_{β_2}, c_γ, C from Equations (B.3) to (B.5). For $d \ge 1$ and $0 < \delta < 1$, the confidence factor in (B.1) satisfies $\Lambda_\delta \ge \log 128 > 1$.

The rule (B.8) and the coefficient (B.3) give $0 < 1 - \beta_2 \le c_{\beta_2} = (1 - \beta_1^2)/4$, hence $\beta_1^2 < \beta_2 < 1$.

First suppose that $\Lambda_\delta \le T$ and use the first branch of (B.8). Lemma B.10, with $H = 16\Delta_0$ in (B.12), gives, with probability at least $1 - 3\delta/16$,

$$\tau_H > T, \qquad I_t = 1 \quad (0 \le t < T), \qquad W_T \le \frac{16\Delta_0}{\gamma}.$$

Here W_T is the preconditioned gradient energy in (B.15), and the second moment scale in (B.12) is $S_H = 32L\Delta_0 + \sigma^2$. The coefficient definitions (B.3) and (B.4), together with (B.2) and $\Lambda_\delta \ge 1$, give

$$\frac{L^2\gamma^2}{(1-\beta_2)^2 \min\{\epsilon, \sqrt{L\Delta_0}\}^2} = \frac{c_\gamma^2}{c_{\beta_2}^2 \Lambda_\delta^2} \le 2^{-44} \frac{d^8}{K_\star^{10}} \le 2^{-13}.$$

Together with the memory and step size bounds established in Lemma B.10, this verifies (B.34). Apply Lemma B.15 at this height with $\rho = \delta/2$. Intersecting its event with the preceding event gives total failure probability at most $3\delta/16 + \delta/2 = 11\delta/16 < \delta$. On their intersection, (B.47) reads

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le 128\frac{K_\star}{d}\left(H\sqrt{\frac{L}{\Delta_0}} + \epsilon\right)\frac{W_T}{T} \\ + \frac{72dLH}{(1-\beta_2)T}\log\frac{2d}{\delta} \\ + \frac{2^{17}K_\star^4 L^3 H\gamma^2}{d(1-\beta_2)\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\left[1 + \log\left(1 + \frac{S_H}{\epsilon^2}\right)\right].\end{aligned}$$

By (B.2), $K_\star/d \ge 1 + \sigma^2/(L\Delta_0)$, and hence $S_H \le 32(K_\star/d)L\Delta_0$. The elementary logarithmic bound gives

$$\log\left(1 + \frac{S_H}{\epsilon^2}\right) \le \frac{\sqrt{S_H}}{\epsilon} \le \sqrt{32K_\star/d}\frac{\sqrt{L\Delta_0}}{\epsilon}.$$

Consequently, since $K_\star/d \geq 1$,

$$1 + \log\left(1 + \frac{S_H}{\epsilon^2}\right) \leq 8\frac{K_\star}{d}\left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right).$$

Also,

$$\frac{\sqrt{L\Delta_0} + \epsilon}{\min\{\epsilon, \sqrt{L\Delta_0}\}} \leq 1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}.$$

Using the bound on W_T, the tuning (B.8), and the value of c_γ in (B.4), the first term satisfies

$$\begin{aligned}
&128\frac{K_\star}{d}\left(H\sqrt{\frac{L}{\Delta_0}} + \epsilon\right)\frac{W_T}{T} \\
&\leq \frac{2^{15}K_\star}{dc_\gamma}L\Delta_0\frac{\sqrt{L\Delta_0} + \epsilon}{\min\{\epsilon, \sqrt{L\Delta_0}\}}\sqrt{\frac{\Lambda_\delta}{T}} \\
&\leq 2^{39}\frac{K_\star^7}{d^6}L\Delta_0\left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right)\sqrt{\frac{\Lambda_\delta}{T}}.
\end{aligned}$$

The definitions in Equations (B.2) and (B.3) give $c_{\beta_2}^{-1} \leq 4K_\star/d$. Together with $\log(2d/\delta) \leq \Lambda_\delta$, this bounds the second term by

$$\begin{aligned}
\frac{72dLH}{(1-\beta_2)T}\log\frac{2d}{\delta} &\leq \frac{1152d}{c_{\beta_2}}L\Delta_0\sqrt{\frac{\Lambda_\delta}{T}} \\
&\leq 4608K_\star L\Delta_0\left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right)\sqrt{\frac{\Lambda_\delta}{T}}.
\end{aligned}$$

For the third term, substitute (B.8), use the logarithmic bound above, and again use the coefficients in Equations (B.3) and (B.4):

$$\begin{aligned}
&\frac{2^{17}K_\star^4L^3H\gamma^2}{d(1-\beta_2)\min\{\epsilon, \sqrt{L\Delta_0}\}^2}\left[1 + \log\left(1 + \frac{S_H}{\epsilon^2}\right)\right] \\
&= \frac{2^{21}K_\star^4c_\gamma^2}{dc_{\beta_2}}L\Delta_0\frac{1 + \log(1 + S_H/\epsilon^2)}{\sqrt{T}\Lambda_\delta^{3/2}} \\
&\leq 2^{-22}\frac{d^7}{K_\star^6}L\Delta_0\left(1 + \frac{\epsilon}{\sqrt{L\Delta_0}} + \frac{\sqrt{L\Delta_0}}{\epsilon}\right)\sqrt{\frac{\Lambda_\delta}{T}}.
\end{aligned}$$

The last inequality also uses $\Lambda_\delta \geq 1$. By (B.5) and $K_\star \geq d \geq 1$,

$$2^{39}\frac{K_\star^7}{d^6} + 4608K_\star + 2^{-22}\frac{d^7}{K_\star^6} \leq 2^{40}\frac{K_\star^9}{d^8} \leq C.$$

Thus (B.9) holds when $\Lambda_\delta \leq T$, with probability at least $1 - 11\delta/16$.

Now suppose that $\Lambda_\delta > T$ and use the second branch of (B.8). The memory condition (B.11) holds because $1 - \beta_2 = c_{\beta_2}$ with c_{β_2} from (B.3). Summing the actual update bound in (B.16), and using $K_\star$ from (B.2), gives on every trajectory and for every $0 \leq t \leq T$,

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_0\| &\leq \frac{2\sqrt{K_\star}\gamma t}{\sqrt{1-\beta_2}} \\
&\leq \frac{2\sqrt{K_\star}c_\gamma}{\sqrt{c_{\beta_2}}}\frac{\min\{\epsilon, \sqrt{L\Delta_0}\}}{L} \\
&\leq 4K_\star c_\gamma\sqrt{\frac{\Delta_0}{L}} \leq \sqrt{\frac{\Delta_0}{L}}.
\end{aligned}$$

Here the final two inequalities follow from Equations (B.3) and (B.4), as $c_{\beta_2}^{-1} \leq 4K_\star$ and $4K_\star c_\gamma \leq 1$. The initial gap bound and Lemma B.3 therefore imply

$$\begin{aligned}
\sqrt{f(\mathbf{x}_t) - f_\star} &\leq \sqrt{\Delta_0} + \sqrt{\frac{L}{2}}\|\mathbf{x}_t - \mathbf{x}_0\| \\
&\leq \sqrt{\Delta_0} + \sqrt{\frac{\Delta_0}{2}} \leq 2\sqrt{\Delta_0}.
\end{aligned}$$

Applying the gradient bound in Lemma B.3 yields, surely,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\sum_{t=0}^{T-1}\big(f(\mathbf{x}_t)-f_\star\big) \le 8L\Delta_0.$$

In this range, $\min\{1,\sqrt{\Lambda_\delta/T}\}=1$. The value of C in (B.5) and $K_\star \ge d \ge 1$ show that this deterministic bound is at most the right side of (B.9). This proves that equation with probability one in the second range and completes the proof for every $T \ge 1$ and $0<\delta<1$. □