THE EXACT TIME-UNIFORM RATE FRONTIER FOR STOCHASTIC GRADIENT DESCENT ON SMOOTH CONVEX OBJECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the time-uniform convergence of the raw iterate of standard stochastic gradient descent (SGD) for unconstrained smooth convex objectives. We prove that, under unbiased norm-sub-Gaussian gradient noise, the time-uniform convergence rate gets arbitrarily close to $\sqrt{\log n/n}$ but never reaches it. More specifically, we prove that for every positive, eventually nondecreasing sequence h satisfying $h(n) = o(\sqrt{n})$, a bound of order $h(n)/\sqrt{n}$, holding simultaneously for all n with probability at least $1-\alpha$ and uniformly over the problem class, is achievable if and only if

$$\sum_{j=1}^{\infty} \frac{1}{h(2^j)^2} < \infty.$$

The constructive sufficiency result follows from a dyadic horizon-free schedule together with an additive conditional-restart inequality. The necessity counterpart applies to every deterministic nonnegative schedule and holds even for a one-dimensional analytic smooth convex objective with Gaussian noise.

1 INTRODUCTION

Stochastic gradient descent (SGD), dating back to at least the stochastic approximation algorithms Robbins & Monro (1951), finds important applications in modern learning (See e.g. Bottou et al., 2018, for a review). In its simplest form, SGD iterates as

$$x_{t+1} = x_t - \eta_t G(x_t, \xi_t), \tag{1}$$

where $G(x_t, \xi_t)$ is an estimate of $\nabla f(x_t)$, and f is the objective function of interest.

Among many types of convergence guarantees, time-uniform guarantees are especially valuable for SGD, because they remain valid under data-dependent stopping. This property is crucial in practice, as stochastic algorithms are often terminated using dynamic, data-dependent rules. From a theoretical perspective, such termination criteria can be formalized as stopping times. Crucially, the stopping time is inherently random, reflecting the stochastic nature of the iterates. A fixed-time guarantee at a predetermined iteration does not directly justify such a stopping rule. To obtain this type of convergence guarantee, we seek to establish bounds of the form:

$$\mathbb{P}\left(f(x_n) - f^* \le b_\alpha(n),\ \forall n \ge 1\right) \ge 1-\alpha, \quad \forall \alpha \in (0,1), \tag{2}$$

where $b_\alpha : \mathbb{N} \to \mathbb{R}_+$ is a boundary depending on α. Unlike a collection of pointwise high-probability bounds, equation 2 is one event over the infinite trajectory and hence remains valid at every almost surely finite data-dependent stopping time. This viewpoint is closely related to confidence sequences in sequential inference (Darling & Robbins, 1967; Lai, 1976; Howard et al., 2021) in statistical inference.

We study this question for standard SGD on unconstrained smooth convex objectives over real Hilbert spaces under conditionally unbiased norm-sub-Gaussian noise. We ask which profiles h permit equation 2 with $b_\alpha(n) = \mathcal{O}(h(n)/\sqrt{n})$ under one deterministic horizon- and confidence-independent schedule, uniformly over the problem class and without dimension dependence.

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.

1.1 CONTRIBUTIONS

Our main result characterizes exactly every achievable overhead profile h.

Theorem 1.1 (Exact profile frontier, informal)**.** *Let $h : \mathbb{N}_+ \to (0, \infty)$ be an eventually nondecreasing profile satisfying $h(n) = o(\sqrt{n})$. There exists a deterministic, confidence-independent, horizon-free stepsize schedule for the SGD recursion equation 1 on smooth convex objectives whose iterates satisfy equation 2 with $b_\alpha(n) = \mathcal{O}(h(n)/\sqrt{n})$, if and only if*

$$\sum_{j=1}^{\infty} \frac{1}{h(2^j)^2} < \infty. \tag{3}$$

In particular, a time-uniform bound of order $\mathcal{O}((\log n)^{(1+\varepsilon)/2}/\sqrt{n})$ is achievable for every $\varepsilon > 0$, whereas $\mathcal{O}\left(\sqrt{\log n/n}\right)$ is not. In addition, the time-uniform bound gets arbitrarily close to the order of $\sqrt{\log n/n}$ through progressively slower iterated-logarithmic overheads, but never reaches it.

The formal version of our main result appears in Theorem 2.3. This criterion is exact at the level of arbitrary slowly varying profiles, not merely a prescribed logarithmic power. Corollary 2.4 establishes that no asymptotically smallest profile h exists. Also, the Cauchy condensation test shows that equation 3 is equivalent to $\sum_{n=1}^{\infty}(nh(n)^2)^{-1} < \infty$. A machine-checked Lean 4 formalization of the targeted mathematical results is provided in the supplementary material.

1.2 RELATED WORKS

For non-smooth stochastic optimization, standard polynomial stepsizes incur logarithmic losses in raw-iterate convergence (Shamir & Zhang, 2013), and Harvey et al. (2019) showed that such losses are unavoidable for the schedules they considered while establishing corresponding high-probability guarantees. Horizon-dependent schedules can recover the optimal rates (Jain et al., 2019), whereas modified methods such as FTRL-based momentum attain the optimal $\mathcal{O}(T^{-1/2})$ expected last-iterate rate without knowing T (Li et al., 2022).

In deterministic non-smooth optimization, Zamani & Glineur (2025) used a moving-comparator argument to derive exact worst-case last-iterate bounds and showed that no universal stepsize sequence attains the optimal guarantee at every horizon. Their moving-comparator argument terminalizes weighted one-step inequalities. Liu & Zhou (2024) extended this mechanism to stochastic composite mirror descent, attaining the optimal fixed-horizon high-probability raw-iterate scale for smooth convex objectives under sub-Gaussian noise. We retain this terminalization mechanism but redesign the remaining-time weights for conditional restarts. Whereas their nonincreasing weights use a nonpositive Bregman remainder to absorb directional martingale variation, ours separate the harmonic and confidence costs, replacing $\mathsf{H}_r(1+\log(1/\delta))$ by $\mathsf{H}_r+\log(1/\delta)$. This additive separation enables simultaneous control over infinitely many epochs and terminal positions.

Recent results provide important time-uniform benchmarks. For smooth convex objectives, Feng et al. (2023) proved an $\mathcal{O}((1 + \log(1/\beta)) \log k/\sqrt{k})$ anytime bound for a particular stochastic gradient descent with momentum (SGDM) scheme; later (Feng et al., 2025) sharpened this to $(\log k)^{(1+\varepsilon)/2}/\sqrt{k}$, $\varepsilon \in (0, 1/2)$, for a variant of SGDM. Under strong convexity or related contractive structure, Chen et al. (2026) and Pham et al. (2025) established sharp $\mathcal{O}((\log\log k + \log(1/\beta))/k)$-type uniform rates. Moreover, Aolaritei & Jordan (2025) derived anytime-valid weighted-suboptimality bounds for projected SGD on general convex domains. These results concern momentum, contractive objectives, or weighted certificates, rather than the raw iterate of standard SGD on the general smooth convex class.

In the non-smooth Lipschitz setting over bounded convex domains, Kornowski & Shamir (2026) proved that no fixed infinite stepsize schedule can achieve an anytime raw-iterate rate $o((\log T)^{1/8}/\sqrt{T})$. To the best of our knowledge, no existing work has characterized the exact time-uniform overhead frontier for the raw last iterate of standard SGD over the general smooth convex class when one deterministic schedule must be independent of both the horizon and the confidence level.

1.3 TECHNICAL OVERVIEW

For sufficiency part, we use dyadic epochs $N_j := 2^j$ with stepsize $\gamma_j \asymp R/(\sigma\sqrt{N_j h(N_j)})$ where σ and R are constants. The stochastic energy of this epoch is $N_j\gamma_j^2 \asymp R^2/(\sigma^2 h(N_j)^2)$, so equation 3 is exactly the condition that the total squared-step energy be finite. The main analytical device is an additive conditional-restart inequality based on remaining-time weights. It controls every terminal position in an epoch while separating the harmonic and confidence terms, after which a summable allocation of failure probabilities yields one event for all n.

To prove the necessity part of Theorem 1.1, we consider a class of one-dimensional analytic quartic-flat convex objectives with Gaussian noise. Such objectives separate the two constraints on the schedule: sufficient cumulative steps are required for deterministic progress, while sufficiently small blockwise squared-step energy is needed to suppress stochastic excursions. Independent Gaussian block events convert these constraints into a deterministic Riccati-type recurrence. If equation 2 held while the series in equation 3 diverged, the recurrence would be forced to remain bounded and diverge simultaneously. Thus the two mechanisms identify the same reciprocal-square frontier.

2 PRELIMINARIES

We use the standard separability convention for real Hilbert spaces. Every space in scope is finite-dimensional or isomorphic to ℓ_2. Accordingly, let $\mathfrak{H} = \{\mathbb{R}^d : d \in \mathbb{N}_+\} \cup \{\ell_2\}$. The problem class is defined as follows.

Definition 2.1. *Fix $L, R, \sigma > 0$. Let $\mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$ denote the class of problem instances $\mathcal{P}$ consisting of the following:*

(i) The ambient space $\mathcal{H} \in \mathfrak{H}$ is selected before the run. We denote the corresponding inner product and norm by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively.

(ii) The objective $f : \mathcal{H} \to \mathbb{R}$ is convex, Fréchet differentiable, and L-smooth, with nonempty minimizer set X^*.

(iii) The initial point $x_0 \in \mathcal{H}$ satisfies $\mathrm{dist}(x_0, X^*) \le R$. Fix $x^* \in X^*$ such that $\|x_0 - x^*\| \le R$, and write $f^* := \min_{x\in\mathcal{H}} f(x) = f(x^*)$.

(iv) Let $\{\xi_t\}_{t\ge 0}$ be fresh, independent random seeds, and let $\mathcal{F}_t := \sigma(\xi_0, \dots, \xi_{t-1})$ be the pre-query filtration. For every $\mathcal{F}_t$-measurable query x_t, the oracle returns a stochastic gradient $G(x_t, \xi_t)$. We require this response to be $\mathcal{F}_{t+1}$-measurable, and define $Z_t := G(x_t, \xi_t) - \nabla f(x_t)$. The stochastic gradient oracle is conditionally unbiased and norm-sub-Gaussian in the following almost-sure sense:

$$\mathbb{E}[Z_t \mid \mathcal{F}_t] = 0, \quad \mathbb{E}\left[\exp\left(\frac{\|Z_t\|^2}{\sigma^2}\right)\middle|\, \mathcal{F}_t\right] \le \mathrm{e}, \quad \forall t \ge 0.$$

We refer to any $\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$ as an admissible problem instance.

For a deterministic infinite schedule $\eta = \{\eta_t\}_{t\ge 0}$ with $\eta_t \ge 0$, consider the standard SGD equation 1 and define

$$\Delta_t := f(x_t) - f^*.$$

The schedule is called horizon-free because the full infinite sequence is determined before the run; it is not retuned for a terminal horizon. More precisely, the schedule is $\eta_t = \eta_t(h, L, R, \sigma)$. Our goal is to establish a time-uniform bound for standard SGD, which is given in the following definition.

Definition 2.2. *Fix $L, R, \sigma > 0$, a confidence level $\alpha \in (0, 1)$, and a deterministic infinite stepsize schedule $\eta = \{\eta_t\}_{t\ge 0}$ with $\eta_t \ge 0$. Let $\mathcal{M}_\eta$ denote the standard SGD method induced by η on the problem class $\mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$, and let $\mathbb{P}^\eta_{\mathcal{P}}$ denote the law of its trajectory on an instance $\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$. A deterministic function $b_{\alpha,\eta} : \mathbb{N}_+ \to \mathbb{R}_+$ is called a time-uniform convergence boundary for $\mathcal{M}_\eta$ over $\mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$ at confidence level $1 - \alpha$ if and only if*

$$\inf_{\mathcal{P}\in\mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)} \mathbb{P}^\eta_{\mathcal{P}}\big(\Delta_n \le b_{\alpha,\eta}(n),\ \forall n \ge 1\big) \ge 1 - \alpha.$$

This probability concerns one joint event over the entire infinite run, and $b_{\alpha,\eta}$ must not depend on the realized sample path.

We now define the class of rate profiles considered in this paper:

$$\mathcal{H}_0 := \{h : \mathbb{N}_+ \to (0,\infty) : h \text{ is eventually nondecreasing and } h(n) = o(\sqrt{n})\}.$$

Here, "eventually nondecreasing" means that there exists $n_0 \in \mathbb{N}_+$ such that $h(n)$ is nondecreasing for all $n \geq n_0$. For any $h \in \mathcal{H}_0$, it acts as a multiplicative overhead on the polynomial rate $\mathcal{O}(1/\sqrt{n})$, which is the optimal fixed-horizon rate for SGD on general smooth convex objectives, since we need $h(n)$ to elevate a fixed-horizon guarantee to a strictly time-uniform guarantee.

With the problem class and operational constraints formally established, we state our main result. The following theorem reveals that the boundary between achievable and unachievable overhead profiles is exactly determined by a reciprocal-square series.

Theorem 2.3 (Exact reciprocal-square profile frontier, formal)**.** *For every fixed $L, R, \sigma > 0$ and $h \in \mathcal{H}_0$, the following statements are equivalent.*

(i) There exists a deterministic infinite schedule $\eta = \eta(h, L, R, \sigma)$, independent of the horizon, problem instance and confidence level, such that for every $\alpha \in (0,1)$, there exists $C_{h,\alpha,L,R,\sigma} < \infty$ satisfying

$$\inf_{\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)} \mathbb{P}_{\mathcal{P}}^{\eta}\left(\Delta_n \leq C_{h,\alpha,L,R,\sigma} \min\left\{1, \frac{h(n)}{\sqrt{n}}\right\},\ \forall n \geq 1\right) \geq 1 - \alpha.$$

(ii)

$$\sum_{j=1}^{\infty} \frac{1}{h(2^j)^2} < \infty.$$

We call profiles satisfying these equivalent conditions achievable, and denote their collection by $\mathcal{H}_{\mathrm{ach}}$

Sections 3 and 4 prove sufficiency ((ii) $\implies$ (i), see Theorem 3.4) and necessity ((i) $\implies$ (ii), see Theorem 4.1). Together, these two directions establish the exact characterization of $\mathcal{H}_{\mathrm{ach}}$. By Theorem 2.3, we have the following corollary, whose proof is deferred to Appendix A.

Corollary 2.4. *The set of achievable profiles $\mathcal{H}_{\mathrm{ach}}$ has no asymptotically smallest element, i.e., if $h \in \mathcal{H}_{\mathrm{ach}}$, then there exists $g \in \mathcal{H}_{\mathrm{ach}}$ such that $g(n) = o(h(n))$.*

To illustrate the corollary, consider the following example. Let $\ell_k(n) = \overbrace{\log \cdots \log}^{k \text{ times}} n$ for all sufficiently large n. Each profile below is understood to be extended to $\mathbb{N}_+$ by an arbitrary positive modification on a finite prefix. Then $\sqrt{\ell_1(n)}$ is not achievable, but $(\ell_1(n))^{1/2+\varepsilon}$ is achievable for any $\varepsilon > 0$. Similarly, $(\ell_1 \cdots \ell_k(n))^{1/2}$ is not achievable, but $(\ell_1 \cdots \ell_{k-1}(n) \cdot \ell_k^{1+\varepsilon}(n))^{1/2}$ is achievable for any $\varepsilon > 0$. Let $h_k(n) = (\ell_1 \cdots \ell_k^{1+\varepsilon}(n))^{1/2}$. Then $h_k \in \mathcal{H}_{\mathrm{ach}}$ and $h_{k+1}(n) = o(h_k(n))$. Hence, there is no asymptotically smallest achievable profile in the family of iterated logarithms, which is consistent with Corollary 2.4. Also, this shows that the time-uniform convergence rate gets arbitrarily close to $\mathcal{O}(\sqrt{\log n/n})$ but never reaches it.

3 Sufficiency of the Reciprocal-Square Profile

In this section, we prove the constructive sufficiency, or upper-bound, direction of Theorem 2.3: every profile $h \in \mathcal{H}_0$ satisfying $\sum_{j\geq 1} h(2^j)^{-2} < \infty$ belongs to $\mathcal{H}_{\mathrm{ach}}$. For arbitrary $L, R, \sigma > 0$, we construct a deterministic infinite schedule $\eta(h, L, R, \sigma)$, independent of the horizon, the problem instance, and the confidence level, whose raw SGD iterates satisfy the $\mathcal{O}(h(n)/\sqrt{n})$ bound simultaneously for all n on a single high-probability event.

3.1 THE HORIZON-FREE DYADIC SCHEDULE

For brevity and clarity, we employ a dyadic epoch-based schedule that maintains a constant stepsize within each epoch to leverage martingale concentration while systematically decreasing the stepsize across epochs to drive optimization error to zero. Let $N_j = 2^j$ and $h_j = h(N_j)$. Algorithm 1 completely specifies our schedule.

Algorithm 1: Horizon-Free Dyadic Schedule for an Admissible Profile

Input: A profile $h \in \mathcal{H}_0$ satisfying $\sum_{j\geq 1} h(2^j)^{-2} < \infty$, and parameters $L, R, \sigma > 0$.
Output: One deterministic infinite schedule η depending only on h, L, R, σ.

$J_h := \inf\{j \in \mathbb{N}_+ : h(n+1) \geq h(n) \text{ for all } n \in \mathbb{N}_+ \text{ with } n \geq N_j\}$.
$c_h := \max\{1, (\sum_{j\geq J_h} h_j^{-2})^{1/2}\}$.
$j_0 := \inf\{j \geq J_h : h_j \geq 1, R/(\sigma c_h \sqrt{N_j} h_j) \leq 1/(2L)\}$.
for $t = 0, 1, 2, \ldots$ **do**
  if $t < N_{j_0}$ **then**
    $\eta_t := 0$.
  end
  if $N_j \leq t < N_{j+1}$ *for some* $j \geq j_0$ **then**
    $\eta_t = \gamma_j := \frac{R}{\sigma c_h \sqrt{N_j} h_j}$.
  end
end
return $\eta = \{\eta_t\}_{t\geq 0}$.

Algorithm 1 is well defined. Indeed, $\sum_{j\geq J_h} h_j^{-2} < \infty$ implies $h_j \to \infty$, and hence $\sqrt{N_j} h_j \to \infty$. Thus the set defining j_0 is nonempty and $j_0 < \infty$. Moreover, since h_j is nondecreasing for $j \geq J_h$, the active stepsizes $\gamma_j = R/(\sigma c_h \sqrt{N_j} h_j)$ are nonincreasing. Crucially, the total squared-step energy of this infinite sequence is bounded above by R^2/σ^2, as guaranteed by the convergence of the reciprocal-square series.

Remark 3.1. *A constant stepsize is adopted within each epoch primarily for simplicity of the sufficiency proof. Alternative stepsize schedules can also be employed.*

3.2 CORE TOOL: ADDITIVE CONDITIONAL RESTART

The sufficiency proof repeatedly applies finite-block last-iterate estimates along the dyadic schedule. Since a block may start from an iterate determined by the preceding trajectory, and since the comparator may be $\mathcal{F}_m$-measurable, we need an estimate that remains valid conditionally on $\mathcal{F}_m$. We refer to such an application as an analytical restart; the SGD trajectory itself is not reset.

A direct specialization of the high-probability last-iterate bound of Liu & Zhou (2024) yields a stochastic term of the form $\sigma^2\gamma \mathsf{H}_r(1 + \log \frac{1}{\delta})$, where $\mathsf{H}_r = \sum_{i=1}^r i^{-1}$ is the r-th harmonic number. Under repeated confidence allocation, this product incurs an additional logarithmic loss. Using the same barycentric last-iterate reduction, but with the remaining weights $w_t = r - t + 1$, we instead obtain the additive dependence $\sigma^2\eta\big(\mathsf{H}_r + \log(1/\delta)\big)$. The following theorem formalizes this conditional last-iterate estimate.

Theorem 3.2. *Fix $L, R, \sigma > 0$, an instance $\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$, and a deterministic nonnegative stepsize schedule η. Let $m \in \mathbb{N}$ and $r \in \mathbb{N}_+$ be deterministic, and suppose that $\eta_m = \cdots = \eta_{m+r-1} = \gamma$, $0 < \gamma \leq 1/(2L)$. Then for every almost surely finite $\mathcal{F}_m$-measurable comparator y and every $\delta \in (0, 1)$,*

$$\mathbb{P}\left(f(x_{m+r}) - f(y) \leq \frac{4D(y, x_m)}{\gamma r} + 6\sigma^2\gamma \mathsf{H}_r + 9\sigma^2\gamma \log \frac{1}{\delta} \middle| \mathcal{F}_m \right) \geq 1 - \delta \quad \textit{almost surely,}$$

where $D : \mathcal{H} \times \mathcal{H} \to [0, +\infty)$ is given by $D(a, b) = \|a - b\|^2/2$.

See Appendix B.1 for a complete proof. Theorem 3.2 supplies the conditional finite-block estimate used in each restart. The time-uniform conclusion will follow by combining it with the global radius control (detailed in Appendix B.2) and the dyadic schedule.

3.3 PROOF OF SUFFICIENCY

Lemma 3.3. *If $\{h_j\}_{j\in\mathbb{N}}$ is positive, eventually nondecreasing, and $\sum_{j=1}^{\infty} h_j^{-2} < \infty$, then $j/h_j^2 \to 0$ as $j \to \infty$.*

Proof. For every sufficiently large j, $\sum_{k=\lfloor j/2 \rfloor}^{j} h_k^{-2} \geq j h_j^{-2}/2$. The left-hand side is a tail of a convergent series and tends to zero. □

Theorem 3.4 (Sufficiency part of Theorem 2.3)**.** *Let $h \in \mathcal{H}_0$ satisfy $\sum_{j=1}^{\infty} h(2^j)^{-2} < \infty$. Then, for every $L, R, \sigma > 0$, there exists a deterministic nonnegative infinite stepsize schedule $\eta = \eta(h, L, R, \sigma)$, independent of the terminal horizon, the problem instance, and the confidence level, such that, for every $\alpha \in (0,1)$, there exists a constant $C_{h,\alpha,L,R,\sigma} < \infty$ satisfying*

$$\inf_{\mathcal{P}\in\mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)} \mathbb{P}_{\mathcal{P}}^{\eta}\left(\Delta_n \leq C_{h,\alpha,L,R,\sigma}\cdot \min\left\{1, \frac{h(n)}{\sqrt{n}}\right\},\ \forall n \geq 1\right) \geq 1-\alpha.$$

The schedule is the one produced by Algorithm 1; it depends on h, L, R, σ alone, and in particular not on α, the horizon, or the instance. The proof has three stages. First, the finite total energy of the schedule confines the entire infinite trajectory to one fixed ball. Epoch j contributes squared-step energy $N_j\gamma_j^2 = \mathcal{O}(h_j^{-2})$, so reciprocal-square summability makes the total energy finite. This yields a single high-probability radius bound for the entire infinite trajectory and makes every active stepsize admissible for Theorem 3.2 (Lemma B.10 and Lemma B.8).

Second, inside each dyadic epoch, two applications of Theorem 3.2 with different comparators control every terminal position. For late terminal positions, enough steps have elapsed that x^* can be used directly as the comparator in Theorem 3.2. But for early terminal positions, since the bound scales like $1/r$ with the number of steps r, the penalty is too large to reach the target bound. Instead, we use the starting point x_{N_j} as the comparator, which eliminates the penalty term. The missing bound on Δ_{N_j} is exactly the late-half bound inherited from epoch $j-1$. Because adjacent dyadic scales differ only by a constant factor, this handoff propagates one constant across all epochs (Lemmas B.12 and B.13).

Finally, uniform control of all 2^j terminal positions in epoch j requires a summable confidence allocation, for which both the harmonic term and $\log(1/\delta)$ are $\mathcal{O}(j)$. The additive dependence in Theorem 3.2 therefore makes the stochastic cost, relative to the target $h_j/\sqrt{N_j}$, of order j/h_j^2, which vanishes by Lemma 3.3. Here the additive structure is crucial: a multiplicative dependence would instead produce j^2/h_j^2 and lose the frontier.

Proof of Theorem 3.4. Fix $\alpha \in (0,1)$ and an arbitrary $\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)$. By Lemma B.10, the schedule of Algorithm 1 satisfies $\gamma_j \leq 1/(2L)$ for every $j \geq j_0$ and $Q_\infty \leq R^2/\sigma^2$. Hence Lemma B.8 applies at level $\alpha/2$, and the global radius event $\mathcal{G}$ of equation 9 in Appendix B.3 satisfies $\mathbb{P}(\mathcal{G}) \geq 1-\alpha/2$. With the allocation $\delta_{j,r} = \alpha/2^{2(j+1)}$ of equation 8 in Appendix B.3, the restart events $\mathcal{A}_{j,r}$ satisfy $\mathbb{P}(\mathcal{A}_{j,r}) \geq 1-\delta_{j,r}$, and

$$\sum_{j=j_0}^{\infty}\sum_{r=1}^{N_j}\delta_{j,r} = \sum_{j=j_0}^{\infty} N_j \frac{\alpha}{2^{2(j+1)}} \leq \frac{\alpha}{2}.$$

A union bound therefore gives $\mathbb{P}(\mathcal{E}) \geq 1-\alpha$ for

$$\mathcal{E} := \mathcal{G} \cap \bigcap_{j\geq j_0}\ \bigcap_{1\leq r\leq N_j} \mathcal{A}_{j,r}.$$

On $\mathcal{E}$, Lemma B.14 controls every $1 \leq t < 3N_{j_0}/2$ with the constant C_{prefix}, and Lemma B.13 controls every $t \geq 3N_{j_0}/2$ with the constant C_{latter}. Finally, L-smoothness at the minimizer and equation 9 give $\Delta_t \leq LB_{\alpha/2} \leq \frac{81}{4}LR^2\log(4/\alpha)$ on $\mathcal{G}$, so the boundary may also be capped at one. Setting

$$C_{h,\alpha,L,R,\sigma} := \max\left\{C_{\mathrm{prefix}}, C_{\mathrm{latter}}, \frac{81}{4}LR^2\log\frac{4}{\alpha}\right\}$$

yields $\Delta_t \leq C_{h,\alpha,L,R,\sigma}\min\{1, h(t)/\sqrt{t}\}$ simultaneously for all $t \geq 1$ on $\mathcal{E}$. Since $\mathcal{P}$ was arbitrary and η does not depend on it, the infimum over $\mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)$ obeys the same bound. □

4 NECESSITY OF RECIPROCAL-SQUARE SUMMABILITY

We prove that reciprocal-square summability is necessary for a deterministic schedule to attain a time-uniform $h(n)/\sqrt{n}$ boundary. It is enough to work with the uncapped envelope $\varepsilon_\alpha(n) \le B \cdot h(n)/\sqrt{n}$, where B is a constant. This is because the above envelope is weaker than the capped boundary in Theorem 2.3.

Theorem 4.1 (Necessity part of Theorem 2.3)**.** *Fix $L, R, \sigma > 0$ and $\alpha \in (0,1)$. Let $h \in \mathcal{H}_0$ and let $\eta = \{\eta_t\}_{t\ge 0}$ be any deterministic nonnegative infinite stepsize schedule, fixed before the run and independent of the terminal horizon and the problem instance. The schedule may depend on h, L, R, σ, and α. Suppose that $\mathcal{M}_\eta$ admits a time-uniform convergence boundary $b_{\alpha,\eta}$ over $\mathfrak{P}_{\mathrm{cvx}}(L, R, \sigma)$ at confidence level $1-\alpha$. If there exist constants $B > 0$ and $n_0 \in \mathbb{N}_+$ such that*

$$b_{\alpha,\eta}(n) \le B \cdot \frac{h(n)}{\sqrt{n}}, \quad \forall n \ge n_0,$$

then $\sum_{j=1}^{\infty} h(2^j)^{-2} < \infty$.

Remark 4.2. *Allowing η to depend on α only strengthens the necessity statement: the conclusion therefore applies directly to the confidence-independent schedules required in the definition of achievability.*

Throughout the proof, the schedule, the profile, and all constants in Theorem 4.1 are fixed. We write $S_n := \sum_{t<n} \eta_t$, $Q_n := \sum_{t<n} \eta_t^2$, and $\varepsilon_\alpha(n) := b_{\alpha,\eta}(n)$. All lemmas and corollaries in the remainder of this section are understood to hold under the conditions of Theorem 4.1.

4.1 FIXED-HORIZON ANALYSIS

Since time-uniform boundaries imply fixed-horizon boundaries:

$$\mathbb{P}\big(\Delta_n \le \varepsilon_\alpha(n)\big) \ge \mathbb{P}\big(\Delta_t \le \varepsilon_\alpha(t),\ \forall t \ge 1\big) \ge 1-\alpha,$$

we can apply the fixed-horizon analysis to obtain some necessary results, which are summarized in the following two lemmas. The first comes from the amount of deterministic progress required to reduce the initial error.

Lemma 4.3. *For any $n \in \mathbb{N}_+$,*

$$\varepsilon_\alpha(n) \ge \frac{R^2}{16} \min\left\{L, \frac{1}{S_n}\right\},$$

where $1/S_n := +\infty$ when $S_n = 0$.

Lemma 4.3 shows that a sufficiently large total stepsize S_n is necessary to reduce the initial error. The second restriction comes from the stochastic energy injected by the oracle noise.

Lemma 4.4. *Suppose $S_n \ge 1/L$. Let $z_\alpha := \Phi^{-1}(1-\alpha/2)$, where Φ is the cumulative distribution function of the standard normal distribution. Then*

$$\varepsilon_\alpha(n) \ge \frac{z_\alpha^2 \sigma^2}{64} \cdot \frac{Q_n}{S_n}.$$

Lemma 4.4 establishes a trade-off constraint: the error bound is governed by the ratio Q_n/S_n. Consequently, accommodating a large Q_n requires a correspondingly large S_n in order to achieve a sufficiently small error bound. The proofs of Lemmas 4.3 and 4.4 are deferred to Appendix C.1.

Now let $N_j = 2^j$, $h_j = h(2^j)$, and $\mathcal{Q}_j = Q_{N_j}$. We define $p_j := S_{N_j}/\sqrt{N_j}$. Then we have the following corollary, whose proof can also be seen in Appendix C.1:

Corollary 4.5. *For sufficiently large $j \in \mathbb{N}_+$, we have*

$$p_j \ge \frac{z_\alpha^2 \sigma^2}{64B} \cdot \frac{\mathcal{Q}_j}{h_j}.$$

These lemmas provide necessary results, but they cannot place infinitely many epoch constraints on one common event. The next subsection uses one fixed instance throughout the infinite run; this is where the joint time-uniform guarantee is essential.

4.2 A FIXED FLAT-ACTIVE WITNESS AND THE DYADIC EPOCH CEILING

Define

$$\phi(y) := \frac{L}{4}\left(\sqrt{R^2+y^2}-R\right)^2$$

It is straightforward to verify that ϕ is convex, L-smooth, and non-PL. We can further verify that ϕ is quartic-flat at the minimizer by Taylor expansion:

$$\phi(y) = \frac{L}{16R^2}y^4 + \mathcal{O}(y^6), \quad \text{as } y \to 0.$$

Moreover,

$$\phi'(y) = \frac{Ly^3}{2\sqrt{R^2+y^2}(\sqrt{R^2+y^2}+R)}, \quad |\phi'(y)| \le \frac{L|y|^3}{4R^2}.$$

Let $\vartheta := (1-\mathrm{e}^{-2})/2$, $\nu^2 := \vartheta\sigma^2$. The recursion is initialized at the minimizer and uses independent $\zeta_t \sim \mathcal{N}(0,\nu^2)$ as the oracle noise:

$$Y_{t+1} = Y_t - \eta_t(\phi'(Y_t)+\zeta_t), \quad Y_0 = 0.$$

The Gaussian oracle satisfies the noise assumption with equality: $\mathbb{E}[\exp(\zeta_t^2/\sigma^2)] = (1-2\nu^2/\sigma^2)^{-1/2} = \mathrm{e}$. Therefore, it is a valid problem instance defined in Definition 2.1. By the assumption of Theorem 4.1, the all-time success event $\mathcal{C} := \{\phi(Y_n) \le \varepsilon_\alpha(n), \forall n \ge 1\}$ satisfies $\mathbb{P}(\mathcal{C}) \ge 1-\alpha$, where $\varepsilon_\alpha(n) \le Bh(n)/\sqrt{n}$ for all sufficiently large n.

Validity on this instance first implies that $\eta_t \to 0$ as $t \to \infty$ (see Lemma C.5 in Appendix C.2).

Now we introduce a few quantities that summarize the behavior of the stepsize schedule and the boundary on each dyadic epoch. First, over the interval $[N_j, N_{j+1}]$, we consider the worst-case rescaling of the overhead function and set

$$\hat{h}_j := \max_{\substack{N_j \le n \le N_{j+1} \\ n \in \mathbb{N}_+}} h(n)\sqrt{\frac{N_j}{n}}.$$

This quantity captures the largest effective value of $h(n)/\sqrt{n}$ within the j-th epoch, expressed at the left endpoint scale N_j. Moreover, there exists a threshold j_h such that, for every $j > i \ge j_h$, eventual monotonicity gives $h_j \le \hat{h}_j \le h_{j+1}$ and $\hat{h}_i \le h_{i+1} \le h_j$. Now we express the corresponding boundary level at epoch j as $\bar{\varepsilon}_j := B\hat{h}_j/\sqrt{N_j}$. On the event $\mathcal{C}$, all iterates in the j-th epoch satisfy $\phi(Y_n) \le \bar{\varepsilon}_j$. This motivates defining

$$\beta_j^2 := \sup\{|y|^2 : \phi(y) \le \bar{\varepsilon}_j\} = r^2(\bar{\varepsilon}_j), \quad r^2(u) := 4R\sqrt{\frac{u}{L}} + \frac{4u}{L}.$$

The following lemma establishes the core time-uniform property of our lower bound.

Lemma 4.6. *Let $\gamma_0 = (\log 2)/4$ and $C_b = 800/(\vartheta\gamma_0)$. Then for every sufficiently large j, we have*

$$H_j := \sum_{t=N_j}^{N_{j+1}-1} \eta_t \le C_b \frac{B}{\sigma^2}\frac{\hat{h}_j\sqrt{N_j}}{j}.$$

The proof is based on a block anti-concentration argument. On the common success event, all iterates in epoch j remain in a shrinking neighborhood of the minimizer, where the quartic-flat objective produces only a small accumulated drift. If the epoch stepsize mass H_j were too large, the epoch could be partitioned into many disjoint blocks, each of which would force an independent Gaussian increment to lie in a prescribed small interval. Gaussian anti-concentration then shows that all these constraints cannot hold with probability at least $1-\alpha$. Balancing the number of available blocks against the corresponding Gaussian tail scale yields the factor $1/j$. The proof is deferred to Appendices C.2 and C.3.

4.3 RECIPROCAL-SQUARE SUMMABILITY

We now combine the fixed-horizon analysis with the dyadic epoch ceiling, yielding the following two lemmas. In Corollary 4.5, we established a lower bound for p_j. The following lemma provides an upper bound for p_j, and together they yield an upper bound for $\mathcal{Q}_j$.

Lemma 4.7. *There exist constants C_p and C_q such that for sufficiently large j, we have*

$$p_j \le C_p \frac{B}{\sigma^2}\frac{h_j}{j}, \quad \mathcal{Q}_j \le C_q \frac{B^2}{\sigma^4}\frac{h_j^2}{j}.$$

The following nonlinear lower recurrence for $\mathcal{Q}_j$ complements the upper bounds above. This recurrence is the key mechanism that connects the dyadic energy growth to the reciprocal-square series.

Lemma 4.8. *Let m be sufficiently large. Then for all $J > m$,*

$$\mathcal{Q}_J \ge (2-\sqrt{2})\mathcal{Q}_m + (\sqrt{2}-1)^2 \frac{z_\alpha^4 \sigma^4}{2^{12} B^2} \sum_{j=m+1}^{J-1} \frac{\mathcal{Q}_j^2}{h_j^2}.$$

The proofs of Lemmas 4.7 and 4.8 are deferred to Appendix C.4. Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. We use contradiction. Suppose $\sum_{j=1}^{\infty} h_j^{-2} = \infty$.

We choose m sufficiently large that all preceding eventual bounds hold. Let $T_{m+1} := (2-\sqrt{2})\mathcal{Q}_m$. By Lemma 4.3, $S_n \to +\infty$. Hence, for sufficiently large m, $\mathcal{Q}_m > 0$, and therefore $T_{m+1} > 0$. By Lemma 4.8, $T_{m+1} \le \mathcal{Q}_{m+1}$. We recursively define

$$T_{J+1} := T_J + \kappa \frac{T_J^2}{h_J^2}, \quad \forall J > m, \quad 0 < \kappa \le (\sqrt{2}-1)^2 \frac{z_\alpha^4 \sigma^4}{2^{12} B^2}.$$

Then by Lemma 4.8 and induction, we have $\mathcal{Q}_J \ge T_J$ for all $J > m$. By Lemma 4.7, we have $\kappa T_J / h_J^2 \le \kappa C_q B^2 \sigma^{-4} / J \le 1$ for every sufficiently large J. Therefore, we have

$$\frac{1}{T_J} - \frac{1}{T_{J+1}} = \frac{T_{J+1} - T_J}{T_J T_{J+1}} = \frac{\kappa T_J^2 / h_J^2}{T_J (T_J + \kappa T_J^2 / h_J^2)} = \frac{\kappa / h_J^2}{1 + \kappa T_J / h_J^2} \ge \frac{\kappa}{2 h_J^2}.$$

Fix an index J_1 from which this inequality holds. Summing from J_1 to ∞ gives

$$\frac{1}{T_{J_1}} \ge \sum_{j=J_1}^{\infty} \frac{1}{T_j} - \frac{1}{T_{j+1}} \ge \sum_{j=J_1}^{\infty} \frac{\kappa}{2 h_j^2} = \infty,$$

which is a contradiction. Therefore, we must have $\sum_{j=1}^{\infty} h_j^{-2} < \infty$. □

5 CONCLUSION

We have characterized the exact time-uniform convergence frontier of standard SGD on general smooth convex objectives: achievable overhead profiles are exactly those satisfying reciprocal-square summability $\sum_{j=1}^{\infty} h(2^j)^{-2} < \infty$. The resulting picture is governed by reciprocal-square summability: rather than admitting a single optimal overhead, the achievable class has no asymptotically smallest element and can approach the $\sqrt{\log n / n}$ arbitrarily closely without attaining it. Our necessity result shows that this obstruction already arises for a one-dimensional analytic smooth convex objective with Gaussian noise, even when the schedule may depend on the confidence level. Thus, the frontier is intrinsic to time-uniform raw-iterate control in the general noncontractive setting, rather than an artifact of nonsmoothness, projection, high dimension, or confidence-independent tuning. In particular, the result characterizes the unavoidable overhead incurred when the raw SGD iterate must remain valid under arbitrary data-dependent stopping. A natural next question is whether the same frontier persists for adaptive or randomized stepsize rules.

AI USE STATEMENT

In this work, we used GPT-5.6 Sol to assist in developing and refining proof strategies for the necessity direction, including suggesting intermediate arguments and technical derivations. The mathematical claims were formulated by the authors, and all AI-assisted proof arguments were independently checked and manually verified by the authors. We did not use generative AI tools for generating the proof in the sufficiency direction. Specifically, GPT-6-Astra was used to assist in translating the mathematical development into Lean 4 and constructing the supplementary formalization. The resulting formalization was checked by the Lean kernel and independently audited for theorem correspondence, dependencies, and unintended axioms. GPT-5.6 Sol and Gemini 3.1 Pro were used to polish the writing of the introduction and related work sections. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs of all mathematical claims are provided in the appendices. We additionally provide an anonymized Lean 4 formalization of the targeted mathematical results as supplementary material, together with theorem-correspondence and proof-integrity audits, as well as instructions for rebuilding the formalization.

REFERENCES

Liviu Aolaritei and Michael I. Jordan. Stopping rules for stochastic gradient descent via anytime-valid confidence sequences, 2025. URL https://arxiv.org/abs/2512.13123.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Kang Chen, Yasong Feng, and Tianyu Wang. Revisiting stochastic gradient descent for strongly convex objectives: Tight uniform-in-time bounds. *Systems & Control Letters*, 212:106419, 2026. doi: 10.1016/j.sysconle.2026.106419. URL https://doi.org/10.1016/j.sysconle.2026.106419.

Donald A Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

Yasong Feng, Yifan Jiang, Tianyu Wang, and Zhiliang Ying. The anytime convergence of stochastic gradient descent with momentum: From a continuous-time perspective, 2023. URL https://arxiv.org/abs/2310.19598.

Yasong Feng, Yifan Jiang, Tianyu Wang, and Zhiliang Ying. Breaking a logarithmic barrier in the stopping time convergence rate of stochastic first-order methods, 2025. URL https://arxiv.org/abs/2506.23335.

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

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1752–1755. PMLR, 2019. URL https://proceedings.mlr.press/v99/jain19a.html.

Guy Kornowski and Ohad Shamir. Gradient descent's last iterate is often (slightly) suboptimal, 2026. URL https://arxiv.org/abs/2604.13870.

Tze Leung Lai. On confidence sequences. *The Annals of Statistics*, pp. 265–280, 1976.

Xiaoyu Li, Mingrui Liu, and Francesco Orabona. On the last iterate convergence of momentum methods. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 699–717. PMLR, 2022. URL https://proceedings.mlr.press/v167/li22a.html.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

Tuan Pham, Alessandro Rinaldo, and Purnamrita Sarkar. Time-uniform concentration bounds for iterative algorithms. *arXiv preprint arXiv:2511.18273*, 2025. URL https://arxiv.org/abs/2511.18273.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 71–79. PMLR, 2013. URL https://proceedings.mlr.press/v28/shamir13.html.

Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025. doi: 10.1137/24M1717762. URL https://doi.org/10.1137/24M1717762.

A Deferred proofs from Section 2

Proof of Corollary 2.4. Let $h \in \mathcal{H}_{\mathrm{ach}}$. By Theorem 2.3, we have

$$\sum_{j=1}^{\infty} \frac{1}{h(2^j)^2} < \infty.$$

Let $h_j = h(2^j)$ and $a_j = h_j^{-2}$. Then $\sum_{j\geq 1} a_j < \infty$. In particular, $a_j \to 0$ and $h_j \to \infty$. Since h is eventually nondecreasing, there exists $J_* \geq 1$ such that h is nondecreasing on $[2^{J_*}, \infty)$. Also, by the convergence of the series, we can choose a strictly increasing sequence of integers $J_* \leq J_1 < J_2 < \cdots$ such that for every $k \geq 1$, $\sum_{j=J_k}^{\infty} a_j < 2^{-k}k^{-2}$. Now we define a positive sequence $\{c_k\}_{k\geq 1}$ by

$$c_j := \begin{cases} 1, & 1 \leq j < J_1 \\ k, & J_k \leq j < J_{k+1}, \quad k \geq 1 \end{cases}$$

Then $c_j \to \infty$. Moreover,

$$\sum_{j=1}^{\infty} c_j a_j = \sum_{j=1}^{J_1-1} a_j + \sum_{k=1}^{\infty} k \sum_{j=J_k}^{J_{k+1}-1} a_j \leq \sum_{j=1}^{J_1-1} a_j + \sum_{k=1}^{\infty} k \sum_{j=J_k}^{\infty} a_j < \sum_{j=1}^{J_1-1} a_j + \sum_{k=1}^{\infty} \frac{2^{-k}}{k} < \infty.$$

Now define $r_j := 1/\sqrt{c_j a_j} = h_j/\sqrt{c_j}$ and $g_j := \max_{1\leq i\leq j} r_i$ for $j \geq 1$. Then $\{g_j\}_{j\geq 1}$ is positive and nondecreasing. Since $g_j \geq r_j$, $g_j^{-2} \leq r_j^{-2} = c_j a_j$. Therefore,

$$\sum_{j=1}^{\infty} \frac{1}{g_j^2} \leq \sum_{j=1}^{\infty} c_j a_j < \infty.$$

We next prove $g_j = o(h_j)$. For any $\varepsilon > 0$, there exists an integer $K \in \mathbb{N}_+$ such that $K^{-1/2} < \varepsilon$. For every $i \geq J_K$, we have $c_i \geq K$. Since $J_K \geq J_*$, for every $J_K \leq i \leq j$ we have $h_i \leq h_j$. Hence,

$$r_i = \frac{h_i}{\sqrt{c_i}} \leq \frac{h_j}{\sqrt{K}} < \varepsilon h_j.$$

Now for the finitely many indices $i < J_K$, define $M_K := \max_{1\le i<J_K} r_i$ with the convention $M_K := 0$ if $J_K = 1$. Since $h_j \to \infty$, there exists $J' \ge J_K$ such that $M_K < \varepsilon h_j$ for any $j \ge J'$. Thus, for every $j \ge J'$,

$$g_j = \max_{1\le i\le j} r_i \le \varepsilon h_j,$$

which implies $g_j/h_j \to 0$ as $j \to \infty$. Finally, we define $g : \mathbb{N}_+ \to (0,\infty)$ by

$$g(n) := g_{\max\{1,\lfloor \log_2 n\rfloor\}}.$$

Now we only need to prove $g(n) = o(h(n))$. For sufficiently large n, let $j = \lfloor \log_2 n\rfloor$. Then $2^j \le n < 2^{j+1}$, and eventual monotonicity of h yields $h(n) \ge h(2^j) = h_j$. Therefore,

$$0 \le \frac{g(n)}{h(n)} = \frac{g_j}{h(n)} \le \frac{g_j}{h_j} \to 0.$$

Hence $g(n) = o(h(n))$. Since $h(n) = o(\sqrt{n})$, it follows that $g(n) = o(\sqrt{n})$ and therefore $g \in \mathcal{H}_0$. Moreover,

$$\sum_{j=1}^{\infty} \frac{1}{g(2^j)^2} = \sum_{j=1}^{\infty} \frac{1}{g_j^2} < \infty.$$

By Theorem 2.3, $g \in \mathcal{H}_{\mathrm{ach}}$. Since $g(n) = o(h(n))$, every $h \in \mathcal{H}_{\mathrm{ach}}$ admits another $g \in \mathcal{H}_{\mathrm{ach}}$ that is asymptotically strictly smaller. Therefore, $\mathcal{H}_{\mathrm{ach}}$ has no asymptotically smallest element. □

B Proofs and auxiliary results for sufficiency

B.1 Additive conditional restart theorem: proof and supporting lemmas

Let $y \in \mathcal{H}$ be an arbitrary comparator. We define the squared-distance function $D : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+ \cup \{0\}$ by $D(a,b) = \|a-b\|^2/2$.

B.1.1 Local smooth geometry

Our analysis is based on two geometric lemmas.

Lemma B.1. *For any $x, u, z \in \mathcal{H}$,*

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

Proof. By the L-smoothness of f, we have

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

By the convexity of f, we have

$$f(x) - f(z) \le \langle \nabla f(x), x - z\rangle.$$

Adding these two inequalities yields

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

□

Lemma B.2. *Let $x^+ = x - \eta(\nabla f(x) + \zeta)$ with $0 < \eta \le 1/(2L)$. Then for any $z \in \mathcal{H}$,*

$$f(x^+) - f(z) \le \frac{D(z,x) - D(z,x^+)}{\eta} + \langle \zeta, z - x\rangle + \eta\|\zeta\|^2.$$

Proof. Applying Lemma B.1 with $u = x^+$ gives

$$f(x^+) - f(z) \le \langle \nabla f(x), x^+ - z\rangle + \frac{L\eta^2}{2}\|\nabla f(x) + \zeta\|^2$$
$$= \frac{\langle \eta(\nabla f(x) + \zeta), x^+ - z\rangle}{\eta} - \langle \zeta, x^+ - z\rangle + \frac{L\eta^2}{2}\|\nabla f(x) + \zeta\|^2$$

By the identity

$$\|a + b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2,$$

we have

$$\langle \eta(\nabla f(x) + \zeta), x^+ - z\rangle = -\frac{1}{2}\eta^2\|\nabla f(x) + \zeta\|^2 - \frac{1}{2}\|x^+ - z\|^2 + \frac{1}{2}\|x - z\|^2.$$

Hence, we have

$$f(x^+) - f(z) \le \frac{D(x, z) - D(x^+, z)}{\eta} - \langle \zeta, x^+ - z\rangle + \frac{\eta}{2}(L\eta - 1)\|\nabla f(x) + \zeta\|^2$$

Notice that

$$\eta\langle \zeta, \nabla f(x)\rangle - \frac{\eta}{4}\|\nabla f(x) + \zeta\|^2 = -\frac{\eta}{4}\|\nabla f(x) - \zeta\|^2 \le 0.$$

Therefore, we have

$$f(x^+) - f(z) \le \frac{D(x, z) - D(x^+, z)}{\eta} + \langle \zeta, z - x\rangle + \eta\|\zeta\|^2.$$

□

Let $\rho \in [0, 1]$ and set $z = (1 - \rho)x + \rho q$. Since $D(z, x) = \rho^2 D(q, x) \le \rho D(q, x)$, Lemma B.2 gives

$$f(x^+) - f(z) \le \rho\langle \zeta, q - x\rangle + \frac{1}{\eta}(\rho D(q, x) - D(z, x^+)) + \eta\|\zeta\|^2.$$

B.1.2 Terminal weights and barycentric identities

In the following, consider the iteration

$$X^{t+1} = X^t - \eta_t(\nabla f(X^t) + Z^t), \quad 1 \le t \le T, 0 < \eta_t \le \frac{1}{2L}.$$

for some $T \ge 1$.

Our objective in this section is to bound the global error $f(X^{T+1}) - f(y)$ evaluated at an arbitrary comparator $y \in \mathcal{H}$. Since the one-step inequality in Lemma B.2 bounds the single-step error $f(X^{t+1}) - f(z^t)$ relative to a reference point z^t, we need to construct a linear combination of these single-step errors that telescopes exactly to the global error.

Let $a_t > 0$ denote the linear-combination coefficient at time t. We consider the corresponding weighted sum of single-step errors. To relate $f(z^t)$ to the iterates $f(X^t)$, we define z^t via a barycentric recursion. Let $\{v_t\}_{t=1}^T$ be a positive, nondecreasing sequence of parameters with $0 < v_0 = v_1$ and $v_T = 1$, and define

$$z^t := \left(1 - \frac{v_{t-1}}{v_t}\right)X^t + \frac{v_{t-1}}{v_t}z^{t-1}, \quad z^0 := y.$$

By applying convexity to this recursion, we have

$$v_t f(z^t) \le (v_t - v_{t-1})f(X^t) + v_{t-1}f(z^{t-1}) \le v_0 f(y) + \sum_{s=1}^{t}(v_s - v_{s-1})f(X^s).$$

Substituting this inequality into the linear combination of single-step errors, we have

$$\begin{aligned}\sum_{t=1}^{T} a_t(f(X^{t+1}) - f(z^t)) &\geq \sum_{t=1}^{T} a_t f(X^{t+1}) - \sum_{t=1}^{T} \frac{a_t}{v_t}\left(v_0 f(y) + \sum_{s=1}^{t}(v_s - v_{s-1}) f(X^s)\right) \\ &= \sum_{t=2}^{T}\left(a_{t-1} - (v_t - v_{t-1})\sum_{k=t}^{T}\frac{a_k}{v_k}\right) f(X^t) \\ &\quad - v_0 f(y)\sum_{t=1}^{T}\frac{a_t}{v_t} + a_T f(X^{T+1}).\end{aligned}$$

Let $a_{t-1} = (v_t - v_{t-1})\sum_{k=t}^{T}\frac{a_k}{v_k}$. Then the first term vanishes. Let $b_t = \frac{a_t}{v_t}$ and $W_t = \sum_{k=t}^{T} b_k$. Then we have $a_{t-1} = (v_t - v_{t-1})W_t$. On the other hand, $a_{t-1} = b_{t-1}v_{t-1}$. Hence, we have the following identity:

$$b_{t-1}v_{t-1} = (v_t - v_{t-1})W_t, \quad 2 \leq t \leq T.$$

Also note that $b_{t-1} = W_{t-1} - W_t$. Substituting this into the previous identity, we have

$$v_{t-1}W_{t-1} = v_t W_t \implies v_t W_t = v_T W_T = a_T, \quad 2 \leq t \leq T.$$

Since $v_0 = v_1$, we have $v_0 W_1 = v_1 W_1 = a_T$. Substituting this back into the previous inequality, we have

$$\sum_{t=1}^{T} a_t(f(X^{t+1}) - f(z^t)) \geq a_T(f(X^{T+1}) - f(y)).$$

Now let $b_t = w_t\eta_t$, where w_t is a nonincreasing, nonnegative weight sequence. Then $a_t = w_t\eta_t v_t$. Let $a_T = C$. Then we have the following lemma.

Lemma B.3. *The following inequality holds:*

$$C(f(X^{T+1}) - f(y)) \leq w_1 v_0 D(y, X^1) + \sum_{t=1}^{T} c_t\|Z^t\|^2 + \sum_{t=1}^{T}\beta_t\langle Z^t, d_t\rangle + \mathcal{V}_T,$$

where $c_t = w_t\eta_t^2 v_t$, $\beta_t = w_t\eta_t v_{t-1}$, $d_t = z^{t-1} - X^t$, *and*

$$\mathcal{V}_T = \sum_{t=2}^{T}(w_t - w_{t-1})v_{t-1}D(z^{t-1}, X^t) \leq 0.$$

Proof. Applying Lemma B.2 to each term in the summation $\sum_{t=1}^{T} a_t(f(X^{t+1}) - f(z^t))$, we have

$$\begin{aligned}&\sum_{t=1}^{T} a_t\big(f(X^{t+1}) - f(z^t)\big) \\ &\leq \sum_{t=1}^{T} w_t v_{t-1} D(z^{t-1}, X^t) - \sum_{t=1}^{T} w_t v_t D(z^t, X^{t+1}) + \sum_{t=1}^{T} c_t\|Z^t\|^2 + \sum_{t=1}^{T}\beta_t\langle Z^t, d_t\rangle \\ &= w_1 v_0 D(y, X^1) + \sum_{t=2}^{T}(w_t - w_{t-1})v_{t-1}D(z^{t-1}, X^t) \\ &\quad - w_T v_T D(z^T, X^{T+1}) + \sum_{t=1}^{T} c_t\|Z^t\|^2 + \sum_{t=1}^{T}\beta_t\langle Z^t, d_t\rangle \\ &\leq w_1 v_0 D(y, X^1) + \sum_{t=1}^{T} c_t\|Z^t\|^2 + \sum_{t=1}^{T}\beta_t\langle Z^t, d_t\rangle + V_T.\end{aligned}$$

where the last inequality follows from the fact that $-w_T v_T D(z^T, X^{T+1}) \leq 0$ and

$$\mathcal{V}_T = \sum_{t=2}^{T}(w_t - w_{t-1})v_{t-1}D(z^{t-1}, X^t).$$

Since w_t is nonincreasing, $\mathcal{V}_T \leq 0$. □

B.1.3 PROOF OF THE ADDITIVE CONDITIONAL RESTART THEOREM

Let $w_t = T - t + 1$. Consider the constant stepsize $\eta_t = \gamma$ for $1 \le t \le T$. Then $b_t = w_t\eta_t = \gamma w_t$ and $W_t = \sum_{k=t}^{T} b_k = \gamma w_t(w_t+1)/2$. We also have $C = a_T = w_T\gamma = \gamma$. Then by the identity $v_t W_t = C$, we have

$$v_t = \frac{C}{W_t} = \frac{2}{w_t(w_t+1)}.$$

Substituting these expressions into Lemma B.3 gives

$$f(X^{T+1}) - f(y) \le \frac{2D(y, X^1)}{\gamma(T+1)} + \sum_{t=1}^{T} w_t\gamma v_t\|Z^t\|^2 + \sum_{t=1}^{T} w_t v_{t-1}\langle Z^t, d_t\rangle - \sum_{t=2}^{T}\frac{v_{t-1}}{2\gamma}\|d_t\|^2.$$

To bound the error, we need the following two lemmas.

Lemma B.4. *For any $\delta \in (0,1)$, the following inequality holds:*

$$\mathbb{P}\left(\sum_{t=1}^{T} w_t\gamma v_t\|Z^t\|^2 \le 2\sigma^2\gamma\mathsf{H}_T + \sigma^2\gamma\log\frac{1}{\delta}\right) \ge 1-\delta,$$

where $\mathsf{H}_T = \sum_{i=1}^{T}\frac{1}{i}$.

Proof. Let $a_* = \max_{1\le i\le T}\{w_i\gamma v_i\}$. Let $\{\mathcal{F}_t\}_{t=1}^{T}$ be a filtration such that Z^t is measurable with respect to $\mathcal{F}_t$. Let

$$N_t = \exp\left(\sum_{i=1}^{t}\frac{a_i}{a_*}\left(\frac{\|Z^i\|^2}{\sigma^2} - 1\right)\right), \quad a_i = w_i\gamma v_i.$$

By the sub-Gaussian property of Z^t, we have

$$\mathbb{E}\left[\exp\left(\frac{\|Z^t\|^2}{\sigma^2} - 1\right)\middle|\mathcal{F}_{t-1}\right] \le 1.$$

Since $a_t/a_* \le 1$, x^{a_t/a_*} is concave for $x \ge 0$. By Jensen's inequality, we have

$$\mathbb{E}\left[\exp\left(\frac{a_t}{a_*}\left(\frac{\|Z^t\|^2}{\sigma^2} - 1\right)\right)\middle|\mathcal{F}_{t-1}\right] \le \left\{\mathbb{E}\left[\exp\left(\frac{\|Z^t\|^2}{\sigma^2} - 1\right)\middle|\mathcal{F}_{t-1}\right]\right\}^{a_t/a_*} \le 1.$$

Since

$$\frac{N_{t+1}}{N_t} = \exp\left(\frac{a_{t+1}}{a_*}\left(\frac{\|Z^{t+1}\|^2}{\sigma^2} - 1\right)\right),$$

we have

$$\mathbb{E}(N_{t+1}/N_t \mid \mathcal{F}_t) \le 1 \implies \mathbb{E}(N_{t+1} \mid \mathcal{F}_t) \le N_t.$$

Therefore, $\{N_t\}$ is a supermartingale. By Ville's inequality, we have

$$\mathbb{P}(N_T \le 1/\delta) \ge 1-\delta, \quad \forall\delta \in (0,1).$$

The event $N_T \le 1/\delta$ implies

$$\sum_{t=1}^{T} w_t\gamma v_t\|Z^t\|^2 \le 2\sigma^2\gamma\mathsf{H}_T + \sigma^2\gamma\log\frac{1}{\delta},$$

where $\mathsf{H}_T = \sum_{i=1}^{T}\frac{1}{i}$. Therefore, we have

$$\mathbb{P}\left(\sum_{t=1}^{T} w_t\gamma v_t\|Z^t\|^2 \le 2\sigma^2\gamma\mathsf{H}_T + \sigma^2\gamma\log\frac{1}{\delta}\right) \ge 1-\delta,$$

□

Lemma B.5. *For any $\delta \in (0,1)$, the following inequality holds:*

$$\mathbb{P}\left(\sum_{t=1}^{T} w_t v_{t-1}\langle Z^t, d_t\rangle - \sum_{t=2}^{T}\frac{v_{t-1}}{2\gamma}\|d_t\|^2 \le \frac{2D(y,X^1)}{\gamma(T+1)^2} + 8\sigma^2\gamma\log\frac{1}{\delta}\right) \ge 1-\delta.$$

Proof. Let $\mathcal{F}_t$ be a filtration generated by $Z^1, \dots, Z^t$. The sub-Gaussian property of Z^t implies that for any $\lambda > 0$,

$$\mathbb{E}[\exp(\langle Z^t, \lambda w_t v_{t-1} d_t \rangle) \mid \mathcal{F}_{t-1}] \le \exp(2\sigma^2 \lambda^2 w_t^2 v_{t-1}^2 \|d_t\|^2).$$

Hence, we can construct a supermartingale M_t as follows:

$$M_t = \exp\left(\lambda \sum_{i=1}^{t} w_i v_{i-1} \langle Z^i, d_i \rangle - 2\sigma^2 \lambda^2 \sum_{i=1}^{t} w_i^2 v_{i-1}^2 \|d_i\|^2 \right)$$

It is straightforward to verify that M_t is a supermartingale. By Ville's inequality, we have

$$\mathbb{P}(M_T \le 1/\delta) \ge 1 - \delta, \quad \forall \delta \in (0, 1).$$

The event $M_T \le 1/\delta$ implies

$$\sum_{t=1}^{T} w_t v_{t-1} \langle Z^t, d_t \rangle \le 2\sigma^2 \lambda \sum_{t=1}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2 + \frac{\log(1/\delta)}{\lambda}.$$

Note that

$$\begin{aligned}
2\sigma^2 \lambda \sum_{t=1}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2 &= 2\sigma^2 \lambda w_1^2 v_0^2 \|d_1\|^2 + 2\sigma^2 \lambda \sum_{t=2}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2 \\
&= 2\sigma^2 \lambda w_1^2 v_1^2 \|d_1\|^2 + 2\sigma^2 \lambda \sum_{t=2}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2 \\
&= \frac{8\sigma^2 \lambda}{(T+1)^2} \|d_1\|^2 + 2\sigma^2 \lambda \sum_{t=2}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2.
\end{aligned}$$

Observe that

$$\frac{2\gamma(w_t^2 v_{t-1}^2)}{v_{t-1}} = \frac{4\gamma w_t^2}{(w_t + 1)(w_t + 2)} \le 4\gamma.$$

Hence, we have

$$2\sigma^2 \lambda \sum_{t=2}^{T} w_t^2 v_{t-1}^2 \|d_t\|^2 \le 8\sigma^2 \lambda \gamma \sum_{t=2}^{T} \frac{v_{t-1}}{2\gamma} \|d_t\|^2.$$

Taking $\lambda = 1/(8\sigma^2\gamma)$, we have

$$\sum_{t=1}^{T} w_t v_{t-1} \langle Z^t, d_t \rangle - \sum_{t=2}^{T} \frac{v_{t-1}}{2\gamma} \|d_t\|^2 \le \frac{2D(y, X^1)}{\gamma(T+1)^2} + 8\sigma^2 \gamma \log \frac{1}{\delta}.$$

This holds with probability at least $1 - \delta$, or equivalently,

$$\mathbb{P}\left(\sum_{t=1}^{T} w_t v_{t-1} \langle Z^t, d_t \rangle - \sum_{t=2}^{T} \frac{v_{t-1}}{2\gamma} \|d_t\|^2 \le \frac{2D(y, X^1)}{\gamma(T+1)^2} + 8\sigma^2 \gamma \log \frac{1}{\delta} \right) \ge 1 - \delta.$$

□

With the master inequality and the two martingale bounds established, we are ready to prove Theorem 3.2. We first prove the following lemma, which is a direct consequence of Lemmas B.3, B.4, and B.5.

Lemma B.6. *The following inequality holds with probability at least* $1 - \delta$*:*

$$f(X^{T+1}) - f(y) \le \frac{4D(y, X^1)}{\gamma T} + 6\sigma^2 \gamma \mathsf{H}_T + 9\sigma^2 \gamma \log \frac{1}{\delta}.$$

Proof. We apply Lemma B.4 with failure probability $p\delta$ and Lemma B.5 with failure probability $(1-p)\delta$ for some $p \in (0,1)$. By the union bound, with probability at least $1-\delta$,

$$\begin{aligned} f(X^{T+1}) - f(y) &\le \frac{2D(y,X^1)}{\gamma(T+1)}\left(1+\frac{1}{T+1}\right) + 2\sigma^2\gamma \mathsf{H}_T + 9\sigma^2\gamma\log\frac{1}{\delta} - \sigma^2\gamma(\log p + 8\log(1-p)) \\ &\le \frac{4D(y,X^1)}{\gamma T} + 2\sigma^2\gamma\mathsf{H}_T + 9\sigma^2\gamma\log\frac{1}{\delta} + \sigma^2\gamma\left(\log 9 + 8\log\frac{9}{8}\right) \\ &\le \frac{4D(y,X^1)}{\gamma T} + 6\sigma^2\gamma\mathsf{H}_T + 9\sigma^2\gamma\log\frac{1}{\delta}, \end{aligned}$$

In the second inequality, we take $p = 1/9$ to minimize the term $-\log p - 8\log(1-p)$. Its minimum has absolute value at most 4, so the last inequality follows. Therefore,

$$\mathbb{P}\left(f(X^{T+1}) - f(y) \le \frac{4D(y,X^1)}{\gamma T} + 6\sigma^2\gamma\mathsf{H}_T + 9\sigma^2\gamma\log\frac{1}{\delta}\right) \ge 1-\delta.$$

□

For clarity, we restate Theorem 3.2.

Theorem B.7. *Fix $L, R, \sigma > 0$, an instance $\mathcal{P} \in \mathfrak{P}_{\mathrm{cvx}}(L,R,\sigma)$, and a deterministic nonnegative stepsize schedule η. Let $m \in \mathbb{N}$ and $r \in \mathbb{N}_+$ be deterministic, and suppose that $\eta_m = \cdots = \eta_{m+r-1} = \gamma$, $0 < \gamma \le 1/(2L)$. Then for every almost surely finite $\mathcal{F}_m$-measurable comparator y and every $\delta \in (0,1)$,*

$$\mathbb{P}\left(f(x_{m+r}) - f(y) \le \frac{4D(y,x_m)}{\gamma r} + 6\sigma^2\gamma\mathsf{H}_r + 9\sigma^2\gamma\log\frac{1}{\delta}\middle|\mathcal{F}_m\right) \ge 1-\delta \quad \textit{almost surely.}$$

Proof. Condition on $\mathcal{F}_m$, set $X^t := x_{m+t-1}$, $Z^t := Z_{m+t-1}$, and $T := r$, and use the local filtration $\mathcal{G}_t := \mathcal{F}_{m+t}$ for $0 \le t \le r$. Then X^1 and y are $\mathcal{G}_0$-measurable, while Z^t is $\mathcal{G}_t$-measurable and satisfies the conditional noise assumptions given $\mathcal{G}_{t-1}$. Lemma B.6 therefore applies conditionally and gives the stated bound. □

B.2 Global radius and calibration

For a time-uniform analysis based on finite stepsize energy, we must first control the entire trajectory. The next lemma obtains this control from two nonnegative supermartingales: one for the accumulated noise energy and one for the stopped directional noise.

Define $D_t := \|x_t - x^*\|^2/2$, $Q_\infty := \sum_{t=0}^\infty \eta_t^2$, and $\eta_* := \sup_{t\ge 0}\eta_t$. We assume $D_0 \le R^2/2$, $Q_\infty < \infty$, and $0 \le \eta_t \le 1/(2L)$.

Lemma B.8. *Let $\beta \in (0,1)$ and $\ell_\beta := \log(2/\beta)$. Define*

$$C_\beta := \frac{R^2}{2} + \frac{\sigma^2 Q_\infty}{2} + \frac{\sigma^2\eta_*^2\ell_\beta}{2}, \quad q_\beta := \sigma^2 Q_\infty \ell_\beta,$$

and set

$$B_\beta := \left(\sqrt{C_\beta + 4q_\beta} + 2\sqrt{q_\beta}\right)^2.$$

Then

$$\mathbb{P}\left(D_t \le B_\beta,\ \forall t \ge 0\right) \ge 1-\beta.$$

Proof. We divide the proof into four steps.

Step 1: A pathwise distance recursion. Set $\bar{x}_{t+1} := x_t - \eta_t\nabla f(x_t)$. Since f is convex and L-smooth and $\nabla f(x^*) = 0$, cocoercivity gives

$$\langle \nabla f(x_t), x_t - x^*\rangle \ge \frac{1}{L}\|\nabla f(x_t)\|^2.$$

Therefore, for every $0 \le \eta_t \le 2/L$,

$$\begin{aligned}\|\bar{x}_{t+1} - x^*\|^2 &= \|x_t - x^*\|^2 - 2\eta_t\langle \nabla f(x_t), x_t - x^*\rangle + \eta_t^2\|\nabla f(x_t)\|^2 \\ &\le \|x_t - x^*\|^2 - \eta_t\left(\frac{2}{L} - \eta_t\right)\|\nabla f(x_t)\|^2 \\ &\le \|x_t - x^*\|^2.\end{aligned}$$

Thus the deterministic gradient step is nonexpansive relative to x^*. Since $x_{t+1} = \bar{x}_{t+1} - \eta_t Z_t$, we obtain the pathwise inequality

$$D_{t+1} \le D_t + \frac{\eta_t^2}{2}\|Z_t\|^2 - \eta_t\langle Z_t, \bar{x}_{t+1} - x^*\rangle. \tag{4}$$

Step 2: Time-uniform control of the accumulated noise energy. If $\eta_* = 0$, then every step is zero and the result is immediate. Hence, assume $\eta_* > 0$ and put $c_t := \eta_t^2/\eta_*^2 \in [0,1]$. The conditional noise assumption and conditional Jensen's inequality give

$$\mathbb{E}\left[\exp\left(c_t\left(\frac{\|Z_t\|^2}{\sigma^2} - 1\right)\right)\middle|\,\mathcal{F}_t\right] \le \left\{\mathbb{E}\left[\exp\left(\frac{\|Z_t\|^2}{\sigma^2} - 1\right)\middle|\,\mathcal{F}_t\right]\right\}^{c_t} \le 1.$$

Consequently,

$$\mathcal{S}_n := \exp\left(\sum_{t=0}^{n-1} c_t\left(\frac{\|Z_t\|^2}{\sigma^2} - 1\right)\right), \quad n \ge 0,$$

is a nonnegative supermartingale with $\mathcal{S}_0 = 1$. Ville's inequality shows that, with probability at least $1 - \beta/2$, simultaneously for every $n \ge 0$,

$$\sum_{t=0}^{n-1}\eta_t^2\|Z_t\|^2 \le \sigma^2\sum_{t=0}^{n-1}\eta_t^2 + \sigma^2\eta_*^2\ell_\beta \le \sigma^2 Q_\infty + \sigma^2\eta_*^2\ell_\beta. \tag{5}$$

Step 3: Time-uniform control of the stopped directional noise. We first record the directional consequence of the conditional noise assumption. For every $a, y \in \mathbb{R}$,

$$\begin{aligned}\mathrm{e}^{ay} - 1 - ay &\le \frac{a^2y^2}{2}\mathrm{e}^{|ay|} \\ &\le \frac{a^2}{2}\mathrm{e}^{a^2/2}y^2\mathrm{e}^{y^2/2} \\ &\le \frac{a^2}{2}\mathrm{e}^{a^2/2}\left(\mathrm{e}^{y^2} - 1\right).\end{aligned}$$

If $u_t \ne 0$, set $Y_t := \langle Z_t, u_t\rangle/(\sigma\|u_t\|)$. Conditional centering gives $\mathbb{E}[Y_t \mid \mathcal{F}_t] = 0$, while the conditional norm-square MGF gives $\mathbb{E}[\mathrm{e}^{Y_t^2} \mid \mathcal{F}_t] \le \mathrm{e}$. Therefore,

$$\mathbb{E}\left[\mathrm{e}^{aY_t}\middle|\,\mathcal{F}_t\right] \le 1 + \frac{\mathrm{e}-1}{2}a^2\mathrm{e}^{a^2/2} \le \mathrm{e}^{2a^2}.$$

Here the last inequality follows from $(\mathrm{e}-1)/2 < 1$ and $s\mathrm{e}^{s/2} \le \mathrm{e}^{2s} - 1$ for $s \ge 0$. The case $u_t = 0$ is immediate. Taking $a = \lambda\sigma\|u_t\|$ yields

$$\mathbb{E}\left[\exp\left(\lambda\langle Z_t, u_t\rangle\right)|\,\mathcal{F}_t\right] \le \exp\left(2\sigma^2\lambda^2\|u_t\|^2\right) \tag{6}$$

for every $\mathcal{F}_t$-measurable $u_t \in \mathcal{H}$ and every $\lambda \in \mathbb{R}$. Fix $b > 0$, and define the stopping time $\tau_b := \inf\{t \ge 0 : D_t > b\}$. On $\{t < \tau_b\}$, Step 1 gives

$$\|\bar{x}_{t+1} - x^*\|^2 \le \|x_t - x^*\|^2 = 2D_t \le 2b.$$

Define the stopped sum

$$M_n^{(b)} := -\sum_{t=0}^{n-1}\eta_t\langle Z_t, \bar{x}_{t+1} - x^*\rangle\mathbf{1}_{\{t<\tau_b\}}.$$

For any fixed $\lambda > 0$, define

$$\mathcal{E}_n^{(b)}(\lambda) := \exp\left(\lambda M_n^{(b)} - 4\sigma^2 b\lambda^2 \sum_{t=0}^{n-1} \eta_t^2\right).$$

The indicator $\mathbf{1}_{\{t<\tau_b\}}$ and the vector $\bar{x}_{t+1} - x^*$ are $\mathcal{F}_t$-measurable. Hence, by equation 6,

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}_{n+1}^{(b)}(\lambda)\middle|\mathcal{F}_n\right] &\le \mathcal{E}_n^{(b)}(\lambda)\exp\left(2\sigma^2\lambda^2\eta_n^2\|\bar{x}_{n+1} - x^*\|^2\mathbf{1}_{\{n<\tau_b\}} - 4\sigma^2 b\lambda^2\eta_n^2\right)\\ &\le \mathcal{E}_n^{(b)}(\lambda).\end{aligned}$$

Thus $\{\mathcal{E}_n^{(b)}(\lambda)\}_{n\ge 0}$ is a nonnegative supermartingale starting from one. Ville's inequality gives, with probability at least $1 - \beta/2$, simultaneously for all $n \ge 0$,

$$M_n^{(b)} \le 4\sigma^2 b\lambda Q_\infty + \frac{\ell_\beta}{\lambda}.$$

Taking $\lambda = \sqrt{\ell_\beta/(4\sigma^2 b Q_\infty)}$ yields

$$M_n^{(b)} \le 4\sigma\sqrt{bQ_\infty\ell_\beta}, \quad \forall n \ge 0. \tag{7}$$

Step 4: Stopping-time contradiction and calibration. Intersect the events in equation 5 and equation 7; this intersection has probability at least $1 - \beta$. Set $b = B_\beta$, and suppose on this event that $\tau_b < \infty$. Summing equation 4 up to τ_b gives

$$\begin{aligned}D_{\tau_b} &\le D_0 + \frac{1}{2}\sum_{t=0}^{\tau_b-1}\eta_t^2\|Z_t\|^2 + M_{\tau_b}^{(b)}\\ &\le C_\beta + 4\sqrt{bq_\beta}.\end{aligned}$$

The definition of B_β is exactly the positive solution of $b = C_\beta + 4\sqrt{bq_\beta}$. Therefore, $D_{\tau_b} \le b$, contradicting the definition of τ_b. Hence $\tau_b = \infty$, and $D_t \le B_\beta$ for every $t \ge 0$ on the same event. □

Remark B.9. *For the dyadic schedule detailed in Algorithm 1, we have*

$$Q_\infty = \sum_{j=j_0}^{\infty} N_j\gamma_j^2 = \sum_{j=j_0}^{\infty} N_j\frac{R^2}{\sigma^2 c_h^2 N_j h_j^2} = \frac{R^2}{\sigma^2}c_h^{-2}\sum_{j=j_0}^{\infty}\frac{1}{h_j^2} \le \frac{R^2}{\sigma^2}.$$

Then Lemma B.8 holds. Suppose that we assign probability $\alpha/2$ to this event. Then we have

$$\begin{aligned}C_{\alpha/2} &= \frac{R^2}{2} + \frac{\sigma^2 Q_\infty}{2} + \frac{\sigma^2\eta_*^2\ell_{\alpha/2}}{2}\\ &\le \frac{R^2}{2} + \frac{R^2}{2} + \frac{\sigma^2\ell_{\alpha/2}}{2}\cdot\frac{R^2}{\sigma^2 c_h^2 N_{j_0} h_{j_0}^2} \le R^2 + \frac{R^2\ell_{\alpha/2}}{2}.\end{aligned}$$

The last inequality follows from the fact that N_{j_0}, h_{j_0}, c_h are all at least 1. Together with the fact that $q_{\alpha/2} = \sigma^2 Q_\infty \ell_{\alpha/2} \le R^2\ell_{\alpha/2}$, we have

$$B_{\alpha/2} = (\sqrt{C_{\alpha/2} + 4q_{\alpha/2}} + 2\sqrt{q_{\alpha/2}})^2 \le R^2\left(\sqrt{1 + \frac{9}{2}\ell_{\alpha/2}} + 2\sqrt{\ell_{\alpha/2}}\right)^2.$$

Since $\ell_{\alpha/2} = \log(4/\alpha) \ge 2\log 2$ and

$$\sqrt{1 + \frac{9}{2}y} \le \frac{5}{2}\sqrt{y}, \quad \forall y \ge 2\log 2,$$

we have

$$B_{\alpha/2} \le \frac{81}{4}R^2\log\frac{4}{\alpha}.$$

B.3 Proof of the sufficiency theorem

Throughout this subsection, h is eventually nondecreasing with $\sum_{j\ge 1} h(2^j)^{-2} < \infty$, and $\eta = \eta(h, L, R, \sigma)$ is the schedule produced by Algorithm 1, with the associated quantities J_h, c_h, j_0, $N_j = 2^j$, $h_j = h(N_j)$, and $\gamma_j = R/(\sigma c_h \sqrt{N_j} h_j)$. Recall $\Delta_t = f(x_t) - f^*$ and $D(y, x) = \|y - x\|^2/2$.

We first record the two calibration properties of the schedule that make Theorem 3.2 and Lemma B.8 applicable.

Lemma B.10 (Schedule calibration and finite energy)**.** *The schedule of Algorithm 1 satisfies:*

(i) $\{\gamma_j\}_{j\ge j_0}$ *is nonincreasing, and* $0 < \gamma_j \le 1/(2L)$ *for every* $j \ge j_0$*;*

(ii) the total squared-step energy is finite, with

$$Q_\infty := \sum_{t\ge 0} \eta_t^2 = \sum_{j=j_0}^{\infty} N_j \gamma_j^2 \le \frac{R^2}{\sigma^2}.$$

Proof. Since $j_0 \ge J_h$, the sequence $\{h_j\}_{j\ge j_0}$ is nondecreasing, and N_j is increasing. Hence $\sqrt{N_j} h_j$ is increasing and $\gamma_j = R/(\sigma c_h \sqrt{N_j} h_j)$ is nonincreasing. The definition of j_0 gives $\gamma_{j_0} = R/(\sigma c_h \sqrt{N_{j_0}} h_{j_0}) \le 1/(2L)$, so $\gamma_j \le \gamma_{j_0} \le 1/(2L)$ for every $j \ge j_0$. This proves (i).

For (ii), the steps vanish before epoch j_0 and are constant on each later epoch, so

$$Q_\infty = \sum_{j=j_0}^{\infty} N_j \gamma_j^2 = \sum_{j=j_0}^{\infty} N_j \frac{R^2}{\sigma^2 c_h^2 N_j h_j^2} = \frac{R^2}{\sigma^2 c_h^2} \sum_{j=j_0}^{\infty} \frac{1}{h_j^2} \le \frac{R^2}{\sigma^2},$$

because $j_0 \ge J_h$ and $c_h^2 \ge \sum_{j\ge J_h} h_j^{-2}$ by construction. □

The next lemma is where reciprocal-square summability enters the constructive direction: it converts Lemma 3.3 into two constants that do not depend on the epoch index.

Lemma B.11 (Epoch constants)**.** *For every* $\alpha \in (0, 1)$*, the quantities*

$$K_{h_1} := \sup_{j\ge j_0} \frac{(24\log 2) j + 18\log 2 + 6}{h_j^2}, \qquad K_{h_2} := \sup_{j\ge j_0} \frac{9\log(1/\alpha)}{h_j^2}$$

are finite. Moreover $K_{h_2} = 9\log(1/\alpha)/h_{j_0}^2 \le 9\log(1/\alpha)$.

Proof. By the definition of j_0 we have $h_{j_0} \ge 1$, and $\{h_j\}_{j\ge j_0}$ is nondecreasing, so $h_j \ge 1$ for every $j \ge j_0$. This gives the stated evaluation of K_{h_2}. For K_{h_1}, Lemma 3.3 gives $j/h_j^2 \to 0$, while $(18\log 2 + 6)/h_j^2 \to 0$. Hence the sequence whose supremum defines K_{h_1} converges to zero, and is therefore bounded. □

We now fix the confidence allocation. For $j \ge j_0$ and $1 \le r \le N_j$, set

$$\delta_{j,r} := \frac{\alpha}{2^{2(j+1)}}, \qquad y_{j,r} := \begin{cases} x_{N_j}, & 1 \le r < N_j/2, \\ x^*, & N_j/2 \le r \le N_j, \end{cases} \tag{8}$$

and let $\mathcal{A}_{j,r}$ denote the event

$$\mathcal{A}_{j,r} := \left\{ f(x_{N_j+r}) - f(y_{j,r}) \le \frac{4D(y_{j,r}, x_{N_j})}{\gamma_j r} + 6\sigma^2 \gamma_j \mathsf{H}_r + 9\sigma^2 \gamma_j \log \frac{1}{\delta_{j,r}} \right\}.$$

Both choices of $y_{j,r}$ are almost surely finite and $\mathcal{F}_{N_j}$-measurable, and $\gamma_j \le 1/(2L)$ by Lemma B.10. Theorem 3.2 therefore applies with $m = N_j$ and this r, giving $\mathbb{P}(\mathcal{A}_{j,r} \mid \mathcal{F}_{N_j}) \ge 1 - \delta_{j,r}$ and hence $\mathbb{P}(\mathcal{A}_{j,r}) \ge 1 - \delta_{j,r}$. Let

$$\mathcal{G} := \left\{ D(x^*, x_t) \le B_{\alpha/2},\ \forall t \ge 0 \right\}, \qquad B_{\alpha/2} \le \frac{81}{4} R^2 \log \frac{4}{\alpha}, \tag{9}$$

which by Lemma B.10 and Lemma B.8 satisfies $\mathbb{P}(\mathcal{G}) \ge 1 - \alpha/2$.

Lemma B.12 (Within-epoch estimate). *Let $j \geq j_0$ and $1 \leq r \leq N_j$. On the event $\mathcal{G} \cap \mathcal{A}_{j,r}$:*

(i) if $1 \leq r < N_j/2$, then

$$f(x_{N_j+r}) - f(x_{N_j}) \leq \frac{K_{h_1} + K_{h_2}}{c_h} \cdot R\sigma \frac{h_j}{\sqrt{N_j}};$$

(ii) if $N_j/2 \leq r \leq N_j$, then

$$\Delta_{N_j+r} \leq \left(162 c_h \log \frac{4}{\alpha} + \frac{K_{h_1} + K_{h_2}}{c_h}\right) R\sigma \frac{h_j}{\sqrt{N_j}}.$$

Proof. We first bound the stochastic terms, which are common to both cases. Since $r \leq N_j$, we have $\mathsf{H}_r \leq \log r + 1 \leq j \log 2 + 1$, and $\log(1/\delta_{j,r}) = 2(j+1)\log 2 + \log(1/\alpha)$ by equation 8. Using $\sigma^2 \gamma_j = R\sigma/(c_h \sqrt{N_j} h_j)$, we obtain

$$\begin{aligned} 6\sigma^2\gamma_j \mathsf{H}_r + 9\sigma^2\gamma_j \log \frac{1}{\delta_{j,r}} &\leq \frac{R\sigma}{c_h\sqrt{N_j}h_j}\left(6(j\log 2 + 1) + 9\left(2(j+1)\log 2 + \log\frac{1}{\alpha}\right)\right) \\ &= \frac{1}{c_h} \cdot R\sigma \frac{h_j}{\sqrt{N_j}}\left(\frac{(24\log 2)j + 18\log 2 + 6}{h_j^2} + \frac{9\log(1/\alpha)}{h_j^2}\right) \\ &\leq \frac{K_{h_1} + K_{h_2}}{c_h} \cdot R\sigma \frac{h_j}{\sqrt{N_j}}, \end{aligned}$$

the last step by Lemma B.11.

For (i), the comparator is $y_{j,r} = x_{N_j}$, so the initial distance term vanishes: $D(x_{N_j}, x_{N_j}) = 0$. On $\mathcal{A}_{j,r}$ the displayed bound is exactly the stochastic estimate above.

For (ii), the comparator is $y_{j,r} = x^*$, so $f(x_{N_j+r}) - f(y_{j,r}) = \Delta_{N_j+r}$. On $\mathcal{G}$ we have $D(x^*, x_{N_j}) \leq B_{\alpha/2}$, and $r \geq N_j/2$, so by equation 9,

$$\frac{4D(x^*, x_{N_j})}{\gamma_j r} \leq \frac{162R^2\log(4/\alpha)}{N_j\gamma_j} = \frac{162R^2\log(4/\alpha)}{N_j} \cdot \frac{\sigma c_h \sqrt{N_j} h_j}{R} = 162 c_h \log\frac{4}{\alpha} \cdot R\sigma\frac{h_j}{\sqrt{N_j}}.$$

Adding the stochastic estimate gives (ii). □

The latter-half estimate at the end of epoch $j-1$ is exactly the quantity needed to initialize the early half of epoch j. This is what propagates a single constant across all active epochs.

Lemma B.13 (Epoch propagation). *Let*

$$C_{\text{latter}} := \left(324 c_h \log\frac{4}{\alpha} + \frac{(2+\sqrt{2})(K_{h_1} + K_{h_2})}{c_h}\right) R\sigma.$$

On the event $\mathcal{G} \cap \bigcap_{j \geq j_0} \bigcap_{1 \leq r \leq N_j} \mathcal{A}_{j,r}$, we have

$$\Delta_t \leq C_{\text{latter}} \frac{h(t)}{\sqrt{t}}, \qquad \forall t \geq \frac{3N_{j_0}}{2}.$$

Proof. Fix $j > j_0$ and $1 \leq r < N_j/2$. Since $N_j = N_{j-1} + N_{j-1}$, the index N_j is the terminal position of epoch $j-1$ reached with $r = N_{j-1}$, so Lemma B.12(ii) applied to epoch $j-1$ gives

$$\Delta_{N_j} \leq \left(162 c_h \log\frac{4}{\alpha} + \frac{K_{h_1} + K_{h_2}}{c_h}\right) R\sigma \frac{h_{j-1}}{\sqrt{N_{j-1}}}.$$

Adding this to the increment bound of Lemma B.12(i) yields

$$\begin{aligned} \Delta_{N_j+r} &= \big(f(x_{N_j+r}) - f(x_{N_j})\big) + \Delta_{N_j} \\ &\leq \frac{K_{h_1} + K_{h_2}}{c_h} R\sigma \frac{h_j}{\sqrt{N_j}} + \left(162 c_h \log\frac{4}{\alpha} + \frac{K_{h_1} + K_{h_2}}{c_h}\right) R\sigma \frac{h_{j-1}}{\sqrt{N_{j-1}}}. \end{aligned}$$

Since $h_{j-1} \le h_j$ and $N_j = 2N_{j-1}$, we have $h_{j-1}/\sqrt{N_{j-1}} \le \sqrt{2}h_j/\sqrt{N_j}$, whence

$$\Delta_{N_j+r} \le \left(162\sqrt{2}c_h \log\frac{4}{\alpha} + \frac{(\sqrt{2}+1)(K_{h_1}+K_{h_2})}{c_h}\right) R\sigma\frac{h_j}{\sqrt{N_j}}. \tag{10}$$

Now let $t \ge 3N_{j_0}/2$ and write $t = N_j + r$ with $j \ge j_0$ and $1 \le r \le N_j$. If $r \ge N_j/2$, then Δ_t obeys Lemma B.12(ii); if $r < N_j/2$, then necessarily $j > j_0$ and Δ_t obeys equation 10, which dominates the former bound. In both cases

$$\Delta_t \le \left(162\sqrt{2}c_h \log\frac{4}{\alpha} + \frac{(\sqrt{2}+1)(K_{h_1}+K_{h_2})}{c_h}\right) R\sigma\frac{h_j}{\sqrt{N_j}}.$$

Finally, $N_j \le t \le 2N_j$ and $h_j \le h(t)$ by monotonicity, so $h_j/\sqrt{N_j} \le \sqrt{2}h(t)/\sqrt{t}$. Multiplying the bracket by $\sqrt{2}$ gives C_{latter}. □

It remains to absorb the finitely many iterates preceding $3N_{j_0}/2$.

Lemma B.14 (Prefix absorption)**.** *Let*

$$C_{\text{prefix}} := \left(\frac{LR^2}{2} + \frac{K_{h_1}+K_{h_2}}{c_h}\cdot R\sigma\frac{h_{j_0}}{\sqrt{N_{j_0}}}\right) \Big/ \min_{1\le n<3N_{j_0}/2} \min\left\{1, \frac{h(n)}{\sqrt{n}}\right\}.$$

On the event $\mathcal{G} \cap \bigcap_{1\le r<N_{j_0}/2} \mathcal{A}_{j_0,r}$, we have

$$\Delta_t \le C_{\text{prefix}} \min\left\{1, \frac{h(t)}{\sqrt{t}}\right\}, \qquad \forall\, 1 \le t < \frac{3N_{j_0}}{2}.$$

Proof. Algorithm 1 sets $\eta_t = 0$ for every $t < N_{j_0}$, so the iterates remain at the initial point and $x_{N_{j_0}} = x_0$. Smoothness at the minimizer gives the deterministic bound

$$\Delta_{N_{j_0}} = \Delta_t = \Delta_0 \le \frac{L}{2}\|x_0 - x^*\|^2 \le \frac{LR^2}{2}, \qquad \forall\, 1 \le t \le N_{j_0}. \tag{11}$$

For $1 \le r < N_{j_0}/2$, combining equation 11 with Lemma B.12(i) applied at $j = j_0$ gives

$$\Delta_{N_{j_0}+r} = \big(f(x_{N_{j_0}+r}) - f(x_{N_{j_0}})\big) + \Delta_{N_{j_0}} \le \frac{LR^2}{2} + \frac{K_{h_1}+K_{h_2}}{c_h}\cdot R\sigma\frac{h_{j_0}}{\sqrt{N_{j_0}}}.$$

The same constant dominates equation 11. Since the range $1 \le t < 3N_{j_0}/2$ is finite and $\min\{1, h(n)/\sqrt{n}\} > 0$ on it, dividing by that minimum gives the claim. □

C Proofs and auxiliary results for necessity

C.1 Fixed-time analysis

This subsection proves Lemmas 4.3 and 4.4. For $a > 0$ and $0 < \lambda \le L$, define

$$\varphi_{\lambda,a}(y) := \frac{\lambda}{2}\left(\sqrt{a^2+y^2} - a\right)^2, \quad q_\lambda(y) := \frac{\lambda}{2}y^2.$$

Lemma C.1 (Geometry of the flat-active family)**.** *The function $\varphi_{\lambda,a}$ is real analytic and convex on $\mathbb{R}$. Moreover, it is λ-smooth, and hence L-smooth. Its unique minimizer is $y^* = 0$; however, it is not strongly convex on any neighborhood of y^* and does not satisfy a PL inequality with any positive constant. Furthermore, for every $y \in \mathbb{R}$,*

$$\left|\varphi'_{\lambda,a}(y) - \lambda y\right| \le \lambda a, \quad 0 \le q_\lambda(y) - \varphi_{\lambda,a}(y) \le \lambda a|y|.$$

Proof. Let $s := \sqrt{a^2+y^2}$. Direct differentiation gives

$$\varphi'_{\lambda,a}(y) = \lambda y\left(1 - \frac{a}{s}\right), \quad \varphi''_{\lambda,a}(y) = \lambda\left(1 - \frac{a^3}{(a^2+y^2)^{3/2}}\right) \in [0,\lambda].$$

Thus, $\varphi_{\lambda,a}$ is convex and λ-smooth, and hence also L-smooth. It is nonnegative and vanishes only at zero, so zero is its unique minimizer. Since $\varphi''_{\lambda,a}(0) = 0$, it is not strongly convex on any neighborhood of zero. For $y \neq 0$,

$$\frac{\varphi'_{\lambda,a}(y)^2}{2\varphi_{\lambda,a}(y)} = \lambda \frac{y^2}{a^2 + y^2} \to 0, \quad \text{as } y \to 0,$$

which rules out a PL inequality with a positive constant. The identity

$$\sqrt{a^2 + y^2} - a = \frac{y^2}{\sqrt{a^2 + y^2} + a}$$

further gives

$$\varphi_{\lambda,a}(y) = \frac{\lambda}{8a^2} y^4 + \mathcal{O}(y^6), \quad \text{as } y \to 0.$$

Therefore, the objective is quartic-flat at its minimizer.

Finally,

$$|\varphi'_{\lambda,a}(y) - \lambda y| = \frac{\lambda a |y|}{\sqrt{a^2 + y^2}} \leq \lambda a.$$

Using $y^2 = (s - a)(s + a)$, we also obtain

$$q_\lambda(y) - \varphi_{\lambda,a}(y) = \lambda a(s - a) \in [0, \lambda a|y|].$$

□

We next show that, over a fixed number of iterations, the recursion for $\varphi_{\lambda,a}$ converges pathwise to the corresponding quadratic recursion as $a \downarrow 0$.

Lemma C.2 (Finite-horizon flat transfer)**.** *Fix $n \in \mathbb{N}_+$ and suppose $\lambda S_n \leq 1/2$. From the same initial point and with the same noise sequence, define*

$$Y_{t+1}^{(a)} = Y_t^{(a)} - \eta_t \left(\varphi'_{\lambda,a}(Y_t^{(a)}) + \zeta_t \right),$$
$$Y_{t+1}^{(0)} = Y_t^{(0)} - \eta_t \left(\lambda Y_t^{(0)} + \zeta_t \right).$$

Then, pathwise,

$$\max_{0 \leq t \leq n} |Y_t^{(a)} - Y_t^{(0)}| \leq \lambda a S_n \leq \frac{a}{2}.$$

In particular,

$$\varphi_{\lambda,a}(Y_n^{(a)}) \to q_\lambda(Y_n^{(0)}), \quad \text{as } a \downarrow 0.$$

Proof. Let $D_t := Y_t^{(a)} - Y_t^{(0)}$ and $r_a(y) := \varphi'_{\lambda,a}(y) - \lambda y$. The coupled recursions imply

$$D_{t+1} = (1 - \lambda\eta_t) D_t - \eta_t r_a(Y_t^{(a)}).$$

Since $0 \leq \lambda\eta_t \leq \lambda S_n \leq 1/2$, Lemma C.1 gives

$$|D_{t+1}| \leq (1 - \lambda\eta_t)|D_t| + \lambda a \eta_t.$$

Starting from $D_0 = 0$, for every $1 \leq k \leq n$, we have

$$|D_k| \leq \lambda a \sum_{i=0}^{k-1} \eta_i \prod_{r=i+1}^{k-1} (1 - \lambda\eta_r) \leq \lambda a S_n \leq \frac{a}{2}.$$

This proves the uniform pathwise bound. In particular, $Y_n^{(a)} \to Y_n^{(0)}$ as $a \downarrow 0$. Lemma C.1 and the continuity of q_λ yield

$$\begin{aligned} |\varphi_{\lambda,a}(Y_n^{(a)}) - q_\lambda(Y_n^{(0)})| &\leq |\varphi_{\lambda,a}(Y_n^{(a)}) - q_\lambda(Y_n^{(a)})| + |q_\lambda(Y_n^{(a)}) - q_\lambda(Y_n^{(0)})| \\ &\leq \lambda a |Y_n^{(a)}| + \frac{\lambda}{2} |Y_n^{(a)} - Y_n^{(0)}| (|Y_n^{(a)}| + |Y_n^{(0)}|) \to 0. \end{aligned}$$

□

The preceding pathwise convergence also transfers any marginal coverage inequality from the quartic-flat family to the limiting quadratic model.

Lemma C.3 (Transfer of marginal coverage)**.** *Fix $n \in \mathbb{N}_+$, an initial point, and a noise distribution, and suppose $\lambda S_n \le 1/2$. If*

$$\mathbb{P}\left(\varphi_{\lambda,a}(Y_n^{(a)}) \le \varepsilon_\alpha(n)\right) \ge 1-\alpha, \quad \forall a > 0,$$

then

$$\mathbb{P}\left(q_\lambda(Y_n^{(0)}) \le \varepsilon_\alpha(n)\right) \ge 1-\alpha.$$

Proof. Lemma C.2 gives almost-sure convergence of the terminal objective values. Since $(-\infty, \varepsilon_\alpha(n)]$ is closed, the closed-set part of the Portmanteau theorem gives

$$\mathbb{P}\left(q_\lambda(Y_n^{(0)}) \le \varepsilon_\alpha(n)\right) \ge \limsup_{a\downarrow 0} \mathbb{P}\left(\varphi_{\lambda,a}(Y_n^{(a)}) \le \varepsilon_\alpha(n)\right) \ge 1-\alpha.$$

□

We are now ready to prove the two fixed-horizon restrictions stated in the main text.

Proof of Lemma 4.3. If $S_n = 0$, set $\lambda_n := L/2$. Otherwise, set $\lambda_n := \min\{L, 1/S_n\}/2$. In both cases, $0 < \lambda_n \le L$ and $\lambda_n S_n \le 1/2$. Start the quartic-flat recursion at $Y_0^{(a)} = R$ and use zero noise. The limiting quadratic recursion satisfies

$$Y_n^{(0)} = R\prod_{t<n}(1-\lambda_n\eta_t).$$

Every factor lies in $[0,1]$. Therefore, the product inequality $\prod_i(1-u_i) \ge 1-\sum_i u_i$ gives

$$|Y_n^{(0)}| \ge R(1-\lambda_n S_n) \ge \frac{R}{2}.$$

The zero-noise oracle is admissible for every $a > 0$. Time-uniform validity thus implies the marginal premise of Lemma C.3, and hence

$$\varepsilon_\alpha(n) \ge q_{\lambda_n}(Y_n^{(0)}) \ge \frac{\lambda_n R^2}{8} = \frac{R^2}{16}\min\left\{L, \frac{1}{S_n}\right\}.$$

□

Proof of Lemma 4.4. Set $\lambda_n := 1/(2S_n)$, start from $Y_0^{(a)} = 0$, and take independent noises $\zeta_t \sim \mathcal{N}(0, \sigma^2/4)$. This oracle is admissible because

$$\mathbb{E}\left[\exp\left(\frac{\zeta_t^2}{\sigma^2}\right)\right] = \sqrt{2} < \mathrm{e}.$$

Since $S_n \ge 1/L$, we have $\lambda_n \le L/2$, while $\lambda_n S_n = 1/2$. Lemma C.3 therefore applies. The limiting quadratic recursion can be unrolled as

$$Y_n^{(0)} = -\sum_{t<n}\eta_t\zeta_t\prod_{s=t+1}^{n-1}(1-\lambda_n\eta_s).$$

For every $t < n$, the product inequality used above gives

$$\prod_{s=t+1}^{n-1}(1-\lambda_n\eta_s) \ge 1-\lambda_n\sum_{s=t+1}^{n-1}\eta_s \ge \frac{1}{2}.$$

Thus, $Y_n^{(0)}$ is centered Gaussian with

$$\mathrm{Var}(Y_n^{(0)}) = \frac{\sigma^2}{4}\sum_{t<n}\eta_t^2\prod_{s=t+1}^{n-1}(1-\lambda_n\eta_s)^2 \ge \frac{\sigma^2 Q_n}{16}.$$

Marginal coverage of $q_{\lambda_n}(Y_n^{(0)})$ is equivalent to

$$\mathbb{P}\left(|Y_n^{(0)}| \le \sqrt{\frac{2\varepsilon_\alpha(n)}{\lambda_n}}\right) \ge 1-\alpha.$$

By the definition $z_\alpha = \Phi^{-1}(1-\alpha/2)$, the two-sided Gaussian quantile requires

$$\frac{2\varepsilon_\alpha(n)}{\lambda_n} \ge z_\alpha^2 \operatorname{Var}(Y_n^{(0)}) \ge \frac{z_\alpha^2 \sigma^2 Q_n}{16}.$$

Substituting $\lambda_n = 1/(2S_n)$ proves the result. □

Corollary 4.5 is a direct result from Lemma 4.3 and 4.4.

Proof of Corollary 4.5. Since $h(n) = o(\sqrt{n})$, the boundary $\varepsilon_\alpha(n) \to 0$ as $n \to \infty$. Therefore, by Lemma 4.3, for sufficiently large n, we have $S_n \ge 1/L$, and additionally,

$$\frac{R^2}{16}\frac{1}{S_n} \le B\frac{h(n)}{\sqrt{n}} \implies S_n \ge \frac{R^2}{16B}\frac{\sqrt{n}}{h(n)} \to \infty.$$

Thus, for sufficiently large j, Lemma 4.4 applies, yielding

$$\frac{z_\alpha^2\sigma^2}{64}\frac{\mathcal{Q}_j}{S_{N_j}} \le B\frac{h_j}{\sqrt{N_j}} \implies p_j \ge \frac{z_\alpha^2\sigma^2}{64B}\frac{\mathcal{Q}_j}{h_j}.$$

□

C.2 Preliminaries for the fixed Gaussian instance

As mentioned in Section 4.2, the instance is given by

$$\phi(y) := \frac{L}{4}(\sqrt{R^2+y^2} - R)^2 = \varphi_{L/2,R}(y).$$

The oracle noise is Gaussian with variance $\nu^2 := \vartheta\sigma^2$, where $\vartheta := (1-\mathrm{e}^{-2})/2$. The recursion is initialized at the minimizer and uses independent $\zeta_t \sim \mathcal{N}(0,\nu^2)$ as the oracle noise:

$$Y_{t+1} = Y_t - \eta_t(\phi'(Y_t) + \zeta_t), \quad Y_0 = 0.$$

We now give some properties of this instance that will be used in the necessity proof.

Lemma C.4. *The function ϕ is convex, L-smooth, and non-PL. It is quartic-flat at the minimizer, and its derivative satisfies*

$$|\phi'(y)| \le \frac{L|y|^3}{4R^2}, \quad \forall y \in \mathbb{R}.$$

Proof. The first three properties follow from Lemma C.1. The derivative is given by

$$\phi'(y) = \frac{Ly^3}{2\sqrt{R^2+y^2}(\sqrt{R^2+y^2}+R)}.$$

Since $\sqrt{R^2+y^2}(\sqrt{R^2+y^2}+R) \ge 2R^2$, we have

$$|\phi'(y)| \le \frac{L|y|^3}{4R^2}.$$

□

This instance forces the stepsize to vanish asymptotically, as stated in the following lemma.

Lemma C.5. *Under the assumptions of Theorem 4.1, the stepsize sequence $\{\eta_t\}$ must satisfy $\eta_t \to 0$ as $t \to \infty$.*

Proof. For every $u \geq 0$, $\phi(y) \leq u$ is equivalent to

$$|y|^2 \leq r^2(u) := 4R\sqrt{\frac{u}{L}} + \frac{4u}{L}.$$

It is easy to see that $r^2(u) \to 0$ as $u \to 0^+$. Since $h(n)/\sqrt{n} \to 0$, we have $\varepsilon_\alpha(n) \to 0$ as $n \to \infty$. Hence, $r^2(\varepsilon_\alpha(n)) \to 0$ as $n \to \infty$. Conditional on the past, Y_{t+1} is a translate of $\mathcal{N}(0, \nu^2\eta_t^2)$. When $\eta_t > 0$, its conditional density is bounded by $(\sqrt{2\pi}\nu\eta_t)^{-1}$, regardless of the conditional mean. Integrating this bound over $[-r(\varepsilon_\alpha(t+1)), r(\varepsilon_\alpha(t+1))]$ yields

$$1 - \alpha \leq \mathbb{P}\left(|Y_{t+1}| \leq r(\varepsilon_\alpha(t+1))\right) \leq \sqrt{\frac{2}{\pi}} \frac{r(\varepsilon_\alpha(t+1))}{\nu\eta_t}.$$

Thus every positive η_t is bounded by a constant multiple of a quantity tending to zero. Zero steps already satisfy the same conclusion, so $\eta_t \to 0$. □

In Section 4.2, we define $N_j = 2^j$, $h_j = h(N_j)$, and

$$\hat{h}_j := \max_{\substack{N_j \leq n \leq N_{j+1} \\ n \in \mathbb{N}_+}} h(n)\sqrt{\frac{N_j}{n}}.$$

We also define $\bar{\varepsilon}_j = B\hat{h}_j/\sqrt{N_j}$ and $\beta_j^2 := r^2(\bar{\varepsilon}_j)$. Now we will give a lower bound and an upper bound for β_j^4.

Lemma C.6. *For sufficiently large $j \in \mathbb{N}_+$, we have*

$$\frac{16R^2\bar{\varepsilon}_j}{L} \leq \beta_j^4 \leq \frac{64R^2\bar{\varepsilon}_j}{L}.$$

Proof. Since $r^2(\varepsilon_j) \geq 4R\sqrt{\varepsilon_j/L}$, the lower bound is immediate. For the upper bound, we notice that

$$\bar{\varepsilon}_j = B\frac{\hat{h}_j}{\sqrt{N_j}} \leq B\frac{h_{j+1}}{\sqrt{N_j}} = \sqrt{2}B\frac{h_{j+1}}{\sqrt{N_{j+1}}} \to 0, \quad \text{as } j \to \infty.$$

Therefore, for sufficiently large j, we have $\bar{\varepsilon}_j \leq LR^2$ and

$$r^2(\bar{\varepsilon}_j) = 4R\sqrt{\frac{\bar{\varepsilon}_j}{L}} + \frac{4\bar{\varepsilon}_j}{L} \leq 8R\sqrt{\frac{\bar{\varepsilon}_j}{L}}.$$

Hence, the desired upper bound holds for sufficiently large j. □

C.3 Greedy step-mass blocks

Now we define the accumulated stepsize over the j-th dyadic epoch as

$$H_j := \sum_{t=N_j}^{N_{j+1}-1} \eta_t.$$

We consider dividing the j-th epoch into several sub-blocks: let $\tau_j > 0$ be a tolerance parameter, and define e_k as the smallest integer such that

$$H_j^{(k)} := \sum_{t=e_{k-1}}^{e_k-1} \eta_t \geq \tau_j, \quad e_0 := N_j.$$

We stop when the remaining step mass is less than τ_j. Let K be the number of complete sub-blocks and $m_k := e_k - e_{k-1}$ be the length of the k-th sub-block. Then we have the following lemma.

Lemma C.7. *For sufficiently large $j \in \mathbb{N}_+$, if $\max_{N_j \le t < N_{j+1}} \eta_t \le \tau_j/4$, we have*

$$\tau_j \le H_j^{(k)} \le \frac{5}{4}\tau_j, \quad \sum_{k=1}^{K} m_k \le N_j.$$

Note that $H_j = \sum_{t=N_j}^{N_{j+1}-1} \eta_t$. We also have

$$K \ge \frac{4}{5} \cdot \frac{H_j}{\tau_j} - \frac{4}{5}.$$

In particular, if $H_j \ge 2\tau_j$, then $K \ge (2H_j)/(5\tau_j)$.

Proof. $\sum_{k=1}^{K} m_k \le N_j$ is trivial since the sub-blocks are disjoint and contained in the j-th epoch. By the definition of e_k, $H_j^{(k)} \ge \tau_j$ for each k. Since $\eta_t \to 0$ as $t \to \infty$, we have $\eta_t \le \tau_j/4$ for sufficiently large j and $t \in [N_j, N_{j+1})$. Therefore, the minimality of e_k implies that the stepsize mass before the last step in the k-th sub-block is less than τ_j. Meanwhile, the last step contributes at most $\tau_j/4$. Hence, $H_j^{(k)} \le 5\tau_j/4$. Finally, since

$$\sum_{k=1}^{K} H_j^{(k)} \le H_j \le \sum_{k=1}^{K} H_j^{(k)} + \tau_j,$$

we have

$$H_j \le \frac{5}{4} K\tau_j + \tau_j \implies K \ge \frac{4}{5} \cdot \frac{H_j}{\tau_j} - \frac{4}{5}.$$

If $H_j \ge 2\tau_j$, then $K \ge (2H_j)/(5\tau_j)$. □

Now we set $\kappa_0 = 32/5$, and $\tau_j := \kappa_0 R^2/(L\beta_j^2)$. Since $\beta_j \to 0$, we have $\tau_j \to \infty$. Together with Lemma C.5, this guarantees $\max_{N_j \le t < N_{j+1}} \eta_t \le \tau_j/4$ for all sufficiently large j.

Lemma C.8. *For a sufficiently large $j \in \mathbb{N}_+$ with $H_j/\tau_j \ge 2$, we define $G_k := \sum_{t=e_{k-1}}^{e_k-1} \eta_t \zeta_t$, $V_k := \mathrm{Var}(G_k)$. Then the variables $G_1, \dots, G_K$ are independent centered Gaussian, and*

$$\mathcal{C} \subset \bigcap_{k=1}^{K} \{|G_k| \le 4\beta_j\}.$$

Moreover, we have

$$\frac{1}{K}\sum_{k=1}^{K} \frac{\beta_j^2}{V_k} \le \frac{25}{\vartheta} B \frac{\hat{h}_j \sqrt{N_j}}{\sigma^2 H_j}.$$

Proof. Independence and zero mean are trivial since the G_k's are linear combinations of disjoint sets of independent centered Gaussian variables. Notice that

$$Y_{e_k} - Y_{e_{k-1}} = -\sum_{t=e_{k-1}}^{e_k-1} \eta_t(\phi'(Y_t) + \zeta_t) = -\sum_{t=e_{k-1}}^{e_k-1} \eta_t \phi'(Y_t) - G_k.$$

Now we control the deterministic term. By Lemma C.4, on the event $\mathcal{C}$, we have

$$\left|\sum_{t=e_{k-1}}^{e_k-1} \eta_t \phi'(Y_t)\right| \le \sum_{t=e_{k-1}}^{e_k-1} \eta_t |\phi'(Y_t)| \le \frac{L\beta_j^3}{4R^2} H_j^{(k)} \le \frac{5}{16}\kappa_0 \beta_j = 2\beta_j.$$

Since $|Y_{e_{k-1}}|, |Y_{e_k}| \le \beta_j$ on the event $\mathcal{C}$, we have

$$|G_k| \le |Y_{e_k} - Y_{e_{k-1}}| + \left|\sum_{t=e_{k-1}}^{e_k-1} \eta_t \phi'(Y_t)\right| \le 4\beta_j.$$

Hence, $\mathcal{C} \subset \bigcap_{k=1}^{K} \{|G_k| \le 4\beta_j\}$.

Finally, by Cauchy–Schwarz, we have

$$V_k = \nu^2 \sum_{t=e_{k-1}}^{e_k-1} \eta_t^2 \ge \frac{\nu^2 (H_j^{(k)})^2}{m_k} \ge \frac{\vartheta \sigma^2 \tau_j^2}{m_k}.$$

Using the fact that $\sum_{k=1}^{K} m_k \le N_j$ and $K \ge (2H_j)/(5\tau_j)$, we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \frac{\beta_j^2}{V_k} &\le \frac{\beta_j^2}{K} \sum_{k=1}^{K} \frac{m_k}{\vartheta \sigma^2 \tau_j^2} \le \frac{\beta_j^2 N_j}{\vartheta \sigma^2 K \tau_j^2} \\
&\le \frac{5\tau_j \beta_j^2 N_j}{2\vartheta \sigma^2 H_j \tau_j^2} = \frac{5\beta_j^2 N_j}{2\vartheta \sigma^2 H_j \tau_j} \\
&= \frac{5LN_j}{2\vartheta \sigma^2 H_j} \cdot \frac{1}{\kappa_0 R^2} \cdot \beta_j^4
\end{aligned}$$

Now use the upper bound in Lemma C.6 to conclude

$$\frac{1}{K}\sum_{k=1}^{K} \frac{\beta_j^2}{V_k} \le \frac{5LN_j}{2\vartheta \sigma^2 H_j} \cdot \frac{1}{\kappa_0 R^2} \cdot \frac{64 R^2 \bar{\varepsilon}_j}{L} = \frac{25}{\vartheta} B \frac{\hat{h}_j \sqrt{N_j}}{\sigma^2 H_j}.$$

□

Now we are ready to prove Lemma 4.6.

Proof of Lemma 4.6. We use contradiction. Suppose

$$H_j > C_b \frac{B}{\sigma^2} \frac{\hat{h}_j \sqrt{N_j}}{j}.$$

Then by the definition of τ_j and Lemma C.6, we have

$$\begin{aligned}
\frac{H_j}{\tau_j} &= H_j \cdot \frac{L\beta_j^2}{\kappa_0 R^2} \\
&\ge C_b \frac{B}{\sigma^2} \frac{\hat{h}_j \sqrt{N_j}}{j} \cdot \frac{L\beta_j^2}{\kappa_0 R^2} \\
&\ge C_b \frac{B}{\sigma^2} \frac{\hat{h}_j \sqrt{N_j}}{j} \cdot \frac{L}{\kappa_0 R^2} \cdot 4R \sqrt{\frac{\bar{\varepsilon}_j}{L}} \\
&= \frac{4 C_b B^{3/2}}{\kappa_0} \frac{L^{1/2}}{R\sigma^2} \frac{\hat{h}_j^{3/2} N_j^{1/4}}{j} \to \infty, \quad \text{as } j \to \infty.
\end{aligned}$$

Then for sufficiently large j, we have $H_j/\tau_j \ge 2$. Then Lemma C.8 applies, and we have

$$\mathcal{C} \subset \bigcap_{k=1}^{K} \{|G_k| \le 4\beta_j\},$$

and

$$\frac{1}{K}\sum_{k=1}^{K} \frac{(4\beta_j)^2}{V_k} \le \frac{400}{\vartheta} B \frac{\hat{h}_j \sqrt{N_j}}{\sigma^2 H_j}.$$

By Markov's inequality, this implies that there exists at least $K/2$ sub-blocks with

$$\frac{(4\beta_j)^2}{V_k} \le \frac{800}{\vartheta} B \frac{\hat{h}_j \sqrt{N_j}}{\sigma^2 H_j} \le \gamma_0 j, \quad j \in \mathcal{G}_j, |\mathcal{G}_j| \ge K/2.$$

where the last inequality follows from the contradiction assumption and the definition of C_b.

Also, Lemma C.7 gives $K \geq (2H_j)/(5\tau_j)$. Therefore, we have

$$K \geq \frac{C_b B^{3/2}}{4} \frac{L^{1/2}}{R\sigma^2} \frac{\hat{h}_j^{3/2} N_j^{1/4}}{j}.$$

Since

$$\mathcal{C} \subset \bigcap_{k=1}^{K} \{|G_k| \leq 4\beta_j\} \subset \bigcap_{k \in \mathcal{G}_j} \{|G_k| \leq 4\beta_j\},$$

we have

$$\begin{aligned}
\mathbb{P}(\mathcal{C}) &\leq \prod_{k \in \mathcal{G}_j} \mathbb{P}(|G_k| \leq 4\beta_j) \\
&= \prod_{k \in \mathcal{G}_j} \mathbb{P}\left(\frac{|G_k|}{\sqrt{V_k}} \leq \frac{4\beta_j}{\sqrt{V_k}} \right) \\
&\leq \prod_{k \in \mathcal{G}_j} \mathbb{P}\left(|Z| \leq \sqrt{\gamma_0 j} \right) \\
&= \prod_{k \in \mathcal{G}_j} (1 - 2\Phi(-\sqrt{\gamma_0 j})) \\
&\leq \exp(-|\mathcal{G}_j| \cdot 2\Phi(-\sqrt{\gamma_0 j})) \\
&\leq \exp\left(- K \cdot \frac{\sqrt{\gamma_0 j}}{1 + \gamma_0 j} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\gamma_0 j/2} \right),
\end{aligned}$$

where the last inequality uses the Mills lower bound

$$\Phi(-x) \geq \frac{x}{1+x^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}, \quad x > 0.$$

Now substituting the lower bound of K gives

$$\mathbb{P}(\mathcal{C}) \leq \exp\left(- \frac{C_b B^{3/2}}{4\sqrt{2\pi}} \frac{L^{1/2}}{R\sigma^2} \cdot \frac{\sqrt{\gamma_0} \hat{h}_j^{3/2} 2^{j/8}}{j^{1/2}(1 + \gamma_0 j)} \right) \to 0, \quad \text{as } j \to \infty.$$

This contradicts the assumption that $\mathbb{P}(\mathcal{C}) \geq 1 - \alpha$. Hence, we must have

$$H_j \leq C_b \frac{B}{\sigma^2} \frac{\hat{h}_j \sqrt{N_j}}{j},$$

for all sufficiently large j. □

C.4 Dyadic sequence estimates

Our main focus here is to prove Lemma 4.7 and 4.8.

Proof of Lemma 4.7. Note that $H_j = S_{N_{j+1}} - S_{N_j}$. By Lemma 4.6, there exists $j_0 \in \mathbb{N}_+$ such that

$$\frac{H_j}{\sqrt{N_j}} = \sqrt{2} p_{j+1} - p_j \leq C_b \frac{B}{\sigma^2} \frac{\hat{h}_j}{j}, \quad \forall j \geq j_0.$$

The above recursion implies that

$$p_j \leq 2^{-(j-j_0)/2} p_{j_0} + \sum_{i=j_0}^{j-1} 2^{-(j-i)/2} C_b \frac{B}{\sigma^2} \frac{\hat{h}_i}{i}.$$

Note that for $j > i$, we have $\hat{h}_i \leq h_j$. Now we split the summation into two parts:

$$\begin{aligned}
p_j &\leq 2^{-(j-j_0)/2} p_{j_0} + \sum_{i=j_0}^{\lfloor j/2 \rfloor - 1} 2^{-(j-i)/2} C_b \frac{B}{\sigma^2} \frac{h_j}{i} + \sum_{i=\lfloor j/2 \rfloor}^{j-1} 2^{-(j-i)/2} C_b \frac{B}{\sigma^2} \frac{h_j}{i} \\
&\leq 2^{-(j-j_0)/2} p_{j_0} + \frac{j}{j_0} 2^{-j/4} C_b \frac{B}{\sigma^2} h_j + C_b \frac{B}{\sigma^2} \frac{2h_j}{j(1 - 2^{-1/2})}.
\end{aligned}$$

It is straightforward to verify that $j2^{-j/4} = o(1/j)$. Thus, the first two terms can be absorbed into the last term multiplied by a constant. Therefore, there exists a constant C_p such that

$$p_j \le C_p \frac{B}{\sigma^2} \frac{h_j}{j}.$$

By Corollary 4.5, we have

$$\mathcal{Q}_j \le \frac{64B}{z_\alpha^2 \sigma^2} p_j h_j \le C_q \frac{B^2}{\sigma^4} \frac{h_j^2}{j},$$

where C_q is a constant. □

Proof of Lemma 4.8. Consider the recursion

$$\mathcal{Q}_{j+1} - \mathcal{Q}_j = \sum_{t=N_j}^{N_{j+1}-1} \eta_t^2 \ge \frac{H_j^2}{N_j}.$$

Note that $H_j/\sqrt{N_j} = \sqrt{2}p_{j+1} - p_j$. Then by Cauchy–Schwarz we have

$$\begin{aligned}
\mathcal{Q}_J - \mathcal{Q}_m &\ge \sum_{j=m}^{J-1} (\sqrt{2}p_{j+1} - p_j)^2 \\
&= \sum_{j=m}^{J-1} (2p_{j+1}^2 + p_j^2 - 2\sqrt{2}p_j p_{j+1}) \\
&\ge \sum_{j=m}^{J-1} ((2-\sqrt{2})p_{j+1}^2 + (1-\sqrt{2})p_j^2) \\
&= (2-\sqrt{2})p_J^2 + (1-\sqrt{2})p_m^2 + (\sqrt{2}-1)^2 \sum_{j=m+1}^{J-1} p_j^2 \\
&\ge (1-\sqrt{2})p_m^2 + (\sqrt{2}-1)^2 \sum_{j=m+1}^{J-1} p_j^2.
\end{aligned}$$

Therefore, we have

$$\mathcal{Q}_J \ge (2-\sqrt{2})\mathcal{Q}_m + (\sqrt{2}-1)^2 \sum_{j=m+1}^{J-1} p_j^2.$$

Then by Corollary 4.5, we have

$$\mathcal{Q}_J \ge (2-\sqrt{2})\mathcal{Q}_m + (\sqrt{2}-1)^2 \frac{z_\alpha^4 \sigma^4}{2^{12} B^2} \sum_{j=m+1}^{J-1} \frac{\mathcal{Q}_j^2}{h_j^2}.$$

□