WHEN MOMENTUM HELPS POLYAK–RUPPERT AVERAGING

Alexandra Suvorikova
Independent researcher

Darina Dvinskikh
HSE University
dmdvinskikh@hse.ru

Yuriy Dorn
MSU AI Institute

Alexander Gasnikov
Innopolis University

arXiv:2610.27990v2 [cs.LG] 25 Oct 2026

ABSTRACT

Polyak–Ruppert averaging gives SGD and momentum methods the same asymptotic covariance. We characterize how fast finite runs reach it and how far they can be parallelized. For strongly convex quadratics with additive noise, we prove an online–offline bridge: the empirical risk minimizer on the samples of a tail average is one Newton step from the tail-averaged query point of any online procedure. For SGD, heavy ball, and Nesterov's method with constant momentum, this step is a boundary term at the two ends of the tail and splits the gap to the offline estimator into center, covariance, and shape errors. Necessary and sufficient conditions show that, on the least curved direction, the center error sets the pace: SGD needs $m \gg \kappa^2$ rounds but momentum only $m \gg \kappa$, where κ is the condition number, so at sample size N the critical batch size grows from $\Theta(\sqrt{N}/\kappa)$ to $\Theta(\sqrt{N/\kappa})$. The covariance error separates the momentum methods: heavy ball has relative variance excess $\Theta(\kappa^{3/2}/m)$ along the stiffest direction, whereas Nesterov has relative error $O(\sqrt{\kappa}/m)$ in every direction. The shape error is method-independent by a Berry–Esseen bound uniform in κ and batch size, so confidence intervals are valid when the other two errors vanish. We further examine how the conclusions change under heavy-tailed or state-dependent noise and clipping, and establish a first-order comparison with ERM for nonlinear Nesterov averaging under explicit conditions.

1 Introduction

Averaging connects stochastic optimization with statistical estimation. Under suitable conditions, an averaged trajectory has an asymptotically Gaussian error with the Polyak–Ruppert covariance, providing a basis for confidence sets as well as point estimates (Polyak and Juditsky, 1992). Yet a limiting law does not specify how many updates are necessary before it accurately describes the output of an algorithm. The distinction matters when observations are processed in large batches: at a fixed sample budget, larger batches leave fewer updates for the algorithm to approach its asymptotic regime. This makes the transient behavior of the method central to statistical inference. Momentum can shorten the persistence of initialization, suggesting that it could make the limiting approximation accurate earlier. However, it also changes how gradient noise accumulates along the trajectory, so faster decay of initialization does not guarantee this benefit. We therefore ask: *Does momentum reduce the number of updates needed for the asymptotic distribution to accurately describe the averaged output?*

Answering this question requires separating three properties of the output: its initialization bias should be negligible on the statistical scale, its covariance should match the target covariance, and its centered, standardized law should be close to Gaussian.

The question builds on several established lines of work. Analyses of averaged linear stochastic approximation characterize asymptotic covariance and nonasymptotic concentration (Mou et al., 2020a).Gaussian approximation and bootstrap methods provide finite-sample inference for averaged stochastic approximation and SGD (Samsonov et al., 2024; Sheshukova et al., 2026); in particular, approximating the finite-sample distribution can be more accurate than using the limiting covariance. Complementary analyses quantify the trade-off between acceleration and noise amplification (Mohammadi et al., 2021), including the effect of averaging on heavy-ball dynamics (Samuelson and Jovanović, 2024). *Our focus is on when the initialization bias becomes negligible relative to statistical fluctuations and the covariance of the averaged output matches its asymptotic prediction up to a small relative error.*

We study constant-parameter SGD, heavy ball (HB), and Nesterov's accelerated gradient method (NAG) on strongly convex quadratics with Hessian H, extreme eigenvalues μ, L, and condition number $\kappa = L/\mu$. The additive oracle noise is independent and centered, with covariance Σ. From $N = mb$ observations in m batches of size b, each method returns the average $\widehat{x}$ of its query points. We derive the pathwise representation

$$\widehat{x} - x_\star = \underbrace{-H^{-1}\bar{\xi}}_{\widehat{x}^{\mathrm{ERM}} - x_\star} + R,$$

where $\bar{\xi}$ is the sample noise mean, $\widehat{x}^{\mathrm{ERM}}$ is the empirical minimizer on the same observations, and R is an explicit method-dependent boundary term involving only trajectory endpoints. The resulting finite-window laws separate initialization bias, covariance corrections, and Gaussian approximation error.

Contribution. Our main contributions are as follows.

- **Bias and variance thresholds.** For full averages from a common deterministic start, we give necessary and sufficient conditions for negligible bias on the statistical scale and vanishing relative variance error in the least-curved eigendirection. Let v_μ be a unit eigenvector of H corresponding to μ, and set $\theta := \frac{\mu\, v_\mu^\top (x_0 - x_\star)}{\sigma_\mu}$, $\sigma_\mu^2 := v_\mu^\top \Sigma v_\mu > 0$. These conditions yield the following thresholds (Section 3).

Table 1: Calibration thresholds on the least-curved direction for full averages from a deterministic start.

Criterion in the μ direction	SGD	HB	NAG
Vanishing relative variance error	$m \gg \kappa$	$m \gg \sqrt{\kappa}$	$m \gg \sqrt{\kappa}$
Negligible normalized bias and relative variance error, $b = 1$, fixed $\theta \neq 0$	$m \gg \kappa^2$	$m \gg \kappa$	$m \gg \kappa$
Initialization-induced batch scale at fixed N, up to constants	$\frac{\sqrt{N}}{\lvert\theta\rvert\kappa}$	$\frac{\sqrt{N}}{\lvert\theta\rvert\sqrt{\kappa}}$	$\frac{\sqrt{N}}{\lvert\theta\rvert\sqrt{\kappa}}$

- **Uniform Gaussian approximation.** Under a finite third-moment assumption, we prove an $O(N^{-1/2})$ Kolmogorov bound for the least-curved projection after centering and scaling, uniformly in κ and b. Approximation by the Polyak–Ruppert law additionally incurs an initialization term and a variance correction of order κ/m for SGD and $\sqrt{\kappa}/m$ for HB and NAG. This decomposition quantifies coverage errors and identifies which terms acceleration improves (Section 3).
- **A spectral separation between HB and NAG.** For the stated HB tuning, the relative variance in the most-curved direction is $1 + \kappa^{3/2}(1 + o(1))/(8m)$ as $\kappa \to \infty$ and $m/\sqrt{\kappa} \to \infty$; calibration there requires $m \gg \kappa^{3/2}$. For the stated NAG tuning, the relative variance error is at most $30\sqrt{\kappa}/m$ uniformly over Hessian eigendirections, with a matching-order correction in the least-curved direction (Section 4).
- **Extensions to heavy tails and nonlinear objectives.** We derive conditions under which the normalized averaged output converges to the same stable law as the ERM in the scalar model, and show that finite variance alone does not guarantee uniformly valid sub-Gaussian confidence bounds centered at the ERM or at the averaged algorithmic output. For smooth strongly convex objectives with additive noise and a Lipschitz Hessian, a trajectory-dependent remainder extends the online–offline comparison. Under explicit burn-in and batch-growth conditions, tail-averaged NAG is first-order equivalent to the ERM at the same total observation budget. (Section 5).

The sharp thresholds concern full averages from a cold start under the specified parameter policies. Burn-in and tail averaging can change these thresholds; the NAG constants refer to averages of query points. The resulting comparison therefore makes the output rule, initialization, and spectral direction explicit when assessing the statistical benefit of acceleration.

Organization. Section 2 specifies the model, algorithms, output, and same-data ERM benchmark, while Sections 3 and 4 develop the finite-budget calibration criteria and spectral comparison. Section 5 treats heavy-tailed noise and nonlinear objectives, Section 6 reports the experiments, and the appendix contains proofs, auxiliary results, and extended experimental details.

2 Problem Statement

This section specifies the common model, algorithms, output, and same-data ERM benchmark.

We study averaged stochastic first-order methods on

$$f(x) := \tfrac{1}{2}\left\langle x - x_\star, H(x - x_\star)\right\rangle, \tag{1}$$

where $H = H^\mathsf{T} \succ 0$, $0 < \mu \le L$, $\mu I \preceq H \preceq LI$, and $\kappa := L/\mu$. In comparisons on the extreme eigendirections, μ and L denote the smallest and largest eigenvalues of the original Hessian. The oracle returns $G(x, \zeta) = H(x - x_\star) + \xi(\zeta)$.

Assumption 2.1 (Additive noise)**.** *Unless stated otherwise, all copies of $\xi := \xi(\zeta)$ are independent, with $\mathbb{E}\xi = 0$ and finite covariance Σ.*

Batching and methods. At round k, a batch of $b \ge 1$ samples gives $\eta_{k+1}^{(b)} := b^{-1}\sum_{j=1}^{b} \xi_{k+1,j}$, with covariance Σ/b. The method queries the oracle at a point s_k and receives the batch-averaged answer $g_k := b^{-1}\sum_{j=1}^{b} G(s_k, \zeta_{k+1,j})$. The two-step family computes

$$s_k = x_k + \gamma(x_k - x_{k-1}), \;\; x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha g_k,$$

with constant $\alpha > 0$, $\beta \in [0, 1)$, and $\gamma \ge 0$. SGD has $(\beta, \gamma) = (0, 0)$, HB has $\gamma = 0$, and NAG has $\gamma = \beta$. The parameter policies used in the comparison are given in Theorem B.2. Writing $e_k := x_k - x_\star$ and $u_k := s_k - x_\star$, so that $g_k = Hu_k + \eta_{k+1}^{(b)}$, we get the error recursion,

$$u_k = e_k + \gamma(e_k - e_{k-1}), \tag{2}$$

$$e_{k+1} = e_k + \beta(e_k - e_{k-1}) - \alpha\big(Hu_k + \eta_{k+1}^{(b)}\big). \tag{3}$$

Unless stated otherwise, all methods use the same deterministic start with $e_{-1} = e_0$, and parameters and averaging windows are fixed before observing the noise. We write $\mathcal{F}_k := \sigma(\xi_{i,j} : i \le k,\ 1 \le j \le b)$; thus u_k is $\mathcal{F}_k$-measurable and $\eta_{k+1}^{(b)}$ is independent of $\mathcal{F}_k$.

Output and comparison. We define the output by averaging the points s_k at which the stochastic gradients are evaluated. For $m \ge 1$ and $t \ge 0$, the resulting estimator of $x_\star$ is

$$\widehat{x}_{t,m} := \frac{1}{m}\sum_{k=t}^{t+m-1} s_k, \;\; \widehat{x}_{t,m} - x_\star = \bar{u}_{t,m} := \frac{1}{m}\sum_{k=t}^{t+m-1} u_k.$$

Here $t = 0$ gives a full average from a cold start, while $t > 0$ gives tail averaging after t burn-in rounds. For SGD and HB, $s_k = x_k$, so this output coincides with the iterate average. For NAG, it averages the extrapolated query points and satisfies

$$\widehat{x}_{t,m} = \frac{1}{m}\sum_{k=t}^{t+m-1} x_k + \frac{\gamma}{m}(x_{t+m-1} - x_{t-1}).$$

All finite-time NAG constants below refer to the query-point average $\widehat{x}_{t,m}$.

Offline reference. Let $N := mb$. The offline reference is the minimizer $\widehat{x}_{t,m}^{\mathrm{ERM}}$ of the cumulative loss $\frac{1}{N}\sum_{k=t}^{t+m-1}\sum_{j=1}^{b} \ell(x, \xi_{k+1,j})$, with $\ell(x, \xi) := f(x) + \langle \xi, x\rangle$,

$$\widehat{x}_{t,m}^{\mathrm{ERM}} - x_\star = -\frac{H^{-1}}{m}\sum_{k=t}^{t+m-1} \eta_{k+1}^{(b)}.$$

Its covariance is $V_\star/N$, where $V_\star := H^{-1}\Sigma H^{-1}$ is the Polyak–Ruppert (PR) covariance. The last batch $\eta_{t+m}^{(b)}$ is included in this ERM and in N, although it does not affect $\widehat{x}_{t,m}$. This benchmark uses the empirical risk, whereas the online methods only see batch-averaged gradients at their query points.

We call a method stable if, in the absence of noise, its iterates x_k converge to $x_\star$ for every initial point; for the stated parameter range this holds if and only if $\alpha\lambda_{\max}(H)(1 + 2\gamma) < 2(1 + \beta)$. Appendix A provides the details.

2.1 Averaged SGD: bounds and risk

Consider SGD with $0 < \alpha \le 1/L$, a deterministic start satisfying $\|x_0 - x_\star\| \le R_0$, and the averaged output $\widehat{x} := \widehat{x}_{0,m} = m^{-1}\sum_{k=0}^{m-1} x_k$. Let $N = mb$, $\sigma^2 := \operatorname{tr}\Sigma$, and $\mathcal{R}_{m,b} := \mathbb{E}\|\widehat{x} - x_\star\|^2$. A descent argument and quadratic contraction, respectively, yield

$$\mathcal{R}_{m,b} \le U^{\mathrm{da}}_{m,b} := \frac{R_0^2}{\alpha\mu m} + \frac{\alpha\sigma^2}{\mu b}, \tag{4a}$$

$$\mathcal{R}_{m,b} \le U^{\mathrm{quc}}_{m,b} := \frac{R_0^2}{\alpha^2\mu^2 m^2} + \frac{\sigma^2}{\mu^2 N}. \tag{4b}$$

Both inequalities concern the same estimator. At fixed α and b, the first leaves a noise term independent of m, whereas the second captures its $1/N$ decay.

Both bounds summarize the problem through scalar parameters and therefore lose information about the relative orientation of H, Σ, and e_0. For example, fixing H and e_0 and placing the same noise variance along a μ-eigendirection instead of an L-eigendirection leaves the bias and both bounds unchanged. Nevertheless, the ratio of the resulting variance contributions tends to κ^2 as $m \to \infty$ at fixed κ (Theorem D.2).

The MSE, its second-order expansion, and a refined finite-window contraction bound are given in Appendix D.1. The refined bound equals the supremum of the risk over the class specified by the scalar parameters (Theorem D.3). Thus, worst-case sharpness is compatible with substantial differences between individual problems sharing those parameters.

From risk to the output distribution. For either upper bound U, Markov's inequality gives a valid confidence ball with radius $\sqrt{U/\delta}$ at level $1-\delta$. However, agreement with the ERM risk does not ensure agreement of the mean and covariance separately, since

$$\mathcal{R}_{m,b} = \|\mathbb{E}\widehat{x} - x_\star\|^2 + \operatorname{tr}\operatorname{Cov}(\widehat{x}).$$

Squared bias can compensate for a variance deficit, even when the relative risk equals one (Section 3). We therefore analyze the finite-window laws of the averaged outputs, separating initialization bias, covariance error, and Gaussian approximation error.

3 Bias, variance, and Gaussian approximation

The exact online–offline bridge reduces finite-window calibration to the analysis of a scalar response sequence. We first obtain sharp conditions for center and variance calibration on the least-curved direction, then separate these effects from Gaussian approximation error.

For every trajectory of Equation (3) and all $t \ge 0$, $m \ge 1$, Theorems A.1 and A.2 ensure

$$\widehat{x}_{t,m} - \widehat{x}^{\mathrm{ERM}}_{t,m} = \frac{H^{-1}\Delta_{t,m}}{\alpha m}, \tag{5}$$

$$\Delta_{t,m} := e_t - e_{t+m} + \beta(e_{t+m-1} - e_{t-1}) \tag{6}$$

Hence, for i.i.d. noise and every member of the family, $\sqrt{N}\,\bar{u}_{t,m} \Rightarrow \mathcal{N}(0, V_\star)$ whenever $\sqrt{b/m}\,\Delta_{t,m} \to 0$ in probability (Theorem A.1). Under constant noise, every stable member would settle at the same point $x_\star - H^{-1}\eta$; the endpoint term carries everything that depends on how a method gets there. Proofs and further results are in Sections A.1, B.1 and B.3.

The least-curved direction. Let v_μ be a unit eigenvector of H for the eigenvalue μ, and set

$$\sigma_\mu^2 := v_\mu^\top \Sigma v_\mu,\ \theta := \frac{\mu\, v_\mu^\top e_0}{\sigma_\mu},\ Z_m := \frac{\mu\sqrt{N}}{\sigma_\mu}\, v_\mu^\top \bar{u}_{0,m}.$$

Since $v_\mu^\top H = \mu v_\mu^\top$, the projection $v_\mu^\top e_k$ follows the scalar recursion with curvature μ, driven by independent noises of variance σ_μ^2. Let r_j be the noiseless response of the oracle to a unit start: $r_j := u_j^\circ$, where $e_{-1}^\circ = e_0^\circ = 1$ and

$$u_j^\circ = e_j^\circ + \gamma(e_j^\circ - e_{j-1}^\circ), \quad e_{j+1}^\circ = e_j^\circ + \beta(e_j^\circ - e_{j-1}^\circ) - \alpha\mu\, u_j^\circ.$$

By linearity, a start $v_\mu^\top e_0$ yields the noiseless oracle points $(v_\mu^\top e_0)\, r_j$. Set

$$S_m := \sum_{j<m} r_j, \qquad Q_m := \sum_{j<m} r_j^2, \qquad B := \sum_{j\ge 0} r_j.$$

By Theorem B.1, applied to the projected recursion, for all $m, b \geq 1$,

$$v_\mu^\mathsf{T} \bar{u}_{0,m} = \frac{(v_\mu^\mathsf{T} e_0)\, S_m}{m} - \frac{1}{\mu m b} \sum_{s=1}^{m} \sum_{\ell=1}^{b} (1 - r_{m-s})\, v_\mu^\mathsf{T} \xi_{s,\ell};$$

multiplying by $\mu\sqrt{N}/\sigma_\mu$ and using $N = mb$ gives

$$Z_m = \theta \sqrt{\frac{b}{m}}\, S_m - \frac{1}{\sqrt{N}} \sum_{s=1}^{m} \sum_{\ell=1}^{b} (1 - r_{m-s})\, \frac{v_\mu^\mathsf{T} \xi_{s,\ell}}{\sigma_\mu}. \tag{7}$$

The method enters only through r_j: through S_m as the memory of the start, and through the weights $1 - r_{m-s}$ as the unfinished response to late noise.

We compare SGD with $\alpha = 1/L$; HB with $\alpha = 4/(\sqrt{L} + \sqrt{\mu})^2$ and $\beta = ((\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1))^2$; and NAG with $\alpha = 1/L$ and $\beta = \gamma = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$. For these policies $0 \leq r_{j+1} \leq r_j \leq 1$, and the memory B equals κ, $\sqrt{\kappa}$, and $2\kappa/(\sqrt{\kappa} + 1)$, respectively (Theorem B.2).

Statement 3.1. *For each policy, set $h_m := \mathbb{E} Z_m = \theta\sqrt{b/m}\, S_m$, $v_m := \mathrm{Var}(Z_m) = 1 - \frac{2S_m - Q_m}{m}$. Along any sequence with $m \to \infty$, allowing $\kappa \geq 1$, $b \geq 1$, and θ to vary,*

$$v_m \to 1 \iff \frac{B}{m} \to 0,$$

$$(h_m, v_m) \to (0, 1) \iff \frac{B}{m} \to 0 \quad \textit{and} \quad |\theta| B \sqrt{\frac{b}{m}} \to 0.$$

Theorem B.3 presents the result in detail. Matching the PR risk is not enough: the relative risk equals $v_m + h_m^2$, and

$$\frac{\mathbb{E}|v_\mu^\mathsf{T} \bar{u}_{0,m}|^2}{\sigma_\mu^2/(\mu^2 N)} = v_m + h_m^2,$$

$$\theta_m^2 = \frac{2S_m - Q_m}{b S_m^2} \implies v_m + h_m^2 = 1,$$

also along sequences with B/m bounded away from zero, where $v_m \not\to 1$.

Statement 3.2 (Critical batch)**.** *Assume that $v_\mu^\mathsf{T} \xi_{s,\ell}/\sigma_\mu$ are i.i.d. with zero mean and unit variance. Their distribution remains fixed along the asymptotic sequence, even when κ, b, and θ vary. If $N \to \infty$, $\frac{bB}{N} \to 0$, and $\frac{\theta B b}{\sqrt{N}} \to h \in \mathbb{R}$, then $Z_m \Rightarrow \mathcal{N}(h, 1)$. The $(1 - \delta)$ interval for $v_\mu^\mathsf{T} x_\star$ centered at $v_\mu^\mathsf{T} \widehat{x}_{0,m}$ with half-width $z_{1-\delta/2}\sigma_\mu/(\mu\sqrt{N})$ has limiting coverage $\Phi(z_{1-\delta/2} - h) - \Phi(-z_{1-\delta/2} - h)$.*

Theorem B.5 presents the result in detail. For $\theta \neq 0$ and $m \gg B$, the scale $b_{\mathrm{crit}}(N) := \sqrt{N}/(|\theta| B)$ marks the batch size at which the initialization bias of the averaged estimator becomes comparable to the ERM standard error along v_μ. At a fixed budget $N = mb$, larger batches leave fewer iterations to reduce this bias, while the ERM standard error remains unchanged. Calibration therefore requires both $b \ll b_{\mathrm{crit}}(N)$, so that the bias is negligible, and $b \ll N/B$, so that the variance matches its PR limit. At $b = 1$ and fixed $\theta \neq 0$, the onset is $m \gg \kappa^2$ for SGD and $m \gg \kappa$ for HB and NAG. Smaller noise at a fixed start increases $|\theta|$ and therefore delays calibrated inference. Since $\theta^{-2} = v_\mu^\mathsf{T} \Sigma v_\mu / (v_\mu^\mathsf{T} \nabla f(x_0))^2$ is a directional gradient noise scale at the initial point, b_{crit} is the geometric mean of the observation budget and this noise scale, divided by B; in the scalar model it is the noise scale of McCandlish et al. (2018) (Theorem B.6).

Statement 3.3 (Gaussian shape versus PR calibration)**.** *Let the noises be i.i.d. with $\nu_3 := \mathbb{E}|v_\mu^\mathsf{T} \xi|^3/\sigma_\mu^3 < \infty$. There is a universal constant C such that, for each of the three policies and all $\kappa \geq 1$, $b \geq 1$, $m \geq 2$,*

$$d_{\mathrm{K}}\Big(\mathcal{L}\Big(\frac{Z_m - h_m}{\sqrt{v_m}}\Big), \mathcal{N}(0, 1)\Big) \leq \frac{C\nu_3}{\sqrt{N}},$$

and, if $m \geq 4B$,

$$d_{\mathrm{K}}\big(\mathcal{L}(Z_m), \mathcal{N}(0, 1)\big) \leq \underbrace{\frac{C\nu_3}{\sqrt{N}}}_{\textit{Gauss. approx.}} + \underbrace{2\phi(1)\frac{B}{m}}_{\textit{variance correction}} + \underbrace{\frac{|h_m|}{\sqrt{2\pi}}}_{\textit{initialization}}, \tag{8}$$

where ϕ is the standard normal density.

This is Theorem B.7 together with Equation (91); the variance correction B/m is sharp even for Gaussian noise (Theorem E.3). The coverage of the interval in Proposition 3.2 is within twice the right-hand side of Equation (8) of $1-\delta$. Gaussian shape is thus uniform in κ and b; the center and the variance are the bottlenecks. These intervals use known curvature and noise scale; multivariate balls and a paid certificate for unknown H and Σ are in Section E.2.

4 Acceleration gains and their limits

We now compare what acceleration changes across the Hessian spectrum. Although HB and NAG have the same order of memory on the least-curved direction, their covariance transients can differ sharply on stiff directions; we also record how burn-in changes the comparison.

Substituting B into Statements 3.1, 3.2 compares the policies on the least-curved direction, for the same full-average output and cold start (see Tab. 1).

Acceleration shortens the memory from κ to order $\sqrt{\kappa}$ and enlarges the admissible batch by the same factor; it does not change the $N^{-1/2}$ term in Statement 3.3.

For the whole spectrum, let (λ, v_λ) be an eigenpair of H with $\sigma_\lambda^2 := v_\lambda^\top \Sigma v_\lambda > 0$, and define the relative variance

$$v_m(\lambda) := \frac{N\lambda^2}{\sigma_\lambda^2}\,\mathrm{Var}\big(v_\lambda^\top \bar{u}_{0,m}\big),$$

so that $v_m(\mu) = v_m$.

Statement 4.1 (Heavy-ball covariance inflation)**.** *For HB with the parameters above and $\sigma_L^2 > 0$,*

$$v_m(L) = 1 + \frac{\kappa^{3/2}}{8m}\{1+o(1)\} \quad \textit{as } \kappa \to \infty,\ m/\sqrt{\kappa} \to \infty,$$

whereas SGD with $\alpha = 1/L$ has $v_m(L) = 1 - 1/m$. In this regime, covariance calibration of HB on this direction is equivalent to $m/\kappa^{3/2} \to \infty$.

Theorem B.9 provides the detail. For example, at $m = \kappa$ and $b = 1$ the stiff-mode HB variance ratio diverges as $\sqrt{\kappa}/8$, while the least-curved HB mode is already close to its PR variance. A Frobenius error weighted by $H^{-1}\Sigma H^{-1}$ can hide this, because the stiff-mode PR variance is smaller by a factor κ^2 when the gradient-noise variances are equal. General noise-amplification tradeoffs (Mohammadi et al., 2021, 2025) and the sign of averaged HB modal corrections (Samuelson and Jovanović, 2024, Theorem 3) are known; the precise averaged-output formula and its relation to the onset criteria are the point of this counterexample.

Statement 4.2 (Uniform spectral calibration of NAG)**.** *For NAG with the parameters above and every $m, b \geq 1$,*

$$\sup_{\lambda:\,\sigma_\lambda^2>0} |v_m(\lambda) - 1| \leq \frac{30\sqrt{\kappa}}{m},$$

and $1 - v_m(\mu) \sim \frac{11}{4}\sqrt{\kappa}/m$ as $\kappa \to \infty$ and $m/\sqrt{\kappa} \to \infty$.

Theorem B.10 provides the result in detail.

Arbitrary directions with noncommuting noise. The result admits a matrix extension. If $\Sigma \succ 0$ and $\chi_\Sigma = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, then for the same NAG output

$$\sup_{c\neq 0}\left|\frac{N\ \mathrm{Var}(c^\top \widehat{x}_{0,m})}{c^\top V_\star c} - 1\right| \leq (18\chi_\Sigma + 6\sqrt{\chi_\Sigma})\frac{\sqrt{\kappa}}{m}.$$

In particular, the error is $O(\sqrt{\kappa}/m)$ in every direction when the noise condition number remains bounded. Theorem D.5 gives the sharper comparison with any positive definite matrix commuting with H, and Theorem D.6 shows why unrestricted noise geometry cannot admit a universal constant.

If H and Σ commute, the same bound holds in every direction with positive PR variance. All these thresholds concern full averages from a cold start. Section E.3 compares full averaging, tail averaging, and switching to SGD at a common budget N', keeping the random state after burn-in. Burn-in can improve the cold-start threshold, so the threshold is not intrinsic to momentum alone, and there is no automatic ranking between continuing accelerated tail averaging and switching to SGD.

What burn-in changes. For a deterministic zero-velocity start $x_{-1} = x_0$, the retained average has mean

$$\mathbb{E}\big[v_\mu^\mathsf{T}(\widehat{x}_{t,m} - x_\star)\big] = \frac{v_\mu^\mathsf{T} e_0}{m} \sum_{j=t}^{t+m-1} r_j. \tag{9}$$

Full averaging uses the prefix S_m; tail averaging instead uses the decaying tail of the initialization response. This explains how burn-in can reduce the cold-start bias. However, the state at round t is random: its covariance should be propagated through the retained window and added to the contribution of the subsequent noises, as in Equation (95). Counting all observations gives $N' = (t+m)b$, so calibration to the retained-window covariance $V_\star/(mb)$ need not imply efficiency relative to the all-sample ERM, whose covariance is $V_\star/N'$. The full covariance formulas and the comparison with switching to SGD remain in Section E.3.

5 Extensions: heavy tails and nonlinearity

We next examine how the calibration picture changes beyond finite-variance additive quadratics. We treat heavy-tailed noise first and then extend the online–offline comparison to smooth strongly convex nonlinear objectives.

5.1 Heavy tails

In the quadratic model, the ERM error is $-H^{-1}\bar{\eta}_{t,m}^{(b)}$, so every deviation bound for the empirical noise mean applies verbatim to the offline estimator, and by Equation (6) the online output adds only the endpoint term (Theorem F.1). Heavy tails thus act on the statistical term identically online and offline; the two differ in the robustifications they admit.

Condition on the noise	Available result
Finite third moment	Uniform normal approximation (Theorem B.7).
Finite variance only	CLT with centering and variance (Theorem B.8); no sub-Gaussian radius uniformly over the class (Theorem F.2).
Strictly stable noise	stable law and calibration criteria (Theorem F.4).
Stable domain of attraction	Stable limit with the correct normalizer, under a boundary condition (Theorem F.3).

Statement 5.1 (High-confidence limit)**.** *In the scalar model with curvature μ and centered i.i.d. noise of variance σ^2, let $N = mb$. For every N and $\delta \in (0, 1/8]$ there is a symmetric noise law with variance σ^2 such that $\mathbb{P}\{|\widehat{x}^{\mathrm{ERM}} - x_\star| \geq \sigma/(\mu\sqrt{2N\delta})\} \geq \delta$. For the full averages of the three policies started at $e_0 = 0$ with $m \geq 4B$, and every $\delta \in (0, 1/16]$, there is such a law with $\mathbb{P}\{|\bar{u}_{0,m}| \geq \sigma/(4\mu\sqrt{N\delta})\} \geq \delta$.*

Theorem F.2 provides the detail. Hence no ball of sub-Gaussian radius centered at the ERM or at the averaged output is valid uniformly over the finite-variance class. A one-pass estimator that clips full gradients at a fixed query does not have this obstruction: with known H and moment bounds, its radius depends on δ only through $\sqrt{\log(2d/\delta)}$ (Theorem F.6). Along a moving trajectory, clipping yields a CLT under checkable conditions on the clipping bias, the boundary, and the conditional covariance (Theorem F.7); clipping, however, leaves the stochastically linear class, and its offline counterpart is a Huber-type M- or Z-estimator rather than a clipped empirical risk (Theorem F.8).

Statement 5.2 (Calibration under strictly stable noise)**.** *In the scalar model, let the noises be i.i.d. symmetric α_0-stable with $1 < \alpha_0 < 2$ and scale σ_{α_0}, let Z_{α_0} be standard symmetric α_0-stable, and put $\theta_{\alpha_0} = \mu e_0/\sigma_{\alpha_0}$. For any two-step parameters and all $m, b \geq 1$,*

$$\frac{\mu N^{1-1/\alpha_0}}{\sigma_{\alpha_0}}\,\bar{u}_{0,m} \overset{d}{=} h_m^{(\alpha_0)} + \Big(\frac{1}{m}\sum_{j<m}|1-r_j|^{\alpha_0}\Big)^{1/\alpha_0} Z_{\alpha_0},$$

$$h_m^{(\alpha_0)} := \theta_{\alpha_0} S_m\, b^{1-1/\alpha_0} m^{-1/\alpha_0}.$$

For the three policies, along any sequence with $m \to \infty$, allowing κ, b, and θ_{α_0} to vary, the scale factor tends to one if and only if $B/m \to 0$, and the output is calibrated if and only if, in addition, $|\theta_{\alpha_0}| B\, b^{1-1/\alpha_0} m^{-1/\alpha_0} \to 0$.

Theorem F.4 provides the details.

For the onset comparison below, keep $\theta_{\alpha_0} \neq 0$ fixed, as in the corresponding specialization of Theorem F.4.

Consequently, at $b = 1$ the onset is $m \gg \kappa^{\alpha_0}$ for SGD and $m \gg \kappa^{\alpha_0/2}$ for HB and NAG, and the critical batch becomes $N^{1/\alpha_0}/(|\theta_{\alpha_0}|B)$. Heavier tails lower the onset and raise the critical batch, because the statistical scale $N^{1/\alpha_0 - 1}$ decays more slowly, although the accuracy is worse. Mean-square guarantees carry no information in this regime, whereas the law yields calibrated intervals from stable quantiles when α_0 and σ_{α_0} are known.

5.2 Nonlinear objectives

Let $f \in C^2$ be μ-strongly convex with an M_2-Lipschitz Hessian, and let the oracle be $G(x, \zeta) = \nabla f(x) + \xi(\zeta)$. For any online procedure that queries $s_t, \ldots, s_{t+m-1}$ and returns their average $\widehat{x}_{t,m}$, with $\bar{g}$ the average of the observed gradients and $\widehat{H} = \nabla^2 f(\widehat{x}_{t,m}^{\mathrm{ERM}})$, pathwise

$$\big\|\widehat{x}_{t,m} - \widehat{H}^{-1}\bar{g} - \widehat{x}_{t,m}^{\mathrm{ERM}}\big\| \le \frac{M_2}{2\mu m} \sum_{k=t}^{t+m-1} \|s_k - \widehat{x}_{t,m}^{\mathrm{ERM}}\|^2$$

(Theorem C.1). For $M_2 = 0$ the right-hand side vanishes, which is the quadratic bridge behind Equation (6). Momentum and step policies enter the gap only through the spread of the queries, which for constant parameters is governed by the noise amplification of the method (Mohammadi et al., 2021); the right-hand side is the online–offline counterpart of the stationary bias of constant-step SGD (Dieuleveut et al., 2020).

Statement 5.3 (First-order ERM equivalence for averaged NAG)**.** *Assume, globally, $\mu I \preceq \nabla^2 f \preceq LI$ with $0 < \mu < L$ and an M_2-Lipschitz Hessian, and let the noises be i.i.d. centered with covariance Σ and $0 < \sigma^2 = \operatorname{tr}\Sigma < \infty$. Fix f, the noise law, the dimension, and x_0. Run NAG with the parameters above from $x_{-1} = x_0$ for $t + m$ rounds, where $t = \left\lceil \sqrt{\kappa}\, \log\!\left(1 + \frac{bL\mathcal{V}_0}{\sigma^2\sqrt{\kappa}}\right)\right\rceil$, $\mathcal{V}_0 = f(x_0) - f(x_\star) + \frac{\mu}{2}\|x_0 - x_\star\|^2$. Let $N' = (t+m)b$, let $\widehat{x}_{N'}^{\mathrm{ERM}}$ be the ERM on all N' observations, and let $H = \nabla^2 f(x_\star)$. Along any sequence with $m \to \infty$, $m/b \to 0$, and $\log b = o(m)$,*

$$\sqrt{N'}\big(\widehat{x}_{t,m} - \widehat{x}_{N'}^{\mathrm{ERM}}\big) \longrightarrow 0 \quad \textit{in } L^1,$$
$$\sqrt{N'}\big(\widehat{x}_{t,m} - x_\star\big) \Rightarrow \mathcal{N}(0, H^{-1}\Sigma H^{-1}).$$

The three costs of nonlinear averaging. For the burn-in, $m \ge \sqrt{\kappa}$ and $t \le m$, the explicit bound in Equation (56) simplifies to

$$\frac{\mu\sqrt{N'}}{\sigma}\mathbb{E}\big\|\widehat{x}_{t,m} - \widehat{x}_{N'}^{\mathrm{ERM}}\big\| \lesssim \underbrace{\frac{\kappa^{1/4}}{\sqrt{m}}}_{\text{endpoint}} + \underbrace{\frac{M_2\sigma\sqrt{\kappa}}{\mu L}\sqrt{\frac{m}{b}}}_{\text{nonlinearity}} + \underbrace{\sqrt{\frac{t}{m}}}_{\text{burn-in cost}} . \tag{10}$$

The hidden constant is universal. The endpoint contribution becomes small once $m \gg \sqrt{\kappa}$, exhibiting the accelerated memory scale. The nonlinear contribution requires $m/b \to 0$ for a fixed problem. Finally, the last term pays for the observations omitted from the average but included in the all-sample ERM. Since the prescribed burn-in is $O(1 + \log b)$ for a fixed problem, $\log b = o(m)$ makes this cost vanish. These three terms explain the conditions of Statement 5.3; the bound with constants is retained in the appendix.

This is Theorem C.3; Theorem C.2 gives an explicit nonasymptotic bound on $\mathbb{E}\|\widehat{x}_{t,m} - \widehat{x}_{N'}^{\mathrm{ERM}}\|$. The class includes the regularized Bernoulli logistic loss, for which the ERM is the usual regularized empirical minimizer (Section C.2). With multiplicative noise, the local covariance at the solution no longer determines the covariance accumulated along the trajectory: fixed-step SGD already acquires a covariance correction (Theorem G.2), and over oracles of arbitrary multiplicative intensity the risk of the two-step family is not bounded by local quantities (Theorem G.4); a second-moment recursion certifies a burn-in for SGD, HB, and NAG (Theorem G.3). Diminishing-gain and increasing-momentum transfers, conditional on moment bounds, and the comparison with Richardson–Romberg extrapolation are in Sections G.2 and G.3.

6 Numerical Experiments

We test the predicted center and covariance transitions in a controlled finite-population study based on Census income data. Holding the observation budget fixed, we compare finite-sample coverage predictions with Monte Carlo coverage as the batch size varies.

Setup. We treat the continuous-income UCI Adult reconstruction of Ding et al. (2021) as a finite population of $49{,}531$ persons with $d = 44$ encoded covariates. A 10^4-person pilot fits the predictor used to construct a prediction-powered, stochastically linear ridge oracle; subsequent calls sample records uniformly with replacement. At the selected ridge level, $\kappa = 1.5 \times 10^4$. Using the policies of Section 2, we evaluate nominal-95% PR intervals over 500 replicates at $N = 2^{16}$ and $b \in \{2^0, \ldots, 2^{12}\}$. Predictions use finite-sample moments, not fitted simulation outputs. The pilot and population geometry are oracle resources; Section H gives the full protocol and other ridge levels.

Throughout the linear experiments, the output is the full query-point average $\widehat{x}_{0,m}$, the observation budget is $N = mb$, and $V_\star = H^{-1}\Sigma H^{-1}$, as in Section 2. For a fixed contrast $c \in \mathbb{R}^d$ with $c^\mathsf{T} V_\star c > 0$, we infer $c^\mathsf{T} x_\star$: $c = v_\mu$ in panel (a), $c = v_L$ in panel (b), and c is the coordinate vector selecting the age coefficient in panel (c).

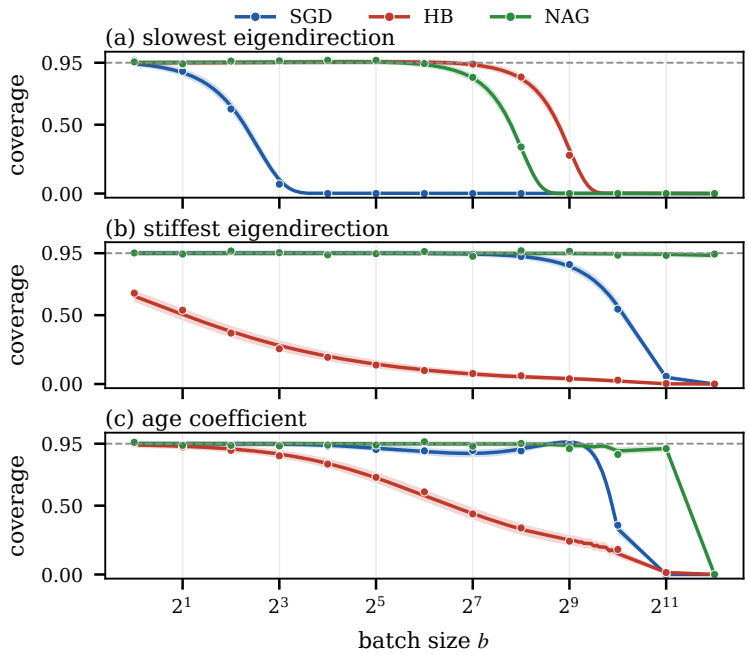

Figure 1: Coverage of nominal 95% intervals against the batch size b at the fixed budget $N = 2^{16}$, with $\kappa = 1.5 \times 10^4$. Panels (a) and (b) show the slowest and the stiffest Hessian eigendirection, and panel (c) the age coefficient. Curves are Gaussian coverage predictions based on finite-sample moments, and markers are simulated coverages. Bands show the 95% range of the Monte Carlo estimate.

Coverage at a Fixed Budget. Figure 1 shows how the coverage of the nominal 95% PR interval $c^\top \widehat{x}_{0,m} \pm 1.96 \left(\frac{c^\top V_\star c}{N}\right)^{1/2}$ changes with the batch size when the budget $N = mb$ is fixed. The half-width is 1.96 times the asymptotic standard error of $c^\top \widehat{x}_{0,m}^{\mathrm{ERM}}$. Under a Gaussian approximation, a negligible bias on this standard-error scale and a variance close to $c^\top V_\star c/N$ in relative terms are sufficient for coverage near 95%. At a single confidence level, bias and scale errors can also compensate (Section E.2.2). Larger batches leave fewer rounds to remove the initialization bias and calibrate the covariance. A bias causes an abrupt loss of coverage once it reaches the half-width, and an excess covariance a gradual one.

Panel (a): Slowest Eigendirection. On the direction $c = v_\mu$, coverage is governed by the normalized bias $h_m = \theta S_m b/\sqrt{N}$, with $h_m \sim \theta B b/\sqrt{N}$ when $m/B \to \infty$ and $\theta \neq 0$, using the notation of Statement 3.1. Coverage falls below 90% from $b = 2$ for SGD, from $b = 128$ for NAG, and from $b = 256$ for HB. The ratios 64 and 128 agree, up to the factor-two grid, with the predicted $(\sqrt{\kappa} + 1)/2 \approx 62$ and $\sqrt{\kappa} \approx 124$. Momentum thus enlarges the critical batch $b_{\mathrm{crit}}(N)$ of Statement 3.2 from order $\sqrt{N}/\kappa$ to order $\sqrt{N/\kappa}$, for fixed nonzero θ.

Panel (b): Stiffest Eigendirection. Here $c = v_L$, and coverage is governed by the relative variance $v_m(L)$ after the start is forgotten. For HB, Statement 4.1 gives $v_m(L) - 1 \sim \kappa^{3/2}/(8m)$ in its stated asymptotic regime. At $b = 1$, the relative variance is $v_m(L) = 4.6$, which implies coverage $2\Phi(1.96/\sqrt{4.6}) - 1 \approx 64\%$; the observed coverage is 66%. Since the excess grows with $b = N/m$, the coverage of HB decreases gradually. NAG, for which $|v_m(L) - 1| \leq 30\sqrt{\kappa}/m$ by Statement 4.2, remains nominal throughout.

Panel (c): Age Coefficient. For an ordinary regression coefficient both effects are present. HB loses coverage first, from $b = 8$, because of its covariance. SGD loses it from $b = 2^{10}$ because of its center. NAG retains coverage of at least 87% up to $b = 2^{11}$.

Agreement with the Theory. In every panel the observed breakdown batch coincides with the predicted one, and 110 of the 117 Monte Carlo estimates lie within their 95% bands, consistent with Monte Carlo variability. Over all 1026 configurations, predicted and observed coverage differ by a median of 0.7 percentage points. These observations

are consistent with the center condition (panel a; Statement 3.1) and the spectral covariance comparison (panel b; Statements 4.1,4.2).

7 Conclusion

Our results give a unified finite-budget account of when averaged stochastic methods become Polyak–Ruppert efficient. The online–offline bridge separates common sampling noise from method-dependent endpoint dynamics, exposing distinct scales for center, covariance, and Gaussian-shape calibration. It explains why momentum permits larger batches and why HB and NAG can approach the same limit through different covariance transients. The nonlinear NAG extension shows that this ERM viewpoint persists beyond quadratics when burn-in is charged. Acceleration therefore shortens the path to a usable statistical limit; it does not improve that limit.

References

Sunyoung An and Xiaoming Huo. A Polyak–Ruppert central limit theorem for SA-Adam with momentum and non-convergent adaptive preconditioning, 2026. URL https://arxiv.org/abs/2606.17364.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454. URL https://www.numdam.org/articles/10.1214/11-AIHP454/.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020. doi: 10.1214/18-AOS1801.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020. doi: 10.1214/19-AOS1850.

Frances Ding, Moritz Hardt, John P. Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems 34*, pages 6478–6490, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

D. H. Fuk and Sergei V. Nagaev. Probability inequalities for sums of independent random variables. *Theory of Probability and Its Applications*, 16(4):643–660, 1971. doi: 10.1137/1116071.

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

Maxime Laborde and Adam Oberman. A lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 602–612, 2020. URL https://proceedings.mlr.press/v108/laborde20a.html.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022. doi: 10.1609/aaai.v36i7.20701.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018. URL https://arxiv.org/abs/1812.06162.

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R. Jovanović. Robustness of accelerated first-order algorithms for strongly convex optimization problems. *IEEE Transactions on Automatic Control*, 66(6):2480–2495, 2021. doi: 10.1109/TAC.2020.3008297.

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R. Jovanović. Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms. *IEEE Transactions on Automatic Control*, 70(2):889–904, 2025. doi: 10.1109/TAC.2024.3453656. URL https://arxiv.org/abs/2209.11920.

Depen Morwani, Alexandru Meterez, Pranav Nair, and Sham Kakade. Compute efficiency and serial runtime tradeoffs for stochastic momentum methods, 2026. URL https://arxiv.org/abs/2606.19179.

Wenlong Mou, Chris Junchi Li, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. On linear stochastic approximation: Fine-grained Polyak–Ruppert and non-asymptotic concentration. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997. PMLR, 2020a. URL https://proceedings.mlr.press/v125/mou20a.html.

Wenlong Mou, Chris Junchi Li, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. On linear stochastic approximation: Fine-grained Polyak–Ruppert and non-asymptotic concentration. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997, 2020b. URL https://proceedings.mlr.press/v125/mou20a.html.

Sergei V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979. doi: 10.1214/aop/1176994938.

Sergei V. Nagaev. Some refinements of probabilistic and moment inequalities. *Theory of Probability and Its Applications*, 42(4):707–713, 1998.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 856–864, 2024. URL https://proceedings.mlr.press/v238/puchkin24a.html.

Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, 1994.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for Polyak–Ruppert averaged linear stochastic approximation with applications to TD learning. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-0396.

Samantha Samuelson and Mihailo R. Jovanović. Tradeoffs between convergence speed and noise amplification in first-order optimization: The role of averaging. In *2024 American Control Conference*, pages 650–655, 2024. doi: 10.23919/ACC60939.2024.10644571.

Marina Sheshukova, Denis Belomestny, Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Nonasymptotic analysis of stochastic gradient descent with the Richardson–Romberg extrapolation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Odtr1rzQMq. Also available as arXiv:2410.05106.

Marina Sheshukova, Sergey Samsonov, Denis Belomestny, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for stochastic gradient descent. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 1378–1386. PMLR, 2026. URL https://proceedings.mlr.press/v300/sheshukova26a.html.

Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015. doi: 10.1214/15-AOS1355.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020. doi: 10.1080/01621459.2018.1543124.

Kejie Tang, Weidong Liu, Yichen Zhang, and Xi Chen. Acceleration of stochastic gradient descent with momentum by averaging: Finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023. URL https://arxiv.org/abs/2305.17665.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Şimşekli, and Murat A. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems 34*, pages 18866–18877, 2021.

Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics*, 46(5):1904–1931, 2018. doi: 10.1214/18-AOS1606.

A Setup Details and the Online–Offline Decomposition

This section fixes the notation and state-space identities used throughout the appendix. It then develops the exact online–offline decomposition and clarifies the output, budget, and same-data ERM conventions.

Notation. The appendix uses the following equivalents of the main-text notation: y_k for the query point s_k; $\widehat{x}^{\mathrm{on}}_{t,m,b}$ and $\widehat{x}^{\mathrm{ERM}}_{t,m,b}$ for $\widehat{x}_{t,m}$ and $\widehat{x}^{\mathrm{ERM}}_{t,m}$; a_λ for the unit eigenvector v_λ; S_∞ (or $S_{\infty,\mathrm{alg}}$) for the memory $B = \sum_{j\geq 0} r_j$ (also written B_{alg}); and $N_{\mathrm{win}} = mb$, $N_{\mathrm{tot}} = (t+m)b$ for the budgets N and N'. The bound U_{en} of Theorem D.1 coincides with $U^{\mathrm{da}}_{m,b}$ in Equation (4a), and $U_{\mathrm{con}} \leq U^{\mathrm{quc}}_{m,b}$ is the sharper finite-window contraction bound.

Outputs and the same-data ERM. For a window $k = t, \ldots, t+m-1$, write $\bar{u}_{t,m} := m^{-1}\sum_{k=t}^{t+m-1} u_k$. The online estimator averages the points at which gradients are evaluated:

$$\widehat{x}^{\mathrm{on}}_{t,m,b} := \frac{1}{m}\sum_{k=t}^{t+m-1} y_k = x_\star + \bar{u}_{t,m}. \tag{11}$$

For any eigenvalue λ and a chosen unit eigenvector a_λ with $Ha_\lambda = \lambda a_\lambda$, write $\bar{u}^{(\lambda)}_{t,m} := a_\lambda^\mathsf{T}\bar{u}_{t,m}$ and $\sigma_\lambda^2 := a_\lambda^\mathsf{T}\Sigma a_\lambda$. The same $N_{\mathrm{win}} = mb$ observations have average noise $\bar{\eta}^{(b)}_{t,m} := m^{-1}\sum_{k=t}^{t+m-1}\eta^{(b)}_{k+1}$ and define

$$\widehat{x}^{\mathrm{ERM}}_{t,m,b} := \arg\min_x \big\{ f(x) + \big\langle \bar{\eta}^{(b)}_{t,m}, x - x_\star \big\rangle \big\} = x_\star - H^{-1}\bar{\eta}^{(b)}_{t,m}. \tag{12}$$

The main question is how quickly Equation (11) approaches the same-data benchmark Equation (12). For SGD and HB, $y_k = x_k$; for NAG, y_k is the extrapolated query. Averaging NAG iterates instead changes an endpoint term.

Let $\mathcal{F}_k$ be the sigma-algebra generated by the samples observed through round k. The query u_k is $\mathcal{F}_k$-measurable, whereas the next batch noise $\eta^{(b)}_{k+1}$ is independent of $\mathcal{F}_k$.

State-space representation. Momentum makes Equation (3) a second-order recursion. Substituting Equation (2) into the update shows that e_{k+1} depends on (e_k, e_{k-1}) and the batch noise. Stack two consecutive errors as $\mathbf{z}_k = [e_k^\mathsf{T},\, e_{k-1}^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{2d}$; then

$$\mathbf{z}_{k+1} = A\mathbf{z}_k + B_\alpha \eta^{(b)}_{k+1}, \qquad u_k = \mathsf{O}_\gamma \mathbf{z}_k = (1+\gamma)e_k - \gamma e_{k-1} = u_k \tag{13}$$

where

$$A := \begin{bmatrix} (1+\beta)I_d - \alpha(1+\gamma)H & -\beta I_d + \alpha\gamma H \\ I_d & 0 \end{bmatrix}, \quad B_\alpha := \begin{bmatrix} -\alpha I_d \\ 0 \end{bmatrix}, \quad \mathsf{O}_\gamma := \big[(1+\gamma)I_d,\, -\gamma I_d\big],$$

with $B_\alpha \in \mathbb{R}^{2d\times d}$ and $\mathsf{O}_\gamma \in \mathbb{R}^{d\times 2d}$. In particular, $\mathsf{O}_0 = [I_d,\, 0]$ selects the current error. The powers A^j propagate the cold start, while $\mathsf{O}_\gamma A^j B_\alpha$ are the impulse-response weights applied to past gradient noise.

We call the method *stable* when $\rho(A) < 1$, so $A^j z_0 \to 0$ and the noise filter is summable. Since the blocks of A are polynomials in H, the dynamics decouple by Hessian eigen-direction. On an eigenvalue λ, the characteristic polynomial is $r^2 - [1+\beta-\alpha\lambda(1+\gamma)]r + (\beta - \alpha\lambda\gamma)$. The Schur test gives the condition

$$\alpha\lambda_{\max}(H)(1+2\gamma) < 2(1+\beta).$$

The form also explains the common Polyak–Ruppert covariance. Set the gradient-noise input identically equal to a constant vector g_{in} in a stable recursion. At equilibrium, $e_{k+1} = e_k = e_{k-1} = e$ and hence $u = e$; Equation (3) reduces to $He + g_{\mathrm{in}} = 0$. Thus the equilibrium response is $e = -H^{-1}g_{\mathrm{in}}$, independently of α, β, γ. In matrix form,

$$\mathsf{O}_\gamma (I_{2d} - A)^{-1} B_\alpha = -H^{-1}. \tag{14}$$

A.1 Online–Offline Decomposition and Finite-Round Laws

Theorem A.1 (boundary decomposition)**.** *For every trajectory of Equation* (3)*, every $t \geq 0$, and every $m \geq 1$,*

$$\bar{u}_{t,m} = -H^{-1}\bar{\eta}^{(b)}_{t,m} + \frac{1}{\alpha m}H^{-1}\Delta_{t,m}, \tag{15}$$

where

$$\bar{\eta}^{(b)}_{t,m} = \frac{1}{m}\sum_{k=t}^{t+m-1}\eta^{(b)}_{k+1} = \frac{1}{m}\sum_{k=t}^{t+m-1}\left[\frac{1}{b}\sum_{j=1}^{b}\xi_{k+1,j}\right],$$
$$\Delta_{t,m} = e_t - e_{t+m} + \beta(e_{t+m-1} - e_{t-1}).$$

Fix H, α, β, γ. Suppose the single-sample noises have a common i.i.d. law with mean zero and finite covariance Σ. Let $m \to \infty$ and allow $b = b_m$ and $t = t_m$ to vary. If

$$\sqrt{b/m}\,\Delta_{t,m} \xrightarrow{\mathbb{P}} 0, \tag{16}$$

then

$$\sqrt{mb}\,\bar{u}_{t,m} \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_\star), \qquad V_\star = H^{-1}\Sigma H^{-1}. \tag{17}$$

The same limit holds for the iterate average if, additionally, $\sqrt{b/m}\,(e_{t+m-1} - e_{t-1}) \xrightarrow{\mathbb{P}} 0$.

Proof of Theorem A.1. Rearranging Equation (3) gives

$$\alpha H u_k = e_k - e_{k+1} + \beta(e_k - e_{k-1}) - \alpha \eta_{k+1}^{(b)}.$$

Summing from $k = t$ to $t + m - 1$ telescopes both finite differences:

$$\alpha H \sum_{k=t}^{t+m-1} u_k = e_t - e_{t+m} + \beta(e_{t+m-1} - e_{t-1}) - \alpha \sum_{k=t}^{t+m-1} \eta_{k+1}^{(b)}.$$

Dividing by αm and multiplying by H^{-1} proves Equation (15).

For the limit, enumerate the $N = mb$ single-sample noises in the averaging window. The first term in Equation (15) satisfies

$$-\sqrt{mb}\, H^{-1} \bar{\eta}_{t,m}^{(b)} = -H^{-1} \frac{1}{\sqrt{mb}} \sum_{k=t}^{t+m-1} \sum_{j=1}^{b} \xi_{k+1,j}.$$

The multivariate iid central limit theorem gives convergence to $\mathcal{N}(0, H^{-1}\Sigma H^{-1})$. The normalized boundary equals

$$\frac{\sqrt{mb}}{\alpha m} H^{-1} \Delta_{t,m} = \frac{1}{\alpha} \sqrt{b/m}\, H^{-1} \Delta_{t,m},$$

which is $o_{\mathbb{P}}(1)$ by Equation (16). Slutsky's theorem proves Equation (17).

Finally, writing $\bar{e}_{t,m} = m^{-1} \sum_{k=t}^{t+m-1} e_k$, the query definition gives

$$\bar{u}_{t,m} - \bar{e}_{t,m} = \frac{\gamma}{m} \left(e_{t+m-1} - e_{t-1} \right). \tag{18}$$

After multiplication by $\sqrt{mb}$, the additional endpoint condition in the theorem makes the right-hand side $o_{\mathbb{P}}(1)$ and proves the assertion for the iterate average. □

Condition Equation (16) also makes the role of burn-in explicit. If $\left\| A^k \right\| \leq K \rho_A^k$ with $\rho_A < 1$, then under a deterministic start and mini-batch noise,

$$\mathbb{E} \left\| \mathbf{z}_t \right\| \lesssim \rho_A^t \left\| \mathbf{z}_0 \right\| + b^{-1/2}. \tag{19}$$

Consequently, it is sufficient that $m \to \infty$ and $\sqrt{b/m}\, \rho_A^t \left\| \mathbf{z}_0 \right\| \to 0$. A larger batch suppresses stochastic state fluctuations, but it does not suppress a deterministic cold start; without extra burn-in, aggressive batching can therefore destroy the CLT at fixed total work.

The Same-Observation ERM. The online and offline estimators in Equations (11) and (12) use the same mb observations. Combining their definitions with Theorem A.1 gives the pathwise identity

$$\hat{x}_{t,m,b}^{\text{on}} - \hat{x}_{t,m,b}^{\text{ERM}} = \frac{H^{-1} \Delta_{t,m}}{\alpha m}. \tag{20}$$

Proposition A.2 (A pathwise gradient bridge)**.** *In the additive quadratic model, let an arbitrary online procedure query the points $x_\star + u_k$, $t \leq k \leq t + m - 1$, with u_k being $\mathcal{F}_k$-measurable, and receive $g_k = H u_k + \eta_{k+1}^{(b)}$. Put $\bar{u}_{t,m} = m^{-1} \sum_{k=t}^{t+m-1} u_k$ and $\bar{g}_{t,m} = m^{-1} \sum_{k=t}^{t+m-1} g_k$. Then, pathwise,*

$$\hat{x}_{t,m,b}^{\text{ERM}} = x_\star + \bar{u}_{t,m} - H^{-1} \bar{g}_{t,m}. \tag{21}$$

For the two-step family, $\alpha m\, \bar{g}_{t,m} = \Delta_{t,m}$, so that Equation (21) *coincides with Equation* (20).

If $\widehat{L}_{t,m}$ is the empirical risk on the window, then $\nabla \widehat{L}_{t,m}(x_\star + \bar{u}_{t,m}) = H \bar{u}_{t,m} + \bar{\eta}_{t,m}^{(b)} = \bar{g}_{t,m}$. Hence Equation (21) is one Newton step from the average.

Indeed, $\bar{g}_{t,m} = H \bar{u}_{t,m} + \bar{\eta}_{t,m}^{(b)}$, while $\hat{x}_{t,m,b}^{\text{ERM}} - x_\star = -H^{-1} \bar{\eta}_{t,m}^{(b)}$. For the two-step family, summing Equation (3) over $t \leq k \leq t + m - 1$ gives $e_{t+m} - e_t = \beta(e_{t+m-1} - e_{t-1}) - \alpha \sum_k g_k$, that is, $\alpha \sum_k g_k = \Delta_{t,m}$. Thus $\bar{g}_{t,m}$ is observable from the queried trajectory, while the Newton correction $H^{-1} \bar{g}_{t,m}$ is directly computable only with known or separately estimated curvature; see Theorem D.9.

B Finite-Budget Calibration Results

This section provides the detailed finite-budget results underlying the calibration claims in the main text. It separates initialization, covariance, and Gaussian-shape effects and then compares their behavior across SGD, HB, and NAG.

B.1 Sharp Calibration Thresholds and Critical Batching

Scalar Law and Risk. In the following scalar result, $H = \mu > 0$, $e_{-1} = e_0$ is deterministic, and the noises are independent with mean zero and variance $\sigma^2 > 0$. Write $\theta = \mu e_0/\sigma$.

Theorem B.1 (cold-start distribution and risk)**.** *For any scalar two-step parameters, set*

$$\iota = (1,1)^\mathsf{T},\ r_j = \mathsf{O}_\gamma A^j \iota,\ S_m = \sum_{j=0}^{m-1} r_j,\ Q_m = \sum_{j=0}^{m-1} r_j^2.$$

Here r_j is the deterministic oracle-point response to a unit start. For every $m, b \geq 1$, without a stability assumption,

$$\bar{u}_{0,m} = \frac{e_0 S_m}{m} - \frac{1}{\mu m b}\sum_{s=1}^{m}\sum_{j=1}^{b}(1 - r_{m-s})\xi_{s,j}. \tag{22}$$

Consequently,

$$\mathbb{E}\bar{u}_{0,m}^2 = \frac{e_0^2 S_m^2}{m^2} + \frac{\sigma^2}{b\mu^2 m^2}(m - 2S_m + Q_m). \tag{23}$$

In particular, for every signed $\theta = \mu e_0/\sigma$,

$$R_{m,b} := \frac{\mathbb{E}\bar{u}_{0,m}^2}{\sigma^2/(\mu^2 m b)} = 1 + \frac{\theta^2 b S_m^2 - 2S_m + Q_m}{m}. \tag{24}$$

Proof of Theorem B.1. In the scalar model write $\mathbf{z}_k = (e_k, e_{k-1})^\mathsf{T}$ and $\iota = (1,1)^\mathsf{T}$. Then $\mathbf{z}_{k+1} = A\mathbf{z}_k + B_\alpha \eta_{k+1}^{(b)}$, $u_k = \mathsf{O}_\gamma \mathbf{z}_k$, and $r_j = \mathsf{O}_\gamma A^j \iota$. Directly from the matrices in Equation (13),

$$(I - A)\iota = (\alpha\mu, 0)^\mathsf{T}, \qquad B_\alpha = (-\alpha, 0)^\mathsf{T}, \qquad (I - A)^{-1}B_\alpha = -\frac{1}{\mu}\iota.$$

Superposition therefore gives $u_k = e_0 r_k + \sum_{s=1}^{k} \mathsf{O}_\gamma A^{k-s} B_\alpha \eta_s^{(b)}$. For $n \geq 0$, the cumulative coefficient of one noise input is

$$\mathsf{O}_\gamma \sum_{j=0}^{n-1} A^j B_\alpha = \mathsf{O}_\gamma (I - A^n)(I - A)^{-1} B_\alpha = -\frac{1}{\mu}\{1 - \mathsf{O}_\gamma A^n \iota\} = -\frac{1}{\mu}(1 - r_n),$$

where the empty sum at $n = 0$ is zero and $r_0 = 1$. Summing $u_0, \ldots, u_{m-1}$, taking $n = m - s$, and expanding each batch average proves Equation (22). Its summands are independent and centered, so squaring the deterministic part and summing their variances gives Equation (23); division by $\sigma^2/(\mu^2 m b)$ gives Equation (24). □

Responses and Tails. Fixed-method asymptotics justify a claim non-uniformly in the condition number: an unspecified mixing constant can grow with κ. We therefore replace the cold-start remainder by an identity and then allow κ, the batch size, and the number of rounds to vary together. For compactness in the displayed parameter formulas, set $\varkappa := \sqrt{\kappa}$.

Table 2: Standard Parameter Policies and Least-Curved-Mode Response Constants

Method	**Parameters**	ϱ	c	$S_\infty := \sum_{j\geq 0} r_j$
SGD	$\alpha = \frac{1}{L}$	$1 - \frac{1}{\kappa}$	0	κ
HB	$\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$, $\beta = \left(\frac{\varkappa-1}{\varkappa+1}\right)^2$	$\frac{\varkappa-1}{\varkappa+1}$	$1 - \varrho$	$\varkappa$
NAG	$\alpha = \frac{1}{L}$, $\beta = \gamma = \frac{\varkappa-1}{\varkappa+1}$	$1 - \frac{1}{\varkappa}$	$\frac{1-\varrho}{2-\varrho}$	$\frac{2\varkappa^2}{\varkappa+1}$

Proposition B.2 (Least-curved-mode responses and tail)**.** *For the standard parameters in Table 2, with $\varkappa = \sqrt{\kappa}$, the least-curved-mode response is $r_j = (1 + cj)\varrho^j$, $0 \le r_{j+1} \le r_j \le 1$. At $\varrho = 0$, use $r_0 = 1$ and $r_j = 0$ for $j \ge 1$. For HB, $\gamma = 0$; for SGD, $\beta = \gamma = 0$. The sum and its tail are*

$$S_\infty = \frac{1-\beta}{\alpha\mu} = \frac{1}{1-\varrho} + \frac{c\varrho}{(1-\varrho)^2}, \tag{25}$$

$$S_m = S_\infty - \varrho^m \left(S_\infty + \frac{cm}{1-\varrho} \right). \tag{26}$$

For all three methods, $1 \le (1-\varrho)S_\infty \le 2$.

Proof of Theorem B.2. Put $\varkappa = \sqrt{\kappa}$. Substitution in the scalar characteristic polynomial on the μ-mode gives the following responses:

	ϱ	c
SGD	$1 - \varkappa^{-2}$	0
HB	$(\varkappa - 1)/(\varkappa + 1)$	$1 - \varrho$
NAG	$1 - \varkappa^{-1}$	$(1-\varrho)/(2-\varrho)$.

The initial conditions $e_{-1} = e_0 = 1$ determine the coefficient of the general repeated-root solution and give $r_j = (1 + cj)\varrho^j$ in each case. For $\varrho = 0$ the asserted convention follows directly. If $\varrho > 0$, then

$$r_j - r_{j+1} = \varrho^j \{(1-\varrho)(1+cj) - c\varrho\} \ge 0.$$

Indeed, the braces equal $1 - \varrho$ for SGD, $(1-\varrho)^2(j+1)$ for HB, and $(1-\varrho)^2(j+2)/(2-\varrho)$ for NAG. Hence $0 \le r_{j+1} \le r_j \le r_0 = 1$.

The geometric-series identities yield

$$\sum_{j \ge 0} (1 + cj)\varrho^j = \frac{1}{1-\varrho} + \frac{c\varrho}{(1-\varrho)^2} =: S_\infty, \quad S_\infty - S_m = \varrho^m \left(S_\infty + \frac{cm}{1-\varrho} \right),$$

which proves Equations (25) and (26). Equivalently, summing the deterministic recursion gives $S_\infty = (1-\beta)/(\alpha\mu)$. Substitution gives $S_\infty = \kappa$, $\sqrt{\kappa}$, and $2\kappa/(\sqrt{\kappa} + 1)$ for SGD, HB, and NAG. Finally, $(1-\varrho)S_\infty = 1 + \frac{c\varrho}{1-\varrho}$, which lies in $[1, 2]$ for all three displayed choices. □

Since $Q_m \le S_m \le S_\infty$, the risk gives

$$1 + \frac{\theta^2 b S_\infty^2}{m} \left(\frac{S_m}{S_\infty} \right)^2 - \frac{2S_\infty}{m} \le R_{m,b} \le 1 + \frac{\theta^2 b S_\infty^2}{m}, \qquad 0 \le 1 - \frac{S_m}{S_\infty} \le (1 + m(1-\varrho))\varrho^m. \tag{27}$$

B.1.1 Mean, Variance, and Risk

Theorem B.3 (Separate mean and covariance calibration)**.** *Choose any one of the three parameter policies above, or switch between policies along a sequence with $m \to \infty$, integer $b \ge 1$, arbitrary $\kappa \ge 1$, and arbitrary signed θ. Define the standardized mean and variance by*

$$h_m = \theta \sqrt{b/m}\, S_m, \qquad v_m = \frac{bm\mu^2}{\sigma^2} \operatorname{Var}(\bar{u}_{0,m}) = 1 - \frac{2S_m - Q_m}{m}. \tag{28}$$

Then

$$v_m \to 1 \quad \Longleftrightarrow \quad S_\infty/m \to 0, \tag{29}$$

$$(h_m, v_m) \to (0, 1) \quad \Longleftrightarrow \quad \begin{array}{c} S_\infty/m \to 0, \\ |\theta| S_\infty \sqrt{b/m} \to 0. \end{array} \tag{30}$$

If $S_\infty/m \to 0$, then $h_m = \theta S_\infty \sqrt{b/m}\,(1 + o(1))$ and

$$R_{m,b} = v_m + \frac{\theta^2 b S_\infty^2}{m} (S_m/S_\infty)^2. \tag{31}$$

In particular, if also $\theta^2 b S_\infty^2/m \to \Lambda_{\text{risk}} \in [0, \infty]$, then $R_{m,b} \to 1 + \Lambda_{\text{risk}}$.

Proof of Theorem B.3. Let $\Delta_m^{\mathrm{cov}} := 2S_m - Q_m = \sum_{j=0}^{m-1} r_j(2 - r_j)$. Since $0 \le r_j \le 1$,

$$S_m \le \Delta_m^{\mathrm{cov}} \le 2S_m, \qquad 0 \le S_m \le S_\infty. \tag{32}$$

We first show that $S_m/m \to 0$ if and only if $S_\infty/m \to 0$. One direction is immediate from $S_m \le S_\infty$. Conversely, suppose along a subsequence $S_\infty/m \ge \epsilon > 0$. With $g_\varrho = 1 - \varrho$, Theorem B.2 gives $1 \le g_\varrho S_\infty \le 2$, hence $g_\varrho m \le 2/\epsilon$ and $g_\varrho \to 0$. For all large m, $(1 - g_\varrho)^m \ge \exp(-2g_\varrho m) \ge \exp(-4/\epsilon)$. Since $r_j \ge \varrho^j$, it follows that $\frac{S_m}{m} \ge \varrho^{m-1} \ge \exp(-4/\epsilon)$, a contradiction. Thus, using $v_m = 1 - \Delta_m^{\mathrm{cov}}/m$ and Equation (32), $v_m \to 1$ if and only if $S_\infty/m \to 0$.

If $S_\infty/m \to 0$, then $mg_\varrho \ge m/S_\infty \to \infty$ and the tail bound gives

$$0 \le 1 - \frac{S_m}{S_\infty} \le (1 + mg_\varrho)\varrho^m \le (1 + mg_\varrho)e^{-mg_\varrho} \longrightarrow 0.$$

Consequently $h_m = \theta S_\infty \sqrt{b/m}\{1 + o(1)\}$, and the joint center–covariance criterion follows. Finally,

$$R_{m,b} = v_m + h_m^2 = v_m + \frac{\theta^2 b S_\infty^2}{m}(S_m/S_\infty)^2,$$

which proves Equation (31) and all its stated limits. □

The two calibration requirements must be kept separate when $\theta = 0$ or $\theta \to 0$. A risk ratio approaching one alone is insufficient for arbitrary starts: taking

$$\theta_m^2 = \frac{2S_m - Q_m}{bS_m^2}$$

makes $R_{m,b} = 1$, including along sequences with S_∞/m bounded away from zero. For fixed $\theta \ne 0$, the mean condition implies $S_\infty/m \to 0$ and the two conditions reduce to $\theta^2 bS_\infty^2/m \to 0$.

Theorem B.4 (Risk equivalence for the special normalization $\theta = 1$)**.** *Choose any one of the three parameter policies above, or switch between them along a sequence. Let $m \to \infty$, $b = b_m \ge 1$ be an integer, and $\kappa = \kappa_m \ge 1$ be arbitrary. Under $e_0 = \sigma/\mu$,*

$$R_{m,b} \longrightarrow 1 \quad \Longleftrightarrow \quad \frac{bS_\infty^2}{m} \longrightarrow 0. \tag{33}$$

More generally, if $bS_\infty^2/m \to \Lambda_{\mathrm{risk}} \in [0, \infty)$, then $R_{m,b} \to 1 + \Lambda_{\mathrm{risk}}$. If $bS_\infty^2/m \to \infty$, then $R_{m,b} \to \infty$.

Proof of Theorem B.4. For $\theta = 1$,

$$R_{m,b} - 1 = \frac{bS_m^2 - \Delta_m^{\mathrm{cov}}}{m}.$$

If $bS_\infty^2/m \to 0$, then, because $b \ge 1$ and $S_\infty \ge 1$, $S_\infty/m \le S_\infty^2/m \to 0$. Hence Section B.1.1 and Equation (32) imply $R_{m,b} \to 1$.

Conversely, suppose $R_{m,b} \to 1$. If S_∞/m failed to vanish, the argument in the preceding proof would give $S_m \ge c_{\mathrm{sub}} m$ along a subsequence for some $c_{\mathrm{sub}} > 0$. Since $\Delta_m^{\mathrm{cov}} \le 2m$ and $b \ge 1$, this would imply $R_{m,b} - 1 \ge c_{\mathrm{sub}}^2 m - 2 \to \infty$, a contradiction. Thus $S_\infty/m \to 0$, $S_m/S_\infty \to 1$, and $\Delta_m^{\mathrm{cov}}/m \to 0$. Therefore

$$\frac{bS_\infty^2}{m} = \frac{bS_m^2}{m}(S_\infty/S_m)^2 \longrightarrow 0.$$

If $bS_\infty^2/m \to \Lambda_{\mathrm{risk}} < \infty$, then $S_\infty/m \to 0$ and the same calculation gives $R_{m,b} \to 1 + \Lambda_{\mathrm{risk}}$. If $bS_\infty^2/m \to \infty$, use $g_\varrho = 1 - \varrho$ and $1 \le g_\varrho S_\infty \le 2$ to obtain

$$S_m \ge \sum_{j=0}^{m-1} \varrho^j = \frac{1 - \varrho^m}{g_\varrho} \ge \frac{S_\infty}{2}\left(1 - e^{-m/S_\infty}\right) \ge c_{\min} \min\{S_\infty, m\}, \qquad c_{\min} := \frac{1 - e^{-1}}{2}.$$

If $S_\infty \le m$, then $bS_m^2/m \ge c_{\min}^2 bS_\infty^2/m \to \infty$; if $S_\infty > m$, then $bS_m^2/m \ge c_{\min}^2 bm \to \infty$. Since $\Delta_m^{\mathrm{cov}}/m \le 2$, in both cases $R_{m,b} \to \infty$ along the full sequence. □

B.1.2 Batch Size and Persistent Bias

Corollary B.5 (The critical batch produces a shifted Gaussian)**.** *Consider $\xi_{s,j}/\sigma$. If $S_\infty/m \to 0$, $\theta S_\infty\sqrt{b/m} \to \mathfrak{m} \in \mathbb{R}$, then*

$$\frac{\mu\sqrt{mb}}{\sigma}\bar{u}_{0,m} \xrightarrow{\mathrm{d}} \mathcal{N}(\mathfrak{m}, 1). \tag{34}$$

For Gaussian noise, the finite-sample law is

$$\mathcal{N}\left(\theta\sqrt{b/m}\, S_m,\ 1 - \frac{2S_m - Q_m}{m}\right). \tag{35}$$

For a fixed $\delta \in (0,1)$, a $(1-\delta)$ interval for $x_\star$, centered at the averaged oracle point with half-width $\Phi^{-1}(1-\delta/2)\sigma/(\mu\sqrt{mb})$, has limiting coverage

$$\Phi\big(\Phi^{-1}(1-\delta/2) - \mathfrak{m}\big) - \Phi\big(-\Phi^{-1}(1-\delta/2) - \mathfrak{m}\big). \tag{36}$$

Proof of Theorem B.5. By Equation (22), the normalized centered output is a triangular weighted sum, with weights $(1-r_j)/\sqrt{N} \in [0, N^{-1/2}]$. Under $S_\infty/m \to 0$, its variance $v_m \to 1$ by Theorem B.3; moreover $N = mb \to \infty$. The Lindeberg condition follows from the common finite-variance law and the vanishing maximal weight. Hence the Lindeberg–Feller theorem and $h_m = \theta S_\infty\sqrt{b/m}\{1+o(1)\} \to \mathfrak{m}$ give Equation (34). For Gaussian noise inputs, the weighted sum is Gaussian at every finite m, b, which gives Equation (35). Evaluating the limiting $\mathcal{N}(\mathfrak{m}, 1)$ cdf at the two endpoints of the interval proves Equation (36). □

The coverage formula assumes the noise scale and curvature are known, or consistently estimated in the normalization used. Coverage is an even function of the signed shift $\mathfrak{m}$. With $\theta = 1$, the critical regime $bS_\infty^2/m \to \tau_{\mathrm{crit}}^2 < \infty$ of Theorem B.4 implies $S_\infty/m \to 0$ and gives $\mathfrak{m} = \tau_{\mathrm{crit}} \geq 0$.

The coverage formula assumes that the noise scale and curvature are known, or that the estimated normalization divided by the true normalization converges to one in probability. Set $N = mb$ and define

$$T_N = \frac{\mu\sqrt{N}}{\sigma}\bar{u}_{0,m}, \qquad \delta_N = \mathbb{E}T_N = \theta S_m\sqrt{\frac{b}{m}} = \frac{\theta b S_m}{\sqrt{N}}.$$

Thus δ_N is the standardized bias. Under the centered normal approximation used in the coverage formula, if $\delta_N \to \mathfrak{m} \in \mathbb{R}$ and the normalized variance tends to one, then

$$T_N \xrightarrow{\mathrm{d}} \mathcal{N}(\mathfrak{m}, 1).$$

The limiting coverage of a symmetric two-sided normal confidence interval depends only on $|\mathfrak{m}|$.

For the batch-size comparisons below, assume $S_\infty > 0$. Under the asymptotic estimates used in Theorem B.4, $S_m/S_\infty \to 1$ when $S_\infty/m \to 0$. In particular, with $\theta = 1$, the regime $bS_\infty^2/m \to \tau_{\mathrm{crit}}^2 < \infty$ implies $S_\infty/m \to 0$ and $\delta_N \to \tau_{\mathrm{crit}} \geq 0$.

With $N = mb$, the asymptotic conditions for the normalized variance to tend to one and for the standardized bias to vanish are, respectively,

$$\frac{bS_\infty}{N} \to 0, \qquad \frac{|\theta| S_\infty b}{\sqrt{N}} \to 0. \tag{37}$$

Assume the first condition holds. For $\theta \neq 0$, define

$$b_{\mathrm{crit}}(N) := \frac{\sqrt{N}}{|\theta| S_\infty} \asymp \begin{cases} \dfrac{\sqrt{N}}{|\theta|\kappa}, & \text{SGD}, \\ \dfrac{\sqrt{N/\kappa}}{|\theta|}, & \text{HB or NAG}. \end{cases} \tag{38}$$

Equivalently, $b_{\mathrm{crit}}(N) = \sigma\sqrt{N}/(\mu|e_0|S_\infty)$. If

$$\frac{b}{b_{\mathrm{crit}}(N)} \to c \in (0, \infty),$$

then $|\delta_N| \to c$: the standardized bias does not vanish. If, in addition, $\delta_N \to \mathfrak{m}$, the limiting normal distribution has mean $\mathfrak{m}$, with $|\mathfrak{m}| = c$. For fixed $\theta \neq 0$, this limit is $\mathfrak{m} = \mathrm{sgn}(\theta)c$. Even when δ_N does not converge, $|\delta_N| \to c$ is sufficient for the coverage of a symmetric two-sided interval to converge under the same normal approximation.

Asymptotically correct nominal coverage requires a negligible standardized bias, so the batch conditions are

$$b = o\big(b_{\mathrm{crit}}(N)\big), \qquad b = o(N/S_\infty).$$

If $b_{\mathrm{crit}}(N)$ remains bounded, the first condition cannot hold for positive integer batch sizes. When $\theta = 0$, the bias is zero, so only the variance condition $bS_\infty/N \to 0$ remains.

At $b = 1$ and fixed $\theta \neq 0$, these conditions reduce to $m/\kappa^2 \to \infty$ for SGD and $m/\kappa \to \infty$ for the accelerated parameter choices. For a general initial error, they are

$$\frac{S_\infty}{m} \to 0, \qquad \frac{\theta^2 b S_\infty^2}{m} \to 0.$$

Reducing σ at fixed $e_0 \neq 0$ decreases the standard deviation but increases the bias relative to that standard deviation. More iterations are therefore needed for nominal coverage. Both reducing the noise and increasing the batch size reduce the variance of an individual update. At fixed observation budget N, increasing the batch size also reduces the number of sequential iterations, since $m = N/b$.

For fixed m and $\theta S_m \neq 0$, treating b as a real variable gives the following balance point between squared bias and variance:

$$b_{\mathrm{balance}}(m) = \frac{m - 2S_m + Q_m}{\theta^2 S_m^2}. \tag{39}$$

It is positive when $m - 2S_m + Q_m > 0$. Under the same asymptotic estimates, if $m/S_\infty \to \infty$, then

$$b_{\mathrm{balance}}(m) \sim \frac{m}{\theta^2 S_\infty^2}.$$

If $\theta S_m = 0$, the bias is zero and there is no positive balance point with a positive variance. If $b_{\mathrm{balance}}(m) < 1$, the squared bias already exceeds the variance at batch size one.

Remark B.6 (Critical batch sizes for inference and optimization). For $\theta \neq 0$, the scalar model gives

$$\theta^{-2} = \frac{\sigma^2}{(\mu e_0)^2} = \frac{\operatorname{tr}\Sigma}{\|\nabla f(x_0)\|^2}.$$

This is the simple gradient noise scale of McCandlish et al. (2018), evaluated at the initial point. With the definition in Equation (38),

$$b_{\mathrm{crit}}(N) = \frac{\sqrt{N\theta^{-2}}}{S_\infty}.$$

Thus the batch scale for asymptotically correct coverage is the geometric mean of the observation budget and the initial gradient noise scale, divided by S_∞. For the accelerated parameter choices, $S_\infty \asymp \sqrt{\kappa}$, whereas for SGD, $S_\infty \asymp \kappa$. At the same N and θ, the accelerated batch scale is therefore larger by a factor of order $\sqrt{\kappa}$.

In a related optimization setting, Morwani et al. (2026) derive spectral-radius lower bounds and interpret them as allowing HB to maintain the compute efficiency of SGD over a larger range of batch sizes. Their model has multiplicative noise, whereas the model considered here has additive noise.

B.2 Gaussian Shape versus Polyak–Ruppert Calibration

Gaussian shape can be accurate before the PR covariance is accurate. Throughout this part, the initial condition is deterministic with zero initial velocity, and the parameters are the standard SGD, HB, or NAG parameters of Section B.1. Noise samples are i.i.d., with variance $\sigma^2 > 0$ and $\nu_3 = \mathbb{E}|\xi/\sigma|^3 < \infty$.

Write $S_\infty = S_{\infty,\mathrm{alg}} = \sum_{j\geq 0} r_j$, $N = mb$, and

$$Z_m = \frac{\mu\sqrt{N}}{\sigma}\bar{u}_{0,m}, \qquad h_m = \mathbb{E}Z_m, \qquad w_j = 1 - r_j, \qquad v_m = \frac{1}{m}\sum_{j=0}^{m-1} w_j^2 = 1 - \frac{2S_m - Q_m}{m}. \tag{40}$$

Thus $0 \leq w_j \leq 1$, and $v_m > 0$ when $m \geq 2$. The last oracle sample has weight zero because the average ends at the query preceding that sample.

Theorem B.7 (Uniform finite-window normal approximation)**.** *For all $\kappa \geq 1$, $b \geq 1$, and $m \geq 2$, there is a universal Berry–Esseen constant C_{BE} such that*

$$d_{\mathrm{W}}\big(\mathcal{L}(Z_m - h_m), \mathcal{N}(0, v_m)\big) \leq \frac{3\nu_3}{\sqrt{N}}, \tag{41}$$

$$d_{\mathrm{K}}\left(\mathcal{L}\left(\frac{Z_m - h_m}{\sqrt{v_m}}\right), \mathcal{N}(0,1)\right) \leq \frac{\sqrt{5}\, C_{\mathrm{BE}}\nu_3}{\sqrt{N}}, \tag{42}$$

$$d_{\mathrm{W}}\left(\mathcal{L}\left(\frac{Z_m - h_m}{\sqrt{v_m}}\right), \mathcal{N}(0,1)\right) \leq \frac{3\sqrt{5}\,\nu_3}{\sqrt{N}}. \tag{43}$$

In particular, these constants do not deteriorate when κ grows, even if $m/S_\infty \to 0$. If $m \geq 4S_\infty$, then

$$d_{\mathrm{W}}\big(\mathcal{L}(Z_m), \mathcal{N}(h_m, 1)\big) \leq \frac{3\nu_3}{\sqrt{N}} + 2\sqrt{\frac{2}{\pi}}\frac{S_\infty}{m}, \tag{44}$$

$$d_{\mathrm{K}}\big(\mathcal{L}(Z_m), \mathcal{N}(h_m, 1)\big) \leq \frac{\sqrt{5}\, C_{\mathrm{BE}}\nu_3}{\sqrt{N}} + 2\phi(1)\frac{S_\infty}{m}, \tag{45}$$

where ϕ denotes the standard normal density. Replacing $\mathcal{N}(h_m, 1)$ by $\mathcal{N}(0,1)$ adds at most $|h_m|$ in Wasserstein distance and $|h_m|/\sqrt{2\pi}$ in Kolmogorov distance.

Proof of Theorem B.7. Put $w_j = 1 - r_j$. The centered statistic is

$$Z_m - h_m = -\frac{1}{\sqrt{N}} \sum_{j=0}^{m-1} \sum_{i=1}^{b} w_j \, \frac{\xi_{j,i}}{\sigma}, \qquad v_m = \frac{1}{m} \sum_{j=0}^{m-1} w_j^2.$$

Let $d_i^{\mathrm{imp}} = w_{i+1} - w_i = r_i - r_{i+1}$. For SGD, d_i^{imp} is a constant multiple of ϱ^i. For HB and NAG it is respectively a constant multiple of $(i+1)\varrho^i$ and $(i+2)\varrho^i$. Thus, with $r_{\mathrm{pole}} = 1$ for SGD and $r_{\mathrm{pole}} = 2$ for HB and NAG,

$$\frac{d_i^{\mathrm{imp}}}{\binom{i+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}-1}} \quad \text{is nonincreasing in } i.$$

Write $\bar{d}_i = d_i^{\mathrm{imp}} / \binom{i+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}-1}$. Because $\bar{d}_i$ is nonincreasing, its weighted average over a prefix is at least its weighted average over the full range, with weights $\binom{i+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}-1}$. Cross-multiplication and the hockey-stick identity therefore imply, for $0 \leq j \leq m-1$,

$$\frac{w_j}{w_{m-1}} \geq \frac{\binom{j+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}}}{\binom{m+r_{\mathrm{pole}}-2}{r_{\mathrm{pole}}}} \geq \left(\frac{j}{m-1}\right)^{r_{\mathrm{pole}}}.$$

Consequently, using the bound $\sum_{j=0}^{m-1} (j/(m-1))^{2r_{\mathrm{pole}}} \geq m/(2r_{\mathrm{pole}}+1)$,

$$\sum_{j=0}^{m-1} w_j^2 \geq \frac{m}{2r_{\mathrm{pole}}+1}\, w_{m-1}^2 \geq \frac{m}{5} \max_{j<m} w_j^2. \tag{46}$$

The independent-summand Berry–Esseen inequality now gives

$$d_{\mathrm{K}}\left(\mathcal{L}\left(\frac{Z_m - h_m}{\sqrt{v_m}}\right), \mathcal{N}(0,1)\right) \leq \frac{C_{\mathrm{BE}}\nu_3}{\sqrt{b}} \frac{\sum_j w_j^3}{(\sum_j w_j^2)^{3/2}} \leq \frac{\sqrt{5}\, C_{\mathrm{BE}}\nu_3}{\sqrt{mb}}.$$

The Wasserstein version of the same inequality and $\sum_j w_j^3 \leq (\max_j w_j) \sum_j w_j^2$, with $\max_j w_j \leq 1$, yield

$$d_{\mathrm{W}}\big(\mathcal{L}(Z_m - h_m), \mathcal{N}(0, v_m)\big) \leq \frac{3\nu_3}{\sqrt{N}},$$

and standardization together with Equation (46) gives Equation (43).

If $m \geq 4S_\infty$, then $0 \leq 1 - v_m = (2S_m - Q_m)/m \leq 2S_\infty/m \leq 1/2$. Coupling centered normals by the same standard Gaussian gives

$$d_{\mathrm{W}}(\mathcal{N}(0, v_m), \mathcal{N}(0,1)) = \sqrt{2/\pi}\,(1 - \sqrt{v_m}) \leq 2\sqrt{2/\pi}\, S_\infty/m.$$

For Kolmogorov distance, differentiation gives, for $v_m \geq 1/2$,

$$d_{\mathrm{K}}(\mathcal{N}(0, v_m), \mathcal{N}(0,1)) \leq \phi(1)(1 - v_m) \leq 2\phi(1) S_\infty / m.$$

The triangle inequality proves Equations (44) and (45). Finally, translation by h_m costs at most $|h_m|$ in Wasserstein distance and $|h_m|/\sqrt{2\pi}$ in Kolmogorov distance. □

Corollary B.8 (Gaussian shape under arbitrary conditioning and parallelism)**.** *Replace the third-moment assumption by a common standardized noise law of mean zero and variance one. Along any sequence with $m \geq 2$, $b \geq 1$, $N = mb \to \infty$, and arbitrary $\kappa \geq 1$, the three parameter policies satisfy*

$$\frac{Z_m - h_m}{\sqrt{v_m}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Proof of Theorem B.8. The same weighted triangular array has total variance one after exact standardization. From Equation (46), its largest standardized coefficient is $O(N^{-1/2})$. A common law therefore implies Lindeberg's condition by uniform integrability, and the triangular-array CLT proves the result without a third-moment assumption. □

B.3 Spectral Separation of Heavy Ball and Nesterov

Now the procedure changes. Momentum is one candidate, but what has to be chosen is the combination of dynamics, steps, averaging and burn-in.

B.3.1 Comparing Methods on the Least-Curved Eigendirection

We first apply the criteria of Section B.1 to the same full-average output, zero-velocity cold start, and additive-noise model on the eigendirection with curvature μ. The parameter choices are those of Theorem B.2.The comparison follows by substituting into Theorems B.3 and B.4,

$$S_{\infty,\mathrm{SGD}} = \kappa, \qquad S_{\infty,\mathrm{HB}} = \sqrt{\kappa}, \qquad S_{\infty,\mathrm{NAG}} = \frac{2\kappa}{\sqrt{\kappa}+1}.$$

For arbitrary θ, centered variance calibration requires $m \gg S_\infty$. Within this regime, negligible initial bias additionally requires $m \gg \theta^2 b S_\infty^2$. Thus the general comparison keeps two conditions,

$$m \gg S_\infty, \qquad m \gg \theta^2 b S_\infty^2.$$

Together they imply relative MSE tending to one. For fixed $\theta \neq 0$ the second condition implies the first; for $\theta = 1$, Theorem B.4 also proves necessity for risk equivalence.

The three method-dependent scales are summarized in the table of Section 4.

At fixed observation budget $N = mb$, both constraints must still hold: $b \ll N/S_\infty$ for covariance and $b \ll \sqrt{N}/(|\theta| S_\infty)$ for a nonzero start. The second is absent when $\theta = 0$. A batch proportional to its critical scale can preserve a nonzero normalized bias and hence a coverage deficit (Theorem B.5).

The distributional comparison has one more component. Theorem B.7 gives the same $O(N^{-1/2})$ bound for all three methods after centering and variance normalization, under a bounded standardized third moment. With the PR variance as target, the extra error is of order S_∞/m, and Theorem E.3 shows that this order is sharp even for Gaussian noise. Momentum changes the center and covariance transients, without improving the sample-size exponent of this shape bound. These conclusions concern the μ eigendirection and the stated full averages. The next subsection tests the spectral extrapolation; Section E.3 then changes the averaging window and charges all discarded rounds to the budget.

B.3.2 Does the Gain Extend across the Spectrum?

The criteria of Theorems B.3 and B.4 describe the soft mode under the stated parameters and start. The following results limit the transfer of this gain to the whole spectrum. The least-curved mode does not control every relative directional inference error. This section gives a counterexample.

Theorem B.9 (Heavy-ball covariance inflation on the stiff mode)**.** *Let $\varkappa = \sqrt{\kappa} > 1$ and use the optimally tuned quadratic heavy-ball parameters from Section B.1. On the eigenmode $\lambda = L$, define $\varrho = (\varkappa - 1)/(\varkappa + 1)$. Its zero-velocity deterministic residual is*

$$r_j^{(L)} = (-\varrho)^j \{1 + (1+\varrho) j\}.$$

If $S_m^{(L)} = \sum_{j<m} r_j^{(L)}$ *and* $Q_m^{(L)} = \sum_{j<m} (r_j^{(L)})^2$*, then*

$$\frac{bmL^2}{\sigma_L^2}\operatorname{Var}(\bar{u}_{0,m}^{(L)}) = 1 + \frac{Q_m^{(L)} - 2S_m^{(L)}}{m}, \qquad S_\infty^{(L)} = \frac{1}{\varkappa}, \qquad Q_\infty^{(L)} = \frac{\varkappa^4 + 2\varkappa^2 + 5}{8\varkappa}. \tag{47}$$

In a joint limit $\varkappa \to \infty$ *and* $m/\varkappa \to \infty$*,*

$$\frac{bmL^2}{\sigma_L^2}\operatorname{Var}(\bar{u}_{0,m}^{(L)}) = 1 + \frac{\varkappa^3}{8m}\{1 + o(1)\}. \tag{48}$$

Thus, in this regime, relative covariance efficiency is equivalent to $m/\kappa^{3/2} \to \infty$*. If* $m/\varkappa^3 \to \tau_{\text{stiff}} \in (0, \infty)$ *and the noise is Gaussian, the centered statistic* $L\sqrt{bm}(\bar{u}_{0,m}^{(L)} - \mathbb{E}\bar{u}_{0,m}^{(L)})/\sigma_L$ *has limiting law* $\mathcal{N}(0, 1 + 1/(8\tau_{\text{stiff}}))$*. In contrast, SGD with* $\alpha = 1/L$ *on this same mode has relative variance* $1 - 1/m$ *.*

Proof of Theorem B.9. On the L-mode of HB the characteristic polynomial has the repeated root $-\varrho$, where $\varrho = (\varkappa - 1)/(\varkappa + 1)$. The zero-velocity initial conditions give

$$r_j^{(L)} = (-\varrho)^j\{1 + (1 + \varrho)j\}.$$

Using $\sum z^j = (1 - z)^{-1}$, $\sum jz^j = z(1 - z)^{-2}$, and $\sum j^2 z^j = z(1 + z)(1 - z)^{-3}$ gives

$$S_\infty^{(L)} = \frac{1}{\varkappa}, \qquad Q_\infty^{(L)} = \frac{\varkappa^4 + 2\varkappa^2 + 5}{8\varkappa}.$$

The scalar law yields

$$\frac{bmL^2}{\sigma_L^2}\operatorname{Var}(\bar{u}_{0,m}^{(L)}) = \frac{1}{m}\sum_{j=0}^{m-1}(1 - r_j^{(L)})^2 = 1 + \frac{Q_m^{(L)} - 2S_m^{(L)}}{m}.$$

If $\varkappa \to \infty$ and $m/\varkappa \to \infty$, the geometric tails of $S_m^{(L)}$ and $Q_m^{(L)}$ are negligible relative to $\varkappa^3$, while $Q_\infty^{(L)} = \varkappa^3\{1 + o(1)\}/8$ and $S_\infty^{(L)} = 1/\varkappa$. This proves Equation (48) and the equivalence with $m/\kappa^{3/2} \to \infty$ in the stated regime. Under Gaussian noise the centered statistic is Gaussian for every $m, \varkappa$; if $m/\varkappa^3 \to \tau_{\text{stiff}}$, its variance tends to $1 + 1/(8\tau_{\text{stiff}})$, proving the final claim. For SGD on the L-mode, $r_0^{(L)} = 1$ and $r_j^{(L)} = 0$ for $j \geq 1$, so the exact relative variance is $(m - 1)/m$. □

What This Rules Out. For example, take $m = \varkappa^2 = \kappa$ and $b = 1$. The stiff-mode HB variance ratio diverges as $\varkappa/8$, whereas the SGD ratio tends to one. The least-curved HB mode at this horizon already has centered variance approaching its PR limit. Thus acceleration of the slow mode cannot imply uniform relative directional covariance acceleration over the spectrum. A Frobenius error weighted by $H^{-1}\Sigma H^{-1}$ can hide the stiff mode, because its limiting variance is smaller by a factor κ^2 when the gradient-noise variances are equal. General noise-amplification tradeoffs (Mohammadi et al., 2021, 2025) and the sign of averaged HB modal corrections (Samuelson and Jovanović, 2024, Theorem 3) are known; the precise averaged-output formula and its relation to the onset criteria above are the point of this counterexample.

Nesterov Covariance Calibration on Every Hessian Eigendirection The stiff-mode obstruction above is not shared by the NAG tuning. The following upper bound complements it with a full-spectrum comparison.

Theorem B.10 (Uniform eigendirection bound for NAG)**.** *Let* $\varkappa = \sqrt{\kappa} \geq 1$*,* $\alpha = 1/L$*,* $\beta = \gamma = (\varkappa - 1)/(\varkappa + 1)$*, and use deterministic zero-velocity initialization. On any eigenvalue* $\lambda \in [\mu, L]$*, let the gradient-noise variance be* $\sigma_\lambda^2 > 0$ *and define*

$$v_m(\lambda) = \frac{bm\lambda^2}{\sigma_\lambda^2}\operatorname{Var}(\bar{u}_{0,m}^{(\lambda)}).$$

Then, for every $m, b \geq 1$*,*

$$\sup_{\lambda \in [\mu, L]} |v_m(\lambda) - 1| \leq \frac{30\sqrt{\kappa}}{m}. \tag{49}$$

For Gaussian noise this also bounds the centered Wasserstein distance to the PR-normalized standard Gaussian by $30\sqrt{2/\pi}\sqrt{\kappa}/m$ *on every eigendirection. The order* $\sqrt{\kappa}/m$ *is sharp: when* $\kappa \to \infty$ *and* $m/\sqrt{\kappa} \to \infty$*, the slow mode obeys*

$$1 - v_m(\mu) \sim \frac{11}{4}\frac{\sqrt{\kappa}}{m}.$$

Thus the worst eigendirection covariance defect for NAG is $\Theta(\sqrt{\kappa}/m)$ *in this regime. For optimally tuned HB the stiff mode alone has asymptotic to* $\kappa^{3/2}/(8m)$ *in the same regime.*

Proof of Theorem B.10. Fix an eigenvalue and put $\lambda_{\mathrm{rel}} = \lambda/L$ and $\beta_\varkappa = (\varkappa-1)/(\varkappa+1)$. For $\varkappa = 1$ the claim reduces to the preceding SGD calculation, so assume $\varkappa > 1$. The deterministic NAG query response obeys

$$r_{j+2}^{(\lambda)} = \mathfrak{a}_\lambda r_{j+1}^{(\lambda)} - \mathfrak{d}_\lambda r_j^{(\lambda)}, \qquad \mathfrak{a}_\lambda = (1+\beta_\varkappa)(1-\lambda_{\mathrm{rel}}), \quad \mathfrak{d}_\lambda = \beta_\varkappa(1-\lambda_{\mathrm{rel}}),$$

with $r_0^{(\lambda)} = 1$ and $r_1^{(\lambda)} = 1-(1+\beta_\varkappa)\lambda_{\mathrm{rel}}$. At $\lambda_{\mathrm{rel}} = \varkappa^{-2}$ the roots coincide at $\varrho_\mu = 1-\varkappa^{-1}$. For $\lambda_{\mathrm{rel}} > \varkappa^{-2}$ they are $\rho e^{\pm i\varphi}$, where $\rho = \sqrt{\beta_\varkappa(1-\lambda_{\mathrm{rel}})} \le \varrho_\mu$. The repeated-root case follows by continuity. If $\lambda_{\mathrm{rel}} = 1$, then $r_0^{(\lambda)} = 1$, $r_1^{(\lambda)} = -\beta_\varkappa$, and $r_j^{(\lambda)} = 0$ for $j \ge 2$; hence $\sum_j |r_j^{(\lambda)}| \le 2$ and $\sum_j (r_j^{(\lambda)})^2 \le 2$, so the uniform estimates below hold directly. For their derivation, assume $\lambda_{\mathrm{rel}} < 1$. Writing the solution in trigonometric form and using $|\sin(j\varphi)| \le j|\sin\varphi|$ gives

$$|r_j^{(\lambda)}| \le \rho^j(1+c_\lambda j), \qquad c_\lambda\rho = \left|r_1^{(\lambda)} - \frac{\mathfrak{a}_\lambda}{2}\right| = \frac{1}{2}\left|\frac{1+\beta_\varkappa}{\beta_\varkappa}\rho^2 - 2\beta_\varkappa\right|. \tag{50}$$

On $0 \le \rho \le \varrho_\mu$, the convex quadratic

$$\frac{1+\beta_\varkappa}{\beta_\varkappa}\rho^2 - 4\rho + 4 - 2\beta_\varkappa$$

has its minimum at ϱ_μ, where it equals $2(1-\beta_\varkappa)(\beta_\varkappa+2)/(1+\beta_\varkappa) \ge 0$. The concave quadratic

$$4 - 4\rho + 2\beta_\varkappa - \frac{1+\beta_\varkappa}{\beta_\varkappa}\rho^2$$

has its minimum over the interval at an endpoint; its values at 0 and ϱ_μ are $4+2\beta_\varkappa$ and $2(1-\beta_\varkappa)(2-\beta_\varkappa)/(1+\beta_\varkappa)$, respectively. Together these two inequalities give $c_\lambda\rho \le 2(1-\rho)$. Define $\mathcal{S}_m(\lambda) := \sum_{j=0}^{m-1} r_j^{(\lambda)}$ and $\mathcal{Q}_m(\lambda) := \sum_{j=0}^{m-1}(r_j^{(\lambda)})^2$, with the subscript ∞ denoting the corresponding infinite sums. Summing Equation (50) therefore gives

$$\sum_{j\ge0}|r_j^{(\lambda)}| \le \frac{1}{1-\rho} + \frac{c_\lambda\rho}{(1-\rho)^2} \le \frac{3}{1-\rho} \le 3\varkappa, \quad \sum_{j\ge0}(r_j^{(\lambda)})^2 \le \frac{2}{1-\rho^2} + \frac{2c_\lambda^2\rho^2(1+\rho^2)}{(1-\rho^2)^3} \le 18\varkappa.$$

For every finite m, the scalar law thus yields

$$|v_m(\lambda) - 1| = \frac{|\mathcal{Q}_m(\lambda) - 2\mathcal{S}_m(\lambda)|}{m} \le \frac{\sum_j (r_j^{(\lambda)})^2 + 2\sum_j |r_j^{(\lambda)}|}{m} \le \frac{24\varkappa}{m} \le \frac{30\varkappa}{m}.$$

For Gaussian noise, $d_{\mathrm{W}}(\mathcal{N}(0,v_m),\mathcal{N}(0,1)) = \sqrt{2/\pi}|\sqrt{v_m}-1| \le \sqrt{2/\pi}|v_m-1|$, which gives the displayed bound.

On the slow mode, $\varrho_\mu = 1-\varkappa^{-1}$, $c = (\varkappa+1)^{-1}$, and $r_j^{(\mu)} = (1+cj)\varrho_\mu^j$. Hence

$$\mathcal{S}_\infty(\mu) = \frac{2\varkappa^2}{\varkappa+1}, \qquad \mathcal{Q}_\infty(\mu) = \frac{1}{1-\varrho_\mu^2} + \frac{2c\varrho_\mu^2}{(1-\varrho_\mu^2)^2} + \frac{c^2\varrho_\mu^2(1+\varrho_\mu^2)}{(1-\varrho_\mu^2)^3}.$$

As $\varkappa \to \infty$, $\mathcal{S}_\infty(\mu) \sim 2\varkappa$, $\mathcal{Q}_\infty(\mu) \sim 5\varkappa/4$, and therefore $2\mathcal{S}_\infty(\mu) - \mathcal{Q}_\infty(\mu) \sim 11\varkappa/4$. If $m/\varkappa \to \infty$, the geometric tails are $o(\varkappa)$, so $1 - v_m(\mu) = \{2\mathcal{S}_m(\mu) - \mathcal{Q}_m(\mu)\}/m \sim 11\varkappa/(4m)$. □

Limits of the Full-Spectrum Statement. The theorem controls each Hessian eigendirection. If H and Σ commute, any directional relative variance is a convex combination of these modal ratios, so the same bound holds in every direction with positive PR variance. Without this commutation, the matrix formula still holds, but a relative bound for arbitrary directions needs control of the noise covariance geometry.

C Beyond Quadratics: First-Order ERM Equivalence for Averaged NAG

The quadratic theory identifies a method-dependent endpoint term as the gap between an averaged online output and the ERM computed from the same data. For smooth strongly convex objectives with additive noise, the corresponding gap acquires a second-order trajectory remainder. We first derive this remainder for an arbitrary online procedure and then control it for tail-averaged constant-parameter NAG, including the observations spent during burn-in in the offline comparison.

C.1 Online–Offline Bridge Beyond Quadratics

We first return to the bridge of Theorem A.2 for a nonquadratic model that remains stochastically linear: the oracle is $G(x,\zeta) = \nabla f(x) + \xi(\zeta)$, where $f \in C^2$ is μ-strongly convex and its Hessian is M_2-Lipschitz.

Proposition C.1 (Trajectory term of the online–offline gap)**.** *Let an arbitrary online procedure query $s_t, \dots, s_{t+m-1}$, with s_k being $\mathcal{F}_k$-measurable, and receive $g_k = \nabla f(s_k) + \eta_{k+1}^{(b)}$. Let $\widehat{x}$ minimize the empirical risk $f(x) + \left\langle \bar{\eta}_{t,m}^{(b)}, x \right\rangle$ of the same $N = mb$ observations, and put $H_{\widehat{x}} = \nabla^2 f(\widehat{x})$, $\bar{s} = m^{-1}\sum_{k=t}^{t+m-1} s_k$ and $\bar{g} = m^{-1}\sum_{k=t}^{t+m-1} g_k$. Then, pathwise,*

$$H_{\widehat{x}}(\bar{s} - \widehat{x}) = \bar{g} - \bar{\rho}, \qquad \bar{\rho} = \frac{1}{m}\sum_{k=t}^{t+m-1}\big[\nabla f(s_k) - \nabla f(\widehat{x}) - H_{\widehat{x}}(s_k - \widehat{x})\big]. \tag{51}$$

$$\big\|\bar{s} - H_{\widehat{x}}^{-1}\bar{g} - \widehat{x}\big\| \le \frac{M_2}{2\mu}\,\frac{1}{m}\sum_{k=t}^{t+m-1}\|s_k - \widehat{x}\|^2. \tag{52}$$

Indeed, $\nabla f(\widehat{x}) = -\bar{\eta}_{t,m}^{(b)}$, hence $\bar{g} = m^{-1}\sum_k\{\nabla f(s_k) - \nabla f(\widehat{x})\}$, which is Equation (51); each bracket in $\bar{\rho}$ has norm at most $M_2\|s_k - \widehat{x}\|^2/2$, and $H_{\widehat{x}} \succeq \mu I$.

In the quadratic model $M_2 = 0$ and Theorem C.1 reduces to Theorem A.2. For a nonquadratic model, the gap between the online average and the offline estimator contains, besides the bridge correction $H_{\widehat{x}}^{-1}\bar{g}$, a trajectory term of second order in the spread of the queried points around the ERM. Beyond the bridge correction, momentum and step policies enter the gap only through this spread. For constant parameters the spread is governed by the stationary second moments of the queries, that is, by the noise amplification of the method (Mohammadi et al., 2021), and the trajectory term is the online–offline counterpart of the stationary bias of constant-step SGD (Dieuleveut et al., 2020). Theorem C.2 below supplies a quantitative control for constant-gain NAG under global curvature assumptions.

C.2 Full-Budget ERM Equivalence for Averaged NAG

We now close the trajectory transfer for a nonlinear accelerated method. The output is an average of the queried points, as in Theorem C.1. Both the burn-in and the averaging window use the classical constant NAG parameters. The Lyapunov stability mechanism is standard; see, for example, Laborde and Oberman (2020, Section 5). Here it is combined with the trajectory identity and a comparison with the ERM on *all* observations.

Theorem C.2 (Nonlinear NAG: verified moments and a full-budget ERM bound)**.** *Let $f \in C^2(\mathbb{R}^d)$ satisfy, globally,*

$$0 < \mu < L, \qquad \mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \|\nabla^2 f(x) - \nabla^2 f(y)\| \le M_2\|x - y\|.$$

Let $\xi_{k+1,j}$ be iid centered vectors with covariance Σ and $0 < \sigma^2 = \operatorname{tr}\Sigma < \infty$. For integers $b, m \ge 1$ and $t \ge 0$, set $N = (t+m)b$ and, for $0 \le k < t+m$, run from a deterministic $x_{-1} = x_0$,

$$\upsilon = \sqrt{\mu/L} = \kappa^{-1/2}, \qquad \beta = \frac{1-\upsilon}{1+\upsilon}, \qquad \eta_{k+1}^{(b)} = \frac{1}{b}\sum_{j=1}^{b}\xi_{k+1,j},$$
$$s_k = x_k + \beta(x_k - x_{k-1}), \qquad x_{k+1} = s_k - \frac{1}{L}\nabla f(s_k) - \frac{1}{L}\eta_{k+1}^{(b)}. \tag{53}$$

Retain $\bar{s}_{t,m} = m^{-1}\sum_{k=t}^{t+m-1} s_k$. For $\ell(x,\xi) = f(x) + \langle \xi, x\rangle$, define the ERM on the same N observations by

$$\widehat{x}_N^{\mathrm{ERM}} = \arg\min_x\{f(x) + \langle \bar{\eta}_N, x\rangle\}, \qquad \bar{\eta}_N = \frac{1}{t+m}\sum_{k=0}^{t+m-1}\eta_{k+1}^{(b)}.$$

Write

$$\psi_k = x_k + \frac{1-\upsilon}{\upsilon}(x_k - x_{k-1}), \qquad \mathcal{V}_k = f(x_k) - f(x_\star) + \frac{\mu}{2}\|\psi_k - x_\star\|^2. \tag{54}$$

Then the natural filtration satisfies

$$\begin{aligned}
&\mathbb{E}[\mathcal{V}_{k+1} \mid \mathcal{F}_k] \le (1-\upsilon)\mathcal{V}_k + \frac{\sigma^2}{Lb}, \qquad \mathbb{E}\mathcal{V}_k \le (1-\upsilon)^k\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon},\\
&\mathbb{E}\|s_k - x_\star\|^2 \le \frac{2}{\mu}\left\{(1-\upsilon)^k\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}\right\}.
\end{aligned} \tag{55}$$

If $(1-\upsilon)^t\mathcal{V}_0 \le \sigma^2/(bL\upsilon)$*, then*

$$\frac{\mu\sqrt{N}}{\sigma}\,\mathbb{E}\|\bar{s}_{t,m} - \widehat{x}_N^{\mathrm{ERM}}\| \le \sqrt{1+\frac{t}{m}}\left\{\frac{4\sqrt{2}}{(1+\upsilon)\sqrt{m\upsilon}} + \frac{M_2\sigma}{\mu L\upsilon}\sqrt{\frac{m}{b}}\left(2+\frac{3}{m\upsilon}\right)\right\} + \sqrt{\frac{t}{m}}. \tag{56}$$

In particular, a sufficient burn-in is

$$t = \left\lceil \frac{1}{\upsilon}\log\left(1+\frac{bL\upsilon\,\mathcal{V}_0}{\sigma^2}\right)\right\rceil. \tag{57}$$

Any known upper bound on $\mathcal{V}_0$ *may replace it in this choice.*

Proof of Theorem C.2. All norms below are Euclidean, and matrix norms are their induced operator norms. Let $\mathcal{F}_k$ contain the start and the first k batches. The next batch is independent of $\mathcal{F}_k$, with conditional mean zero and conditional squared norm σ^2/b.

Global Stability with the Accelerated Parameters. The definitions in Equations (53) and (54) give the identities

$$s_k = \frac{x_k + \upsilon\psi_k}{1+\upsilon}, \qquad \psi_{k+1} = (1-\upsilon)\psi_k + \upsilon s_k - \frac{\upsilon}{\mu}\{\nabla f(s_k) + \eta_{k+1}^{(b)}\}. \tag{58}$$

First suppress the noise in one step, and denote the resulting states by x^+ and ψ^+. With $g_s = \nabla f(s)$ and $\omega = (1-\upsilon)\psi + \upsilon s$, L-smoothness and $\upsilon^2/\mu = 1/L$ imply

$$\begin{aligned} f(x^+) - f(x_\star) + \frac{\mu}{2}\|\psi^+ - x_\star\|^2 &\le f(s) - f(x_\star) - \frac{\|g_s\|^2}{2L} + \frac{\mu}{2}\left\|\omega - x_\star - \frac{\upsilon}{\mu}g_s\right\|^2 \\ &= f(s) - f(x_\star) + \frac{\mu}{2}\|\omega - x_\star\|^2 - \upsilon\langle \omega - x_\star, g_s\rangle. \end{aligned}$$

Use strong convexity at s, first with x and then with $x_\star$, with weights $1-\upsilon$ and υ. The linear terms cancel because $(1+\upsilon)s = x + \upsilon\psi$. Expanding the squared norm of $\omega - x_\star$ yields

$$f(x^+) - f(x_\star) + \frac{\mu}{2}\|\psi^+ - x_\star\|^2 \le (1-\upsilon)\mathcal{V} - \frac{\mu(1-\upsilon)}{2}\{\|x-s\|^2 + \upsilon\|\psi - s\|^2\} \le (1-\upsilon)\mathcal{V}. \tag{59}$$

Restoring the noise changes x^+ by $-\eta_{k+1}^{(b)}/L$ and ψ^+ by $-\upsilon\eta_{k+1}^{(b)}/\mu$. Conditional expectation eliminates the linear terms. Smoothness bounds the first additional contribution by $\sigma^2/(2Lb)$; the second is $\upsilon^2\sigma^2/(2\mu b) = \sigma^2/(2Lb)$. This proves the first inequality in Equation (55). Iteration gives $\mathbb{E}\mathcal{V}_k \le (1-\upsilon)^k\mathcal{V}_0 + \sigma^2/(bL\upsilon)$.

Strong convexity also gives, pathwise,

$$\mathcal{V}_k \ge \frac{\mu}{2}\{\|x_k - x_\star\|^2 + \|\psi_k - x_\star\|^2\}.$$

Since $s_k = (x_k + \upsilon\psi_k)/(1+\upsilon)$, it follows that

$$\mathbb{E}\|s_k - x_\star\|^2 \le \frac{2}{\mu}\left\{(1-\upsilon)^k\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}\right\}.$$

In particular,

$$\frac{1}{m}\sum_{k=t}^{t+m-1}\mathbb{E}\|s_k - x_\star\|^2 \le \frac{2}{\mu}\left\{\frac{(1-\upsilon)^t\mathcal{V}_0}{m\upsilon} + \frac{\sigma^2}{bL\upsilon}\right\}. \tag{60}$$

Endpoint Control and the Retained-Window ERM. Set $\Delta x_k = x_k - x_{k-1}$ and use the corrected state

$$x_k^{\mathrm{corr}} = x_k + \frac{\beta}{1-\beta}\Delta x_k = \frac{x_k + \psi_k}{2}.$$

Then

$$\mathbb{E}\|x_k^{\mathrm{corr}} - x_\star\|^2 \le \frac{1}{\mu}\left\{(1-\upsilon)^k\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}\right\},$$

and the recursion gives

$$x_{k+1}^{\mathrm{corr}} - x_k^{\mathrm{corr}} = -\frac{\nabla f(s_k) + \eta_{k+1}^{(b)}}{L(1-\beta)}.$$

Consequently, for the observed average gradient on the retained window,

$$\bar{g}_{t,m} = \frac{1}{m}\sum_{k=t}^{t+m-1}\{\nabla f(s_k) + \eta_{k+1}^{(b)}\} = \frac{2L\upsilon}{(1+\upsilon)m}(x_t^{\mathrm{corr}} - x_{t+m}^{\mathrm{corr}}), \tag{61}$$

and

$$\mathbb{E}\|\bar{g}_{t,m}\| \le \frac{2L\upsilon}{(1+\upsilon)m\sqrt{\mu}}\left[\sqrt{(1-\upsilon)^t\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}} + \sqrt{(1-\upsilon)^{t+m}\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}}\right].$$

Let $\bar{\eta}_{t,m} = m^{-1}\sum_{k=t}^{t+m-1}\eta_{k+1}^{(b)}$ and $\widehat{x}_{t,m}^{\mathrm{ERM}} = \arg\min_x\{f(x) + \langle\bar{\eta}_{t,m}, x\rangle\}$. Strong monotonicity and independence of the noises imply

$$\|\widehat{x}_{t,m}^{\mathrm{ERM}} - x_\star\| \le \frac{\|\bar{\eta}_{t,m}\|}{\mu}, \qquad \mathbb{E}\|\widehat{x}_{t,m}^{\mathrm{ERM}} - x_\star\|^2 \le \frac{\sigma^2}{\mu^2 bm}. \tag{62}$$

Apply Theorem C.1 with this ERM. Since its Hessian is bounded below by μI, and $\|s_k - \widehat{x}_{t,m}^{\mathrm{ERM}}\|^2 \le 2\|s_k - x_\star\|^2 + 2\|\widehat{x}_{t,m}^{\mathrm{ERM}} - x_\star\|^2$, we obtain

$$\begin{aligned}\mathbb{E}\|\bar{s}_{t,m} - \widehat{x}_{t,m}^{\mathrm{ERM}}\| \le{}& \frac{2L\upsilon}{(1+\upsilon)\mu^{3/2}m}\left[\sqrt{(1-\upsilon)^t\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}} + \sqrt{(1-\upsilon)^{t+m}\mathcal{V}_0 + \frac{\sigma^2}{bL\upsilon}}\right] \\ &+ \frac{2M_2}{\mu^2}\left\{\frac{(1-\upsilon)^t\mathcal{V}_0}{m\upsilon} + \frac{\sigma^2}{bL\upsilon}\right\} + \frac{M_2\sigma^2}{\mu^3 bm}.\end{aligned} \tag{63}$$

If $(1-\upsilon)^t\mathcal{V}_0 \le \sigma^2/(bL\upsilon)$, each quantity under a square root is at most $2\sigma^2/(bL\upsilon)$. Multiply Equation (63) by $\mu\sqrt{mb}/\sigma$ and use $\mu = L\upsilon^2$ to get

$$\frac{\mu\sqrt{mb}}{\sigma}\mathbb{E}\|\bar{s}_{t,m} - \widehat{x}_{t,m}^{\mathrm{ERM}}\| \le \frac{4\sqrt{2}}{(1+\upsilon)\sqrt{m\upsilon}} + \frac{M_2\sigma}{\mu L\upsilon}\sqrt{\frac{m}{b}}\left(2 + \frac{3}{m\upsilon}\right). \tag{64}$$

Charging the Burn-In Observations. The map $u \mapsto \arg\min_x\{f(x) + \langle u, x\rangle\}$ is $1/\mu$-Lipschitz by strong monotonicity. Using the same noise batches for both ERMs gives the second-moment identity

$$\mathbb{E}\|\bar{\eta}_{t,m} - \bar{\eta}_N\|^2 = \frac{\sigma^2 t}{bm(t+m)}.$$

Thus

$$\mathbb{E}\|\widehat{x}_{t,m}^{\mathrm{ERM}} - \widehat{x}_N^{\mathrm{ERM}}\| \le \frac{\sigma}{\mu}\sqrt{\frac{t}{bm(t+m)}}. \tag{65}$$

For $t = 0$, both differences are zero. Combining Equations (64) and (65) and $N = (t+m)b$ proves Equation (56). Finally, $(1-\upsilon)^t \le e^{-\upsilon t}$ proves that Equation (57) ensures $(1-\upsilon)^t\mathcal{V}_0 \le \sigma^2/(bL\upsilon)$. The upper bound $\mathcal{V}_0 \le (L+\mu)\|x_0 - x_\star\|^2/2$ follows from smoothness and $\psi_0 = x_0$. □

The three terms in Equation (56) have different roles. The boundary term becomes small for $m \gg \sqrt{\kappa}$ after burn-in. The term proportional to $M_2\sigma/(\mu L\upsilon)$ pays for nonlinear trajectory spread. The final $\sqrt{t/m}$ compares the retained-window ERM with the ERM using also the tb burn-in observations. Thus initialization is neither free nor an assumed warm start. The burn-in has the accelerated contraction scale $\upsilon^{-1} = \sqrt{\kappa}$. A bound $\|x_0 - x_\star\| \le R_{\mathrm{init}}$ gives the bound $\mathcal{V}_0 \le (L+\mu)R_{\mathrm{init}}^2/2$ in Equation (57).

Corollary C.3 (Nonlinear first-order equivalence on all observations)**.** *Fix f, the noise law, the dimension and x_0 under the assumptions of Theorem C.2. Along any sequence with*

$$m \longrightarrow \infty, \qquad \frac{m}{b} \longrightarrow 0, \qquad \log b = o(m),$$

choose t by Equation (57)*. Then, with $H = \nabla^2 f(x_\star)$ and $N = (t+m)b$,*

$$\sqrt{N}(\bar{s}_{t,m} - \widehat{x}_N^{\mathrm{ERM}}) \xrightarrow{L_1} 0, \qquad \sqrt{N}(\bar{s}_{t,m} - x_\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, H^{-1}\Sigma H^{-1}). \tag{66}$$

More, there is a remainder $R_{t,m,b}$ such that

$$\begin{aligned}\sqrt{N}(\bar{s}_{t,m} - x_\star) &= -\frac{H^{-1}}{\sqrt{N}}\sum_{k=0}^{t+m-1}\sum_{j=1}^{b}\xi_{k+1,j} + R_{t,m,b}, \\ \mathbb{E}\|R_{t,m,b}\| &\le \frac{\sigma}{\mu}\sqrt{1 + \frac{t}{m}}\left\{\frac{4\sqrt{2}}{(1+\upsilon)\sqrt{m\upsilon}} + \frac{M_2\sigma}{\mu L\upsilon}\sqrt{\frac{m}{b}}\left(2 + \frac{3}{m\upsilon}\right)\right\} + \frac{\sigma}{\mu}\sqrt{\frac{t}{m}} + \frac{M_2\sigma^2}{2\mu^3\sqrt{N}}.\end{aligned} \tag{67}$$

Proof of Theorem C.3. For a fixed problem the prescribed burn-in is $O(1 + \log b)$. Hence $t/m \to 0$, $mv \to \infty$, and $\{M_2\sigma/(\mu L\upsilon)\}\sqrt{m/b} \to 0$. Thus the right-hand side of Equation (56) tends to zero. The first conclusion in Equation (66) follows directly from Equation (56).

To identify the limiting law, strong monotonicity gives $\|\widehat{x}_N^{\mathrm{ERM}} - x_\star\| \le \|\bar{\eta}_N\|/\mu$. Taylor expansion at $x_\star$, using $H \succeq \mu I$ and the global Lipschitz Hessian, gives

$$\|\widehat{x}_N^{\mathrm{ERM}} - x_\star + H^{-1}\bar{\eta}_N\| \le \frac{M_2}{2\mu}\|\widehat{x}_N^{\mathrm{ERM}} - x_\star\|^2 \le \frac{M_2}{2\mu^3}\|\bar{\eta}_N\|^2.$$

Since $\mathbb{E}\|\bar{\eta}_N\|^2 = \sigma^2/N$, this proves Equation (67). The usual iid vector CLT needs only the finite second moment assumed here, and completes the proof. □

For example, $b = m^2$ and Equation (57) give an admissible sequence, with $t = O(\log m)$ and $N \asymp m^3$. A statistical example is the regularized Bernoulli logistic loss

$$\ell(x, Y) = \frac{\lambda_{\mathrm{reg}}}{2}x^2 + \log(1 + e^x) - Yx, \qquad Y \sim \mathrm{Bernoulli}(\pi_{\mathrm{Ber}}), \qquad \lambda_{\mathrm{reg}} > 0, \quad 0 < \pi_{\mathrm{Ber}} < 1. \tag{68}$$

Here $f = \mathbb{E}\ell$, the additive gradient noise is $\xi = \pi_{\mathrm{Ber}} - Y$, and $\mu = \lambda_{\mathrm{reg}}$, $L = \lambda_{\mathrm{reg}} + 1/4$, $M_2 = 1/(6\sqrt{3})$ and $\sigma^2 = \pi_{\mathrm{Ber}}(1 - \pi_{\mathrm{Ber}})$ are valid global constants. Indeed, writing $\pi_x = (1 + e^{-x})^{-1}$ gives $f''(x) = \lambda_{\mathrm{reg}} + \pi_x(1 - \pi_x)$ and $|f'''(x)| = \pi_x(1 - \pi_x)|1 - 2\pi_x| \le 1/(6\sqrt{3})$. Thus $\widehat{x}_N^{\mathrm{ERM}}$ is the usual regularized empirical minimizer for the observed Bernoulli sample, and $x_\star$ is its population counterpart.

In arbitrary dimension the class also includes, for any positive-definite matrix H_0,

$$f(x) = \tfrac{1}{2}x^\top H_0 x + \varepsilon \sum_{i=1}^{d}(\sin x_i - x_i), \qquad 0 < \varepsilon < \lambda_{\min}(H_0). \tag{69}$$

Here $x_\star = 0$, $\mu = \lambda_{\min}(H_0) - \varepsilon$, $L = \lambda_{\max}(H_0) + \varepsilon$, and $M_2 = \varepsilon$ are valid global constants. A complete proof is in Section C.2.

D Additional Quadratic Identities and Risk Comparisons

Comparable guarantees refer to the same output of the procedure, the same initialization, the same available information and the same total budget. The main criteria are the mean squared error and the confidence ball:

$$\mathcal{R}_N = \mathbb{E}\,\|\hat{x}_N - x_\star\|^2, \qquad \mathbb{P}\{\|\hat{x}_N - x_\star\| \le r_N(\delta)\} \ge 1 - \delta. \tag{70}$$

Individual directions and ellipsoids provide additional information, but they do not replace the comparison in the chosen common metric.

We distinguish the following targets. They use different normalizations and therefore favor different parameter policies.

Table 3: Four Non-Equivalent Meanings of Convergence

Target	**Typical statement**	**What it resolves / what it hides**
Optimization transient	$\|x_k - x_\star\|^2 \le \rho^{2k}\|x_0 - x_\star\|^2 + \cdots$	Burn-in and conditioning; not the law of the output.
Finite-budget risk	$\mathbb{E}\,\|\hat{x}_N - x_\star\|^2 \le c_0/N + \varepsilon_N$	Valid at every N; constants and directional uncertainty may be hidden.
Distributional limit	$\sqrt{N}(\hat{x}_N - x_\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V)$	local geometry and inference; says little about how large N must be.
Offline expansion	$\widehat{x}_N^{\mathrm{ERM}} - x_\star = -[\nabla^2 f(x_\star)]^{-1}\bar{\eta} + \text{remainder}$	Law of the offline estimator on the same data, $\bar{\eta}$ being the empirical noise mean; says nothing about how an online method reaches it.

D.1 One SGD Output: Two Nonasymptotic Guarantees and the Risk

Fix $\beta = \gamma = 0$, $0 < \alpha \le 1/L$, a deterministic start x_0 with $\|x_0 - x_\star\| \le R_0$, and the output $\widehat{x} = m^{-1}\sum_{k=0}^{m-1} x_k$. Let $N = mb$, and set

$$\sigma^2 := \operatorname{tr}\Sigma, \quad F_\alpha := I - \alpha H, \quad \rho_\mu := 1 - \alpha\mu, \quad \mathsf{S}_{\alpha,m} := \sum_{j=0}^{m-1} F_\alpha^j, \quad \omega_m(\rho_\mu) := \sum_{j=0}^{m-1}(1-\rho_\mu^j)^2.$$

Proposition D.1 (Comparable guarantees for the full average)**.** *In the additive quadratic model, we get*

$$\mathcal{R}_{m,b} = \|\mathsf{S}_{\alpha,m} e_0/m\|^2 + \operatorname{tr} W_m, \tag{71}$$

$$W_m := \frac{1}{bm^2}\sum_{j=0}^{m-1}(I - F_\alpha^j)V_\star(I - F_\alpha^j). \tag{72}$$

It satisfies

$$\mathcal{R}_{m,b} \le U_{\mathrm{en}} := \frac{R_0^2}{\alpha\mu m} + \frac{\alpha\sigma^2}{\mu b}, \tag{73}$$

$$\mathcal{R}_{m,b} \le U_{\mathrm{con}} := R_0^2\left(\frac{1-\rho_\mu^m}{\alpha\mu m}\right)^2 + \frac{\sigma^2\omega_m(\rho_\mu)}{b\mu^2 m^2} \le \frac{R_0^2}{\alpha^2\mu^2 m^2} + \frac{\sigma^2}{\mu^2 N}. \tag{74}$$

Bound Equation (73) *also holds for a general smooth strongly convex f with the same μ, L and conditional noise variance at most σ^2/b. For either $U \in \{U_{\mathrm{en}}, U_{\mathrm{con}}\}$, the radius $r = \sqrt{U/\delta}$ guarantees $\mathbb{P}\{\|\widehat{x} - x_\star\| \le r\} \ge 1 - \delta$.*

Proof of Theorem D.1. Unrolling SGD and summing its pre-update iterates gives

$$\widehat{x} - x_\star = \frac{\mathsf{S}_{\alpha,m} e_0}{m} - \frac{1}{m}\sum_{s=1}^{m}(I - F_\alpha^{m-s})H^{-1}\eta_s^{(b)}.$$

The last summand is zero. Independence, centering, and $\operatorname{Cov}(\eta_s^{(b)}) = \Sigma/b$ prove Equations (71) and (72). For $\lambda \in [\mu, L]$,

$$\frac{1-(1-\alpha\lambda)^j}{\lambda} = \alpha\sum_{i=0}^{j-1}(1-\alpha\lambda)^i \le \frac{1-\rho_\mu^j}{\mu}.$$

Thus the bias is at most $R_0(1-\rho_\mu^m)/(\alpha\mu m)$ and the trace of each covariance summand is at most $\sigma^2(1-\rho_\mu^j)^2/\mu^2$. This proves both inequalities in Equation (74).

For the energy bound, the gradient of any smooth strongly convex f satisfies $\|\nabla f(x)\|^2 \le L\langle x - x_\star, \nabla f(x)\rangle$ and $\langle x - x_\star, \nabla f(x)\rangle \ge \mu\|x - x_\star\|^2$. For $0 < \alpha \le 1/L$ these imply

$$\|x - x_\star - \alpha\nabla f(x)\|^2 \le (1-\alpha\mu)\|x - x_\star\|^2.$$

Conditional centering of the oracle noise therefore yields, with $p_k = \mathbb{E}\|e_k\|^2$,

$$\alpha\mu p_k \le p_k - p_{k+1} + \alpha^2\sigma^2/b.$$

Sum over $0 \le k < m$, discard $-p_m$, and use Jensen's inequality for the average. This gives Equation (73). Finally, Markov's inequality gives the stated confidence radius for either valid risk bound. □

The energy bound U_{en} follows from the standard descent argument, whereas U_{con} uses the linear recursion and has a noise term that decays at the statistical rate $1/N$.

For a reproducible second-order comparison, take an orthonormal eigenbasis φ_i of H and set $\lambda_i = \varphi_i^\top H\varphi_i$, $c_i = \varphi_i^\top e_0$, $s_i^2 = \varphi_i^\top\Sigma\varphi_i$ and $\rho_i = 1 - \alpha\lambda_i$. Then, ,

$$\mathcal{R}_{m,b} = \frac{1}{m^2}\sum_i \frac{c_i^2(1-\rho_i^m)^2}{\alpha^2\lambda_i^2} + \frac{1}{bm^2}\sum_i \frac{s_i^2}{\lambda_i^2}\left[m - \frac{2(1-\rho_i^m)}{\alpha\lambda_i} + \frac{1-\rho_i^{2m}}{\alpha\lambda_i(2-\alpha\lambda_i)}\right]. \tag{75}$$

For fixed H, α the second-order approximation is

$$\mathcal{R}_{m,b}^{(2)} := \frac{\operatorname{tr} V_\star}{bm} + \frac{1}{m^2}\left\{\frac{\|H^{-1}e_0\|^2}{\alpha^2} + \frac{1}{b}\sum_i \frac{s_i^2}{\lambda_i^2}\left[\frac{1}{\alpha\lambda_i(2-\alpha\lambda_i)} - \frac{2}{\alpha\lambda_i}\right]\right\}. \tag{76}$$

Table 4: Comparable Guarantees for One Averaged SGD Output

Description	**MSE and confidence radius**	**Assumptions and required knowledge**
Descent	U_{en}; $\sqrt{U_{\mathrm{en}}/\delta}$	Smoothness, strong convexity, conditional variance; μ, L, R_0, σ^2.
Contraction	U_{con}; $\sqrt{U_{\mathrm{con}}/\delta}$	Additive quadratic model, finite variance; the same scalar knowledge.
moments	Equation (71); MSE without an asymptotic remainder	In addition H, Σ, e_0; for unknown e_0 the bound R_0 is used.
PR approximation	$\operatorname{tr} V_\star/N$; a Gaussian quantile only after controlling the center and the covariance	Not a finite-time upper bound.

A remainder bound, valid also for varying parameters, is

$$|\mathcal{R}_{m,b} - \mathcal{R}^{(2)}_{m,b}| \le \frac{2\rho_\mu^m \|H^{-1}e_0\|^2}{\alpha^2 m^2} + \frac{3\rho_\mu^m \operatorname{tr} V_\star}{\alpha\mu b m^2}. \tag{77}$$

It is obtained by subtracting Equation (76) from Equation (75) and using $0 \le \rho_i \le \rho_\mu < 1$.

We begin by comparing descriptions of the same averaged SGD output. Placing the same noise variance along the μ-eigendirection instead of the L-eigendirection leaves the bias and both scalar bounds unchanged, but increases the variance term by a factor tending to κ^2 for fixed κ as $m \to \infty$. Thus, the risk depends on how the noise covariance is aligned with the eigendirections of H.

Proposition D.2 (Same scalar parameters, different risks)**.** *Let $0 < \mu < L$, $H = \operatorname{diag}(\mu, L)$, $e_0 = (r, 0)^\mathsf{T}$, $\alpha = 1/L$, and let the two covariances be*

$$\Sigma_{\mathrm{soft}} = \sigma^2 \operatorname{diag}(1, 0), \qquad \Sigma_{\mathrm{stiff}} = \sigma^2 \operatorname{diag}(0, 1).$$

The two problems share $\mu, L, \operatorname{tr}\Sigma, R_0 = |r|$, the start, the procedure and the budget. With $\rho_\mu = 1 - \mu/L$ they have the same squared bias $\mathrm{Bias}_m^2 = r^2[(1-\rho_\mu^m)/(\alpha\mu m)]^2$ and the risks

$$\mathcal{R}_{\mathrm{soft}} = \mathrm{Bias}_m^2 + \frac{\sigma^2\omega_m(\rho_\mu)}{b\mu^2 m^2}, \qquad \mathcal{R}_{\mathrm{stiff}} = \mathrm{Bias}_m^2 + \frac{\sigma^2(m-1)}{bL^2m^2}.$$

For $m > 1$ the ratio of the variance parts equals $\kappa^2\omega_m(\rho_\mu)/(m-1)$ and tends to κ^2 for fixed κ and $m \to \infty$. The PR terms equal $\sigma^2/(\mu^2 N)$ and $\sigma^2/(L^2 N)$, respectively, whereas both scalar bounds Equations (73) *and* (74) *coincide.*

Proof of Theorem D.2. Apply Equations (71) and (72) in the eigenbasis of $H = \operatorname{diag}(\mu, L)$. The starting vector lies on the first coordinate in both problems, so their biases coincide. On the second coordinate, $1 - \alpha L = 0$, and hence $\sum_{j=0}^{m-1}(1-(1-\alpha L)^j)^2 = m-1$. On the first coordinate this sum is $\omega_m(\rho_\mu)$. These identities give the two displayed risks and their variance ratio. For fixed κ, $\omega_m(\rho_\mu) = m + O(1)$ as $m \to \infty$, which proves the limit. Both scalar upper bounds depend only on $\mu, L, R_0, \operatorname{tr}\Sigma, m, b, \alpha$, which are the same in the two models. □

A nondegenerate variant is obtained by replacing the covariances with $\sigma^2 \operatorname{diag}(1-\varepsilon, \varepsilon)$ and $\sigma^2 \operatorname{diag}(\varepsilon, 1-\varepsilon)$, $0 < \varepsilon < 1/2$; this variant is used in the experiments. The example proves the loss of information in the chosen scalar bounds. Nonasymptotic matrix estimates, including the equality above, retain this information. When the common bias dominates, the ratio of the total risks can be close to one even when the PR covariances differ strongly.

Loss of problem-specific geometry is compatible with exact worst-case sharpness over the class described by those scalar parameters. The following result fixes both the algorithm and its output.

Proposition D.3 (worst-case risk of fixed SGD)**.** *Let $d \ge 2$, $0 < \mu < L$, and let the class of additive quadratic problems fix $\lambda_{\min}(H) = \mu$, $\lambda_{\max}(H) = L$, $\|e_0\| \le R_0$ and $\operatorname{tr}\Sigma = \sigma^2$. For fixed $\alpha \le 1/L$, m and b, the supremum of the risk of the considered SGD average over this class equals U_{con} from Equation* (74)*. The equality holds at every finite budget and therefore persists along any sequences of these parameters satisfying the stated constraints.*

Proof of Theorem D.3. The first expression for U_{con} in Equation (74) is an upper bound for every member of the stated class. It is attained by taking $H = \operatorname{diag}(\mu, L, \mu, \ldots, \mu)$, $e_0 = R_0\mathbf{e}_1$, and $\Sigma = \sigma^2\mathbf{e}_1\mathbf{e}_1^\mathsf{T}$. Indeed, the formula in Equation (72) then has only the μ-mode contribution, and both its squared bias and its variance equal the corresponding terms of U_{con}. The construction also covers $R_0 = 0$ or $\sigma^2 = 0$. □

D.2 Laws for a Specified Initial Point

Proposition D.4 (matrix formula from a cold start)**.** *Let $J = (I, I)^\mathsf{T}$, $R_j = \mathsf{O}_\gamma A^j J$, and $\mathsf{S}_m = \sum_{j=0}^{m-1} R_j$. For deterministic initialization $e_{-1} = e_0$,*

$$\mathbb{E}\bar{u}_{0,m} = m^{-1}\mathsf{S}_m e_0, \qquad \mathrm{Cov}(\bar{u}_{0,m}) = \frac{1}{bm^2}\sum_{j=0}^{m-1}(I - R_j)V_\star(I - R_j)^\mathsf{T}. \tag{78}$$

For fixed stable parameters, put

$$\mathsf{S}_\infty := \frac{1-\beta}{\alpha}H^{-1}, \qquad \mathsf{Q}_\infty := \sum_{j\geq 0} R_j V_\star R_j^\mathsf{T}.$$

Then the cold-start second-order coefficient is

$$\mathrm{Cov}(\bar{u}_{0,m}) = \frac{V_\star}{bm} + \frac{\mathsf{Q}_\infty - \mathsf{S}_\infty V_\star - V_\star \mathsf{S}_\infty^\mathsf{T}}{bm^2} + \textit{exponentially small remainder}.$$

This coefficient differs from the stationary coefficient in Theorem D.7.

Proof of Theorem D.4. The state-space matrices satisfy $(I - A)J = -B_\alpha H$, and $I - A$ is invertible since $H \succ 0$ and $\alpha > 0$. Consequently, for every $j \geq 0$,

$$\sum_{i=0}^{j-1}\mathsf{O}_\gamma A^i B_\alpha = -(I - R_j)H^{-1}.$$

Unrolling the state recursion from $\mathbf{z}_0 = Je_0$ and summing the output therefore gives

$$\bar{u}_{0,m} = \frac{\mathsf{S}_m e_0}{m} - \frac{1}{m}\sum_{s=1}^{m}(I - R_{m-s})H^{-1}\eta_s^{(b)}.$$

Taking its mean and covariance proves Equation (78). For fixed stable parameters, summing the noiseless endpoint identity over the infinite trajectory gives $\mathsf{S}_\infty = (1-\beta)H^{-1}/\alpha$. Expand the covariance as

$$\frac{mV_\star - \mathsf{S}_m V_\star - V_\star \mathsf{S}_m^\mathsf{T} + \sum_{j<m} R_j V_\star R_j^\mathsf{T}}{bm^2}.$$

This proves the second-order coefficient. More explicitly, choose $K_R < \infty$ and $q_R \in (0,1)$ with $\|R_j\| \leq K_R q_R^j$, which is possible for fixed stable A. The norm of the displayed remainder is at most

$$\frac{\|V_\star\|}{bm^2}\left\{\frac{2K_R q_R^m}{1-q_R} + \frac{K_R^2 q_R^{2m}}{1-q_R^2}\right\}.$$

This also specifies the asserted exponential decay. □

Theorem D.5 (NAG covariance in arbitrary directions)**.** *Use the NAG parameters and deterministic zero-velocity start of Theorem B.10. Suppose there are a matrix $Q_\Sigma \succ 0$ commuting with H and numbers $0 < \ell_\Sigma \leq u_\Sigma$ such that*

$$\ell_\Sigma Q_\Sigma \preceq \Sigma \preceq u_\Sigma Q_\Sigma, \qquad \chi_\Sigma := u_\Sigma/\ell_\Sigma.$$

Then, for every $m, b \geq 1$, with $N = mb$,

$$\sup_{c\neq 0}\left|\frac{N\ \mathrm{Var}(c^\mathsf{T}\widehat{x}_{0,m})}{c^\mathsf{T}V_\star c} - 1\right| \leq (18\chi_\Sigma + 6\sqrt{\chi_\Sigma})\frac{\sqrt{\kappa}}{m}. \tag{79}$$

Equivalently, the same bound holds for $\|V_\star^{-1/2}\{N\,\mathrm{Cov}(\widehat{x}_{0,m}) - V_\star\}V_\star^{-1/2}\|_{\mathrm{op}}$. For arbitrary $\Sigma \succ 0$, one may take $Q_\Sigma = I_d$ and $\chi_\Sigma = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. If Σ commutes with H, one may instead take $Q_\Sigma = \Sigma$ and $\chi_\Sigma = 1$.

Proof. The response matrices R_j of Theorem D.4 are symmetric functions of H. The estimates proved in Theorem B.10 hold for every eigenvalue and imply

$$\left\|\sum_{j<m} R_j\right\|_{\mathrm{op}} \leq 3\sqrt{\kappa}, \qquad \sum_{j<m} R_j^2 \preceq 18\sqrt{\kappa}\, I_d.$$

These statements include $\kappa = 1$, where $R_0 = I_d$ and $R_j = 0$ for $j \geq 1$. Put $\mathsf{S}_m = \sum_{j<m} R_j$ and $\mathsf{T}_m = \sum_{j<m} R_j \Sigma R_j$. The covariance formula gives

$$H\{N \operatorname{Cov}(\widehat{x}_{0,m}) - V_\star\}H = \frac{1}{m}\{\mathsf{T}_m - \mathsf{S}_m\Sigma - \Sigma\mathsf{S}_m\}.$$

Since Q_Σ commutes with every R_j,

$$0 \preceq \mathsf{T}_m \preceq u_\Sigma Q_\Sigma^{1/2}\left(\sum_{j<m} R_j^2\right) Q_\Sigma^{1/2} \preceq 18\chi_\Sigma\sqrt{\kappa}\,\Sigma.$$

Also, commuting S_m with Q_Σ gives

$$\begin{aligned}\|\Sigma^{-1/2}\mathsf{S}_m\Sigma^{1/2}\|_{\mathrm{op}} &\leq \|\Sigma^{-1/2}Q_\Sigma^{1/2}\|_{\mathrm{op}}\|\mathsf{S}_m\|_{\mathrm{op}}\|Q_\Sigma^{-1/2}\Sigma^{1/2}\|_{\mathrm{op}}\\ &\leq 3\sqrt{\chi_\Sigma}\sqrt{\kappa}.\end{aligned}$$

Whitening the covariance difference by $\Sigma^{-1/2}$ therefore gives the bound $(18\chi_\Sigma + 6\sqrt{\chi_\Sigma})\sqrt{\kappa}/m$ in operator norm. For $d_c = H^{-1}c$, its Rayleigh quotient is the relative variance error in Equation (79). Taking the supremum over $c \neq 0$ proves both equivalent formulations. □

Remark D.6 (Why noise geometry enters an all-direction bound). A universal constant independent of Σ is impossible. Take $H = \operatorname{diag}(1, 4)$, $\Sigma = \mathbf{1}\mathbf{1}^\mathsf{T} + \varepsilon I_2$, $c = (1, -4)^\mathsf{T}$, $b = 1$, $m = 2$, and zero initial error. NAG has $\alpha = 1/4$, $\beta = \gamma = 1/3$, and its averaged query error is $-\xi_1/6$. Thus

$$c^\mathsf{T}V_\star c = 2\varepsilon, \qquad \frac{N\ \operatorname{Var}(c^\mathsf{T}\widehat{x}_{0,2})}{c^\mathsf{T}V_\star c} = \frac{9 + 17\varepsilon}{36\varepsilon}.$$

This diverges as $\varepsilon \downarrow 0$ while $\kappa = 4$ and $m = 2$ stay fixed. The order $O(\sqrt{\kappa}/m)$ in arbitrary directions is consequently uniform on classes with controlled noise geometry, not over all positive definite noise covariances with one geometry-independent constant. Here $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) = (2 + \varepsilon)/\varepsilon$, so even the linear order of dependence on the noise condition number is unavoidable.

D.3 Stationary Initialization

Let the stable chain start from stationarity, and let P_1 solve $P_1 = AP_1A^\mathsf{T} + B_\alpha\Sigma B_\alpha^\mathsf{T}$. For batch b, its state covariance is P_1/b. Define

$$\Delta_{\mathrm{stat}} := \mathsf{O}_\gamma A(I - A)^{-2}P_1\mathsf{O}_\gamma^\mathsf{T} + \mathsf{O}_\gamma P_1 A^\mathsf{T}(I - A^\mathsf{T})^{-2}\mathsf{O}_\gamma^\mathsf{T}. \tag{80}$$

Theorem D.7 (covariance and second-order expansion)**.** *With* $\Gamma_j^{\mathrm{lag}} = b^{-1}\mathsf{O}_\gamma P_1(A^\mathsf{T})^j\mathsf{O}_\gamma^\mathsf{T}$,

$$\operatorname{Cov}(\bar{u}_{0,m}) = \frac{1}{m^2}\left[m\Gamma_0^{\mathrm{lag}} + \sum_{j=1}^{m-1}(m - j)(\Gamma_j^{\mathrm{lag}} + (\Gamma_j^{\mathrm{lag}})^\mathsf{T})\right], \tag{81}$$

$$= \frac{V_\star}{bm} - \frac{\Delta_{\mathrm{stat}}}{bm^2} + R_m. \tag{82}$$

If $\|A^j\| \leq K\rho_A^j$*, then*

$$\|R_m\| \leq \frac{2K\,\|\mathsf{O}_\gamma\|^2\,\|P_1\|}{bm^2}\frac{\rho_A^{m+1}}{(1 - \rho_A)^2}.$$

Proof of Theorem D.7. At stationarity, $\operatorname{Cov}(\mathbf{z}_k) = P_1/b$, and for $j \geq 0$,

$$\Gamma_j^{\mathrm{lag}} = \operatorname{Cov}(u_0, u_j) = b^{-1}\mathsf{O}_\gamma P_1(A^\mathsf{T})^j\mathsf{O}_\gamma^\mathsf{T}.$$

Expanding the covariance of $m^{-1}\sum_{k=0}^{m-1} u_k$ and grouping pairs by lag gives Equation (81).

Let $\mathcal{G}_j^{\mathrm{lag}} = \Gamma_j^{\mathrm{lag}} + (\Gamma_j^{\mathrm{lag}})^\mathsf{T}$ and

$$\Sigma_{\mathrm{LR}} = \Gamma_0^{\mathrm{lag}} + \sum_{j=1}^{\infty}\mathcal{G}_j^{\mathrm{lag}}.$$

The moving-average representation $u_k = \sum_{r=0}^{\infty} \mathsf{O}_\gamma A^r B_\alpha \eta_{k-r}^{(b)}$ implies

$$\Sigma_{\mathrm{LR}} = \left(\sum_{r=0}^{\infty} \mathsf{O}_\gamma A^r B_\alpha\right) \frac{\Sigma}{b} \left(\sum_{r=0}^{\infty} \mathsf{O}_\gamma A^r B_\alpha\right)^{\mathsf{T}} = \frac{1}{b} \mathsf{O}_\gamma (I-A)^{-1} B_\alpha \Sigma B_\alpha^{\mathsf{T}} (I-A^{\mathsf{T}})^{-1} \mathsf{O}_\gamma^{\mathsf{T}} = \frac{V_\star}{b},$$

where the last equality uses Equation (14). Hence

$$m^2 \operatorname{Cov}(\bar{u}_{0,m}) = m\Sigma_{\mathrm{LR}} - \sum_{j=1}^{m-1} j\mathcal{G}_j^{\mathrm{lag}} - m \sum_{j=m}^{\infty} \mathcal{G}_j^{\mathrm{lag}} = m\Sigma_{\mathrm{LR}} - \sum_{j=1}^{\infty} j\mathcal{G}_j^{\mathrm{lag}} + \sum_{j=m}^{\infty} (j-m)\mathcal{G}_j^{\mathrm{lag}}. \tag{83}$$

Since $\sum_{j\geq 1} jA^j = A(I-A)^{-2}$, $\sum_{j=1}^{\infty} j\mathcal{G}_j^{\mathrm{lag}} = \Delta_{\mathrm{stat}}/b$. Dividing Equation (83) by m^2 proves Equation (82). Finally,

$$\|R_m\| \leq \frac{2K\|\mathsf{O}_\gamma\|^2\|P_1\|}{bm^2} \sum_{j=m}^{\infty} (j-m)\rho_A^j = \frac{2K\|\mathsf{O}_\gamma\|^2\|P_1\|}{bm^2} \frac{\rho_A^{m+1}}{(1-\rho_A)^2},$$

which is the stated remainder bound. □

Thus the leading covariance is method-independent while the integrated autocovariance moment Δ_{stat} is not. In a finite sample, a method can approach $V_\star/(bm)$ from above, from below, or anisotropically.

D.4 Meaning and Limits of the ERM Comparison

Let $\mathfrak{b}_{\mathrm{out}} = \mathbb{E}(\hat{x} - x_\star)$ and $W = \operatorname{Cov}(\hat{x})$ for any output with finite second moments. Its two quadratic risks are

$$\mathbb{E}\|\hat{x} - x_\star\|^2 = \|\mathfrak{b}_{\mathrm{out}}\|^2 + \operatorname{tr} W, \tag{84}$$

$$\mathbb{E}[f(\hat{x}) - f(x_\star)] = \tfrac{1}{2}\mathfrak{b}_{\mathrm{out}}^{\mathsf{T}} H \mathfrak{b}_{\mathrm{out}} + \tfrac{1}{2}\operatorname{tr}(HW). \tag{85}$$

For the ERM, these are respectively $\operatorname{tr}(H^{-1}\Sigma H^{-1})/N$ and $\operatorname{tr}(H^{-1}\Sigma)/(2N)$. Thus function risk and parameter risk can weight the spectrum differently. For the online estimator the moment formulas justify any corresponding leading-risk statement. In a regular correctly specified likelihood model, the additional information identity $\Sigma = H$ gives covariance H^{-1}/N.

Remark D.8 (offline expansions and two confidence sets). With $\bar{\eta} = \bar{\eta}_{t,m}^{(b)}$, the empirical risk of the retained-window observations is $L_N(x) = \frac{1}{2}\langle x - x_\star, H(x - x_\star)\rangle + \langle \bar{\eta}, x - x_\star\rangle$. Hence the Fisher expansion $\hat{x}_{t,m,b}^{\mathrm{ERM}} - x_\star = -H^{-1}\bar{\eta}$ and the Wilks expansion $2N\{L_N(x_\star) - L_N(\hat{x}_{t,m,b}^{\mathrm{ERM}})\} = N\bar{\eta}^{\mathsf{T}} H^{-1}\bar{\eta}$ hold without remainder. For a cutoff $c_{\mathrm{Wilks}} \geq 0$, the likelihood-type set $\{x : N(x - \hat{x}_{t,m,b}^{\mathrm{ERM}})^{\mathsf{T}} H (x - \hat{x}_{t,m,b}^{\mathrm{ERM}}) \leq c_{\mathrm{Wilks}}\}$ has the shape of H, and under Gaussian noise its quantile is that of a weighted chi-square law with weights $\operatorname{spec}(H^{-1/2}\Sigma H^{-1/2})$. The PR ellipsoid has the shape of $V_\star^{-1} = H\Sigma^{-1}H$. The two shapes coincide if and only if Σ is proportional to H, which includes the information identity $\Sigma = H$; otherwise the likelihood-type quantile has to be estimated, for instance by the bootstrap under misspecification of Spokoiny and Zhilova (2015). By Theorem A.2, every such offline statement transfers to the online output up to the bridge correction $H^{-1}\bar{g}_{t,m}$, measured in the corresponding metric. Computing this correction requires known curvature or a separately charged estimate of it.

Remark D.9 (Information classes and a curvature correction). The ERM uses the Hessian of the empirical risk, whereas the online methods considered here use only first-order information and, at most, the scalar parameters μ and L. For any symmetric $\widehat{H} \succ 0$, the corrected output $x_\star + \bar{u}_{t,m} - \widehat{H}^{-1}\bar{g}_{t,m}$ is computable from the queried points and the observed gradients, and by Equation (21)

$$\left(x_\star + \bar{u}_{t,m} - \widehat{H}^{-1}\bar{g}_{t,m}\right) - \hat{x}_{t,m,b}^{\mathrm{ERM}} = (I - \widehat{H}^{-1}H)\, H^{-1}\bar{g}_{t,m}.$$

The correction thus reduces the online–offline gap $H^{-1}\bar{g}_{t,m}$ by the relative curvature error $\|I - \widehat{H}^{-1}H\|_{\mathrm{op}}$; oracle calls used to estimate H, as in Theorem E.9, belong to the budget. With known H, a single query point already attains the minimax risk Equation (86): the estimator $x_0 - H^{-1}\bar{G}(x_0)$, where $\bar{G}(x_0)$ averages N oracle answers at x_0, equals the ERM of these observations. The comparison of online methods in this paper therefore concerns procedures that do not know H.

Proposition D.10 (Gaussian lower bound in a specified class)**.** *For known $H \succ 0$, $\Sigma \succ 0$ and N independent calls of the additive Gaussian oracle, including calls at adaptively chosen points,*

$$\inf_{\widetilde{x}} \sup_{x_\star \in \mathbb{R}^d} \mathbb{E}_{x_\star}\|\widetilde{x} - x_\star\|^2 = \frac{\operatorname{tr}(H^{-1}\Sigma H^{-1})}{N}. \tag{86}$$

The infimum is taken over all estimators using these calls.

Proof of Theorem D.10. At any possibly adaptive query x_i, form the observable vector

$$Y_i = x_i - H^{-1}G(x_i) = x_\star - H^{-1}\xi_i.$$

The Y_i are independent $\mathcal{N}(x_\star, V_\star)$ observations. Conversely, given these observations and the procedure's random seed, one reconstructs its successive queries and answers by $G(x_i) = H(x_i - Y_i)$. Thus the adaptive transcript contains this Gaussian location information; independent algorithmic randomization cannot decrease the optimal squared-error Bayes risk.

Give $x_\star$ the prior $\mathcal{N}(0, \tau^2 I_d)$. The posterior covariance after N observations is $(\tau^{-2}I_d + NV_\star^{-1})^{-1}$, so every estimator has maximal risk at least its trace. Letting $\tau \to \infty$ gives the lower bound $\mathrm{tr}(V_\star)/N$. The estimator $N^{-1}\sum_i Y_i$ has that risk for every $x_\star$, proving equality. □

The class in Equation (86) does not restrict $\|e_0\|$. For example, for $e_0 = 0$ the cold average can have risk below the PR term. For the full average at finite m, the supremum of the risk over unbounded $x_\star$ is even infinite because of the bias. In the joint small-noise regime, the formula gives a lower bound only for the stated Gaussian location problem at each N.

E Further Calibration and Inference Results

This section develops additional inference guarantees beyond the main calibration results. It covers normal approximation, confidence-set accuracy, the effects of burn-in and switching, and data-independent selection among specified procedures.

E.1 Normal Approximation and the Actual Variance

E.1.1 A Boundary-Based Normal Approximation

For a unit direction a, set $\sigma_a^2 = a^\mathsf{T} V_\star a$ and $\rho_a = \mathbb{E}\left|a^\mathsf{T} H^{-1}\xi\right|^3$. The triangle inequality for Wasserstein distance and the classical Berry–Esseen bound applied to the first term in Equation (15) give the following finite-m statement.

Corollary E.1 (Directional Gaussian approximation)**.** *If $\sigma_a^2 > 0$ and $\rho_a < \infty$, then for $N = mb$ there is a universal constant C_{WBE} such that*

$$d_{\mathrm{W}}\left(\mathcal{L}\left(\sqrt{N}\, a^\mathsf{T}\bar{u}_{t,m}\right), \mathcal{N}(0, \sigma_a^2)\right) \le \frac{C_{\mathrm{WBE}}\rho_a}{\sigma_a^2\sqrt{N}} + \frac{\sqrt{b/m}}{\alpha}\,\mathbb{E}\left|a^\mathsf{T} H^{-1}\Delta_{t,m}\right|. \tag{87}$$

Proof of Theorem E.1. Set $Y_a = a^\mathsf{T} H^{-1}\xi$, so that $\mathbb{E}Y_a = 0$, $\mathrm{Var}(Y_a) = \sigma_a^2$, and $\mathbb{E}|Y_a|^3 = \rho_a$. The one-dimensional Wasserstein Berry–Esseen inequality gives

$$d_{\mathrm{W}}\left(\mathcal{L}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} Y_{a,i}\right), \mathcal{N}(0, \sigma_a^2)\right) \le \frac{C_{\mathrm{WBE}}\rho_a}{\sigma_a^2\sqrt{N}}.$$

The sign of the sum is immaterial. By Equation (15), the remaining scalar perturbation is

$$R_{\mathrm{BE}} = \frac{\sqrt{N}}{\alpha m} a^\mathsf{T} H^{-1}\Delta_{t,m}.$$

For random variables X, R_{BE} on the same probability space, $d_{\mathrm{W}}(\mathcal{L}(X + R_{\mathrm{BE}}), \mathcal{L}(X)) \le \mathbb{E}|R_{\mathrm{BE}}|$. The triangle inequality now yields Equation (87). □

Here d_{W} is the Wasserstein distance of order one. This general perturbation bound is valid but can be loose; Section B.2 instead uses the independent weights. The first term depends only on the total number of oracle samples. The second depends on the number of *sequential rounds* and on the optimization dynamics. This split is the distributional analogue of "statistical error plus optimization error," but now the statistical term has its covariance and a quantitative normal-approximation scale.

Proposition E.2. *Let $r_{\mathrm{pole}} \ge 1$ be fixed and let $d_i^{\mathrm{imp}} \ge 0$, $d_0^{\mathrm{imp}} > 0$, satisfy*

$$i \longmapsto \frac{d_i^{\mathrm{imp}}}{\binom{i+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}-1}} \quad \textit{is nonincreasing}.$$

Set $w_j = \sum_{i=0}^{j-1} d_i^{\mathrm{imp}}$, $m \geq 2$, and let $\varepsilon_{j,k}$ be i.i.d. standardized noises. The weighted statistic

$$T_{m,b} = \frac{\sum_{j=0}^{m-1} \sum_{k=1}^{b} w_j \varepsilon_{j,k}}{\sqrt{b \sum_{j=0}^{m-1} w_j^2}}$$

obeys

$$d_{\mathrm{K}}(\mathcal{L}(T_{m,b}), \mathcal{N}(0,1)) \leq \frac{C_{\mathrm{BE}} \sqrt{2r_{\mathrm{pole}} + 1}\, \nu_3}{\sqrt{mb}}, \qquad d_{\mathrm{W}}(\mathcal{L}(T_{m,b}), \mathcal{N}(0,1)) \leq \frac{3\sqrt{2r_{\mathrm{pole}} + 1}\, \nu_3}{\sqrt{mb}}.$$

With a common finite-variance noise law, $T_{m,b} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ whenever $mb \to \infty$, even if the impulse sequence varies arbitrarily subject to the displayed assumption. This includes repeated-pole impulses $d_i^{\mathrm{imp}} = c_{\mathrm{imp}} \binom{i+r_{\mathrm{pole}}-1}{r_{\mathrm{pole}}-1} \rho_{\mathrm{imp}}^i$, $c_{\mathrm{imp}} > 0$, $\rho_{\mathrm{imp}} \in [0,1)$, and their positive mixtures. SGD admits $r_{\mathrm{pole}} = 1$, giving $\sqrt{3}$ in place of the common $\sqrt{5}$ constant for the three methods.

Proof of Theorem E.2. Write $r = r_{\mathrm{pole}}$ and $D_j = \sum_{i=0}^{j-1} \binom{i+r-1}{r-1} = \binom{j+r-1}{r}$ for $j \geq 1$, with $D_0 = 0$. The nonincreasing normalized impulses imply that their weighted average over any prefix is at least their weighted average over $0, \ldots, m-2$. Hence

$$\frac{w_j}{w_{m-1}} \geq \frac{D_j}{D_{m-1}} \geq \left(\frac{j}{m-1}\right)^r, \qquad 0 \leq j < m.$$

For completeness, the second inequality follows because $D_{j+1}/D_j = 1 + r/j$ and $(1+1/j)^r \geq 1 + r/j$ for $r \geq 1$, so D_j/j^r is nonincreasing. This argument also applies if the binomial coefficients are interpreted through the gamma function for noninteger r. Convexity of x^{2r} and the trapezoidal rule give

$$\sum_{j=0}^{m-1} \left(\frac{j}{m-1}\right)^{2r} \geq \frac{m-1}{2r+1} + \frac{1}{2} \geq \frac{m}{2r+1}.$$

It follows that $\sum_j w_j^2 \geq m w_{m-1}^2/(2r+1)$. The individual standardized coefficients $a_{j,k} = w_j / \sqrt{b \sum_\ell w_\ell^2}$ thus satisfy

$$\sum_{j,k} a_{j,k}^2 = 1, \qquad \max_{j,k} |a_{j,k}| \leq \sqrt{\frac{2r+1}{mb}}, \qquad \sum_{j,k} |a_{j,k}|^3 \leq \sqrt{\frac{2r+1}{mb}}.$$

The independent-summand Kolmogorov and Wasserstein Berry–Esseen inequalities yield the two claimed bounds. With only the common finite second moment, the Lindeberg sum is at most

$$\mathbb{E}\Big[\varepsilon^2 \mathbf{1}_{\{|\varepsilon| > \epsilon\sqrt{mb/(2r+1)}\}}\Big] \longrightarrow 0$$

for every $\epsilon > 0$, which proves the triangular-array CLT. Repeated-pole impulses and their positive mixtures satisfy the assumed monotonicity directly. □

Proposition E.3 (The covariance-calibration term is sharp)**.** *Suppose the noise is Gaussian. Define $Q_\infty = \sum_{j \geq 0} r_j^2$ and $\Delta_\infty^{\mathrm{cov}} = 2S_\infty - Q_\infty$. For fixed κ and $m \to \infty$,*

$$d_{\mathrm{W}}(\mathcal{L}(Z_m - h_m), \mathcal{N}(0,1)) = \frac{\sqrt{2/\pi}}{2} \frac{\Delta_\infty^{\mathrm{cov}}}{m} + o(m^{-1}),$$
$$d_{\mathrm{K}}(\mathcal{L}(Z_m - h_m), \mathcal{N}(0,1)) = \frac{\phi(1)}{2} \frac{\Delta_\infty^{\mathrm{cov}}}{m} + o(m^{-1}).$$

Here $S_\infty \leq \Delta_\infty^{\mathrm{cov}} \leq 2S_\infty$. If simultaneously $\kappa \to \infty$ and $m/S_\infty \to \infty$, the same formulas hold with $\Delta_\infty^{\mathrm{cov}}/m$ replaced by $c_{\mathrm{alg}} S_\infty/m$ and remainder $o(S_\infty/m)$, where

$$c_{\mathrm{SGD}} = \frac{3}{2}, \qquad c_{\mathrm{HB}} = c_{\mathrm{NAG}} = \frac{11}{8}.$$

Therefore the term S_∞/m cannot be replaced uniformly by $o(S_\infty/m)$ when the target variance is the PR variance, even for Gaussian noise and after exact centering.

Proof. For Gaussian noise, the centered statistic is $\mathcal{N}(0, v_m)$, where

$$1 - v_m = \frac{A_m}{m}, \qquad A_m := 2S_m - Q_m.$$

For fixed κ, summability gives $A_m \to \Delta_\infty^{\mathrm{cov}}$. Coupling two centered normals with the same standard Gaussian and expanding the square root yield

$$d_{\mathrm{W}}(\mathcal{N}(0, v_m), \mathcal{N}(0,1)) = \sqrt{\frac{2}{\pi}}\,(1 - \sqrt{v_m}) = \frac{\sqrt{2/\pi}}{2}\frac{\Delta_\infty^{\mathrm{cov}}}{m} + o(m^{-1}).$$

Similarly, uniformly in x, a first-order expansion of $\Phi(x/\sqrt{v_m}) - \Phi(x)$ gives

$$d_{\mathrm{K}}(\mathcal{N}(0, v_m), \mathcal{N}(0,1)) = \frac{1}{2}\sup_x |x|\phi(x)\frac{\Delta_\infty^{\mathrm{cov}}}{m} + o(m^{-1}) = \frac{\phi(1)}{2}\frac{\Delta_\infty^{\mathrm{cov}}}{m} + o(m^{-1}).$$

Since $0 \le r_j \le 1$, one has $0 \le Q_\infty \le S_\infty$, and hence $S_\infty \le \Delta_\infty^{\mathrm{cov}} \le 2S_\infty$.

In the joint regime $m/S_\infty \to \infty$, the response tails in Theorem B.2 give $S_m = S_\infty + o(S_\infty)$ and $Q_m = Q_\infty + o(S_\infty)$. Direct summation of $(1+cj)^2\varrho^{2j}$ gives

$$\frac{Q_\infty}{S_\infty} \longrightarrow \begin{cases} 1/2, & \text{SGD}, \\ 5/8, & \text{HB or NAG}, \end{cases}$$

so $A_m/S_\infty \to 3/2$ for SGD and $A_m/S_\infty \to 11/8$ for HB and NAG. Substitution in the two normal-distance expansions proves the joint claims. □

Proposition E.4 (A continuum of finite-mixing Gaussian limits)**.** *Fix one of the standard policies* $\mathrm{alg} \in \{\mathrm{SGD}, \mathrm{HB}, \mathrm{NAG}\}$ *and let* $S_\infty = S_{\infty,\mathrm{alg}}$*. Assume that the standardized noise inputs* ξ/σ *have a common iid law of mean zero and variance one across the sequence of experiments. Allow* $\kappa \to \infty$*,* $m/S_\infty \to \tau_{\mathrm{mix}} \in (0,\infty)$*, and any integer batch sequence* $b \ge 1$*. Define the limiting response profiles by*

$$R_{\mathrm{SGD}}(x) = e^{-x}, \qquad R_{\mathrm{HB}}(x) = R_{\mathrm{NAG}}(x) = (1+2x)e^{-2x}, \qquad x \ge 0,$$

and define

$$a_{\mathrm{alg}}(\tau_{\mathrm{mix}}) = \frac{1}{\tau_{\mathrm{mix}}}\int_0^{\tau_{\mathrm{mix}}} R_{\mathrm{alg}}(x)\,dx, \qquad v_{\mathrm{alg}}(\tau_{\mathrm{mix}}) = \frac{1}{\tau_{\mathrm{mix}}}\int_0^{\tau_{\mathrm{mix}}} (1 - R_{\mathrm{alg}}(x))^2\,dx.$$

Then $S_m/m \to a_{\mathrm{alg}}(\tau_{\mathrm{mix}})$*,* $v_m \to v_{\mathrm{alg}}(\tau_{\mathrm{mix}}) \in (0,1)$*, and*

$$Z_m - h_m \xrightarrow{\mathrm{d}} \mathcal{N}(0, v_{\mathrm{alg}}(\tau_{\mathrm{mix}})).$$

If the start is local in the sense that $\mu\sqrt{N}e_0/\sigma \to \vartheta \in \mathbb{R}$*, the uncentered limit is* $\mathcal{N}(\vartheta a_{\mathrm{alg}}(\tau_{\mathrm{mix}}), v_{\mathrm{alg}}(\tau_{\mathrm{mix}}))$*. HB and NAG have identical limiting profiles after time is measured in units of their respective* $S_{\infty,\mathrm{alg}}$*.*

Proof of Theorem E.4. Write $B = S_{\infty,\mathrm{alg}}$. By Theorem B.2, $r_j = (1+cj)\varrho^j$. As $\kappa \to \infty$, the constants in Table 2 satisfy

$$(B\log\varrho, Bc) \longrightarrow \begin{cases} (-1, 0), & \text{SGD}, \\ (-2, 2), & \text{HB or NAG}. \end{cases}$$

Consequently $r_{\lfloor Bx \rfloor} \to R_{\mathrm{alg}}(x)$ uniformly on each compact interval of $[0,\infty)$. Since $m/B \to \tau_{\mathrm{mix}}$, the corresponding Riemann sums give

$$\frac{S_m}{m} = \frac{B}{m}\frac{1}{B}\sum_{j=0}^{m-1} r_j \longrightarrow a_{\mathrm{alg}}(\tau_{\mathrm{mix}}),$$

$$v_m = \frac{B}{m}\frac{1}{B}\sum_{j=0}^{m-1}(1 - r_j)^2 \longrightarrow v_{\mathrm{alg}}(\tau_{\mathrm{mix}}).$$

For each $x > 0$, $0 < R_{\mathrm{alg}}(x) < 1$, so the limiting variance lies strictly between zero and one.

Let $\varepsilon_{j,i}$ have the common law. The law Equation (22) writes the centered statistic as

$$Z_m - h_m = -\sum_{j=0}^{m-1}\sum_{i=1}^{b} a_{j,i}\varepsilon_{j,i}, \qquad a_{j,i} = \frac{1-r_j}{\sqrt{mb}}.$$

The total variance is $\sum_{j,i} a_{j,i}^2 = v_m$, and $\max_{j,i}|a_{j,i}| \le N^{-1/2} \to 0$, because $B \to \infty$, $m/B \to \tau_{\mathrm{mix}} > 0$, and $N = mb \ge m$. For every $\epsilon > 0$, the Lindeberg sum is bounded by

$$\sum_{j,i} a_{j,i}^2 \mathbb{E}\big[\varepsilon_{j,i}^2 \mathbf{1}_{\{|a_{j,i}\varepsilon_{j,i}|>\epsilon\}}\big] \le v_m \mathbb{E}\big[\varepsilon^2 \mathbf{1}_{\{|\varepsilon|>\epsilon\sqrt{N}\}}\big] \longrightarrow 0.$$

The last step uses only the common finite second moment. The Lindeberg–Feller theorem proves the centered Gaussian limit. Finally,

$$h_m = \frac{\mu\sqrt{N}e_0}{\sigma}\frac{S_m}{m} \longrightarrow \vartheta a_{\mathrm{alg}}(\tau_{\mathrm{mix}}),$$

so Slutsky's theorem gives the stated uncentered limit. The identical HB and NAG profiles yield their asserted agreement in units of B. □

E.2 Confidence Sets and Coverage Accuracy

The next guarantees use two distinct approximation targets. The multivariate balls below concern the SGD output and matrix geometry of Theorem D.1; the scalar interval bounds use the least-curved-mode law of Theorem B.7.

E.2.1 Multidimensional Confidence Balls

Proposition E.5 (Finite-time Gaussian ball)**.** *Under independent Gaussian additive noise, set $\mu_m^{\mathrm{out}} = \mathsf{S}_{\alpha,m}e_0/m$ and take W_m from Equation* (72)*. Then $\widehat{x} - x_\star \overset{\mathrm{d}}{=} \mu_m^{\mathrm{out}} + W_m^{1/2}Z$, $Z \sim \mathcal{N}(0, I_d)$. If*

$$\mathrm{Quant}_{\pi_{\mathrm{cov}}}(m_0, W) := \inf\{r \ge 0 : \mathbb{P}(\|m_0 + W^{1/2}Z\| \le r) \ge \pi_{\mathrm{cov}}\},$$

then the ball of radius $\mathrm{Quant}_{1-\delta}(\mu_m^{\mathrm{out}}, W_m)$ has coverage at least $1-\delta$ ($1-\delta$ when $W_m \succ 0$). When only the bound $\|e_0\| \le R_0$ on the start is known, the computable radius

$$r_{\mathrm{G}} = R_0\|\mathsf{S}_{\alpha,m}/m\|_{\mathrm{op}} + \mathrm{Quant}_{1-\delta}(0, W_m) \tag{88}$$

guarantees the same lower coverage. Without computing the quantile one can take

$$r_{\mathrm{G}}^{\mathrm{up}} = R_0\|\mathsf{S}_{\alpha,m}/m\|_{\mathrm{op}} + \sqrt{\mathrm{tr}\, W_m + 2\sqrt{\mathrm{tr}(W_m^2)\log(1/\delta)} + 2\|W_m\|_{\mathrm{op}}\log(1/\delta)}. \tag{89}$$

Proof of Theorem E.5. The linear representation in the proof of Theorem D.1 is Gaussian, with mean μ_m^{out} and covariance W_m. This proves the asserted law. The quantile definition gives coverage at least $1-\delta$. If $W_m \succ 0$, the Gaussian has a positive density, and the distribution of its distance from the origin is continuous and strictly increasing on positive radii; the coverage is then exact. Also $\|\mu_m^{\mathrm{out}}\| \le R_0\|\mathsf{S}_{\alpha,m}/m\|_{\mathrm{op}}$, so the triangle inequality proves Equation (88).

Here is a direct concentration proof of the upper radius. First assume $W_m \ne 0$. If λ_i are the eigenvalues of W_m, set $Q = \sum_i \lambda_i Z_i^2$, $a_W = \mathrm{tr}(W_m^2)$ and $c_W = 2\|W_m\|_{\mathrm{op}}$. For $0 < s < c_W^{-1}$,

$$\log \mathbb{E}e^{s(Q-\mathrm{tr}\, W_m)} = \frac{1}{2}\sum_i\{-\log(1-2s\lambda_i) - 2s\lambda_i\} \le \frac{a_W s^2}{1-c_W s}.$$

The inequality follows by the power series of $-\log(1-u)$. For $a_W > 0$, Chernoff's inequality with $s = \sqrt{x}/(\sqrt{a_W} + c_W\sqrt{x})$ gives

$$\mathbb{P}\{Q > \mathrm{tr}\, W_m + 2\sqrt{a_W x} + c_W x\} \le e^{-x}.$$

If $W_m = 0$ the same conclusion is immediate. Take $x = \log(1/\delta)$, then a square root and the same triangle inequality, to obtain Equation (89). □

The noncentral quantile uses $e_0 = x_0 - x_\star$ and is therefore usually a diagnostic benchmark. Known H, Σ do not by themselves make e_0 known. The radius Equation (88) uses only the stated bound R_0.

Proposition E.6 (Approximation control on the class of balls)**.** *Let the noise be independent with a finite third moment and $W_m \succ 0$. Denote the independent contributions of individual samples by*

$$Y_{si} = -\frac{1}{mb}(I - F_\alpha^{m-s})H^{-1}\xi_{si}, \qquad 1 \le s \le m, \quad 1 \le i \le b,$$

and set

$$\varepsilon_m^{\text{ball}} = \min\left\{1, (42d^{1/4} + 16)\sum_{s,i} \mathbb{E}\|W_m^{-1/2}Y_{si}\|^3\right\}.$$

Then, uniformly in $r \ge 0$,

$$\left|\mathbb{P}\{\|\widehat{x} - x_\star\| \le r\} - \mathbb{P}\{\|\mu_m^{\text{out}} + W_m^{1/2}Z\| \le r\}\right| \le \varepsilon_m^{\text{ball}}. \tag{90}$$

If $\varepsilon_m^{\text{ball}} < \delta$, then $R_0\|\mathsf{S}_{\alpha,m}/m\|_{\text{op}} + \text{Quant}_{1-\delta+\varepsilon_m^{\text{ball}}}(0, W_m)$ is a valid radius of level $1-\delta$.

Proof of Theorem E.6. Put $X_{si} = W_m^{-1/2}Y_{si}$. The variables are independent and centered and satisfy $\sum_{s,i}\text{Cov}(X_{si}) = I_d$. The convex-set Berry–Esseen inequality of Raič (2019, Theorem 1.1) therefore bounds the Gaussian approximation error by $(42d^{1/4} + 16)\sum_{s,i}\mathbb{E}\|X_{si}\|^3$. The set

$$\{z : \|\mu_m^{\text{out}} + W_m^{1/2}z\| \le r\}$$

is convex for every $r \ge 0$. Applying the inequality to this set, and also using the trivial bound one, proves Equation (90). For the enlarged radius in the proposition, the Gaussian probability is at least $1 - \delta + \varepsilon_m^{\text{ball}}$ by the triangle inequality and the bound on $\|\mu_m^{\text{out}}\|$. Subtracting the approximation error proves coverage at least $1-\delta$. □

E.2.2 Scalar Intervals and High Confidence Levels

The scalar bounds below assume known curvature and noise scale. A small error probability is kept separate from heavy tails of the oracle; the latter are treated in Section F. For $m \ge 4S_\infty$, Theorem B.7 gives the bound

$$d_{\text{K}}(\mathcal{L}(Z_m), \mathcal{N}(0,1)) \le \epsilon_{N,m} := \underbrace{\frac{\sqrt{5}C_{\text{BE}}\nu_3}{\sqrt{N}}}_{\text{shape}} + \underbrace{2\phi(1)\frac{S_\infty}{m}}_{\text{covariance}} + \underbrace{\frac{|\theta|S_m\sqrt{b/m}}{\sqrt{2\pi}}}_{\text{start}}, \tag{91}$$

where $\theta = \mu e_0/\sigma$. For the interval with half-width $\Phi^{-1}(1-\delta/2)\sigma/(\mu\sqrt{N})$, its coverage differs from $1-\delta$ by at most $2\epsilon_{N,m}$. This follows by subtracting the cdf at the two endpoints (using the left limit when there is an atom). For Gaussian noise inputs the shape term is zero and the law $\mathcal{N}(h_m, v_m)$ gives the coverage at every δ. Consequently, vanishing center and covariance errors together with the shape bound imply nominal coverage at every fixed confidence level.

For a vanishing error probability δ_N, a sufficient condition for *relative* miscoverage accuracy is $\epsilon_{N,m} = o(\delta_N)$, which is stronger than convergence in law. As a separate concentration statement, if the standardized noise inputs satisfy $\mathbb{E}e^{s\xi/\sigma} \le e^{s^2/2}$ for all real s, then for every $0 < \delta < 1$ the deterministic weights give

$$\mathbb{P}\Big\{|Z_m - h_m| > \sqrt{2v_m\log(2/\delta)}\Big\} \le \delta. \tag{92}$$

Indeed, independence bounds the mgf of $Z_m - h_m$ by $\exp(s^2v_m/2)$, and Chernoff's inequality proves the result. Covering $x_\star$ also requires a bound on $|h_m|$.

Theorem E.7 (Coverage of the entire PR interval family)**.** *Consider the least-curved direction, the three standard policies, and a common iid standardized noise law of mean zero and variance one as in Theorem B.8. Let $m \to \infty$; allow $\kappa \ge 1$, $b \ge 1$, and θ to vary. With $B = S_\infty$, the following assertions are equivalent:*

1. *For every fixed $\delta \in (0,1)$, $\mathbb{P}\{|Z_m| \le z_{1-\delta/2}\} \to 1-\delta$.*
2. *$h_m \to 0$ and $v_m \to 1$.*
3. *$B/m \to 0$ and $|\theta|B\sqrt{b/m} \to 0$.*

Equivalently in the first assertion, the supremum of the absolute coverage error over $\delta \in (0,1)$ tends to zero.

Proof. The equivalence of the second and third assertions is Theorem B.3. By Theorem B.8, $T_m = (Z_m - h_m)/\sqrt{v_m} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. Thus the second assertion implies the first by Slutsky's theorem.

Conversely, $0 < v_m \le 1$. If $|h_m|$ were unbounded along a subsequence, then for every fixed $z > 0$, Chebyshev's inequality would give

$$\mathbb{P}\{|Z_m| \le z\} \le \frac{v_m}{(|h_m| - z)^2} \longrightarrow 0,$$

contradicting the first assertion. Hence h_m is bounded. From any subsequence extract a further one with $h_m \to h$ and $v_m \to v \in [0,1]$. The Gaussian shape limit gives $Z_m \xrightarrow{\mathrm{d}} h + \sqrt{v}\, G$ for $G \sim \mathcal{N}(0,1)$. The first assertion identifies the law of its absolute value as that of $|G|$; this follows first at the continuity points and then everywhere by right continuity. In particular, its second and fourth moments equal those of G:

$$h^2 + v = 1, \qquad h^4 + 6h^2 v + 3v^2 = 3.$$

Subtracting three times the square of the first equation from the second yields $-2h^4 = 0$, so $h = 0$ and $v = 1$. Every convergent subsequence has this limit, proving the second assertion. Finally, convergence in distribution of $|Z_m|$ to the continuous law of $|G|$ implies uniform convergence of their distribution functions, which gives the asserted uniform absolute coverage formulation. □

Corollary E.8 (One confidence level with a nonvanishing start)**.** *Under the assumptions of Theorem E.7, suppose in addition that* $\liminf |\theta| > 0$*; in particular,* $\theta \neq 0$ *may be fixed. For any one fixed* $\delta \in (0,1)$*,*

$$\mathbb{P}\{|Z_m| \le z_{1-\delta/2}\} \to 1 - \delta \quad \Longleftrightarrow \quad \frac{B}{m} \to 0, \qquad |\theta| B \sqrt{\frac{b}{m}} \to 0.$$

For a fixed nonzero θ*, the right-hand side is equivalently* $bB^2/m \to 0$*, or* $b = o(\sqrt{N}/(|\theta| B))$ *with* $N = mb$*. If* $\theta = 0$ *identically, the same one-level equivalence holds with the sole condition* $B/m \to 0$*.*

Proof. Sufficiency follows from Theorem E.7. For necessity, put $z = z_{1-\delta/2}$. The same Chebyshev argument, using the positive limit $1 - \delta$, shows that h_m is bounded. If B/m did not vanish, the argument in the proof of Theorem B.3 would give $S_m/m \ge c > 0$ along a subsequence. But then

$$|h_m| = |\theta| \sqrt{mb}\, \frac{S_m}{m} \ge c|\theta|\sqrt{mb} \longrightarrow \infty,$$

a contradiction. Hence $v_m \to 1$. Along any subsequence where $h_m \to h$, the limiting coverage is $\Phi(z-h) - \Phi(-z-h)$. For $z > 0$ this even function is strictly decreasing in $|h| > 0$, so it equals $2\Phi(z) - 1$ only if $h = 0$. Therefore $h_m \to 0$, and Theorem B.3 proves the required conditions. For fixed nonzero θ, $bB^2/m \to 0$ also implies $B/m \to 0$ because $b \ge 1$ and $B \ge 1$. This proves the last equivalences. If $\theta = 0$ identically, then $h_m = 0$. Along a subsequence with $v_m \to v \in [0,1]$, the limiting coverage equals $2\Phi(z/\sqrt{v}) - 1$ for $v > 0$, and equals one for $v = 0$. It is nominal only for $v = 1$, proving the zero-start case. □

The additional hypothesis in Theorem E.8 is essential for a one-level converse. If the start is allowed to shrink, the finite-mixing limits of Theorem E.4 permit any limiting variance $v \in (0,1)$ in the corresponding range and an adjustable limiting mean h. For a fixed $z > 0$, continuity and strict monotonicity in $|h|$ give a nonzero h such that $\mathbb{P}\{|h + \sqrt{v}\, G| \le z\} = 2\Phi(z) - 1$ for $G \sim \mathcal{N}(0,1)$. Thus one-level cancellation is compatible with the general model but excluded by a nonvanishing standardized start.

E.2.3 Unknown Curvature and Covariance: Independent Calibration

The next result keeps the center of the ball equal to the original output $\widehat{x}$ and explicitly pays for the additional oracle calls. It requires the oracle $G(x) = H(x - x_\star) + \xi$ with independent $\xi \sim \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$.

Proposition E.9 (Finite-time certificate with curvature estimation)**.** *After training with* N_{train} *calls, fix a probe spacing* $\delta_H > 0$ *and* $n > d$ *and the probe points* $y_0^{\mathrm{probe}} = \widehat{x}$ *and* $y_j^{\mathrm{probe}} = \widehat{x} + \delta_H \mathbf{e}_j$*, where* $\mathbf{e}_j$ *is the* j*th coordinate vector,* $1 \le j \le d$*. At each point obtain* n *new independent gradients, their mean* $\bar{G}_j$ *and the unbiased sample covariance* $\widehat{\Sigma}_j$*. All new noises are independent of the training; conditionally on the training sample the calibration groups are independent. For* $\delta_j = \delta/(d+1)$ *set*

$$c_j = \frac{d(n-1)}{n(n-d)} F^{-1}_{d,n-d}(1 - \delta_j), \qquad \epsilon_j = \sqrt{c_j \lambda_{\max}(\widehat{\Sigma}_j)}, \qquad \widetilde{H}_{\cdot j} = \frac{\bar{G}_j - \bar{G}_0}{\delta_H},$$

$$\widehat{H} = \tfrac{1}{2}(\widetilde{H} + \widetilde{H}^{\mathsf{T}}), \qquad \epsilon_H = \frac{1}{\delta_H} \sqrt{\sum_{j=1}^{d} (\epsilon_j + \epsilon_0)^2}, \qquad \widehat{\mu} = \lambda_{\min}(\widehat{H}). \tag{93}$$

Here $F^{-1}_{d,n-d}$ is the quantile of the Fisher distribution. When $\widehat{\mu} > \epsilon_H$, define

$$r_{\mathrm{cal}} = \frac{\widehat{\mu}\|\widehat{H}^{-1}\bar{G}_0\| + \epsilon_0}{\widehat{\mu} - \epsilon_H}; \qquad r_{\mathrm{cal}} = +\infty \quad \textit{otherwise}. \tag{94}$$

Then $\mathbb{P}\{\|\widehat{x} - x_\star\| \le r_{\mathrm{cal}}\} \ge 1 - \delta$ at the total budget $N_{\mathrm{train}} + (d+1)n$. If a bound $\lambda_{\min}(H) \ge \mu_0 > 0$ is known, one group at $\widehat{x}$ suffices: the radius $(\|\bar{G}_0\| + \epsilon_0)/\mu_0$ with $\delta_0 = \delta$ has the same coverage and requires only n additional calls.

Proof of Theorem E.9. Condition on the training data. The probe points are then fixed. Write $g_j = H(y_j^{\mathrm{probe}} - x_\star)$ and $r_j^{\mathrm{probe}} = \bar{G}_j - g_j$. Gaussian sampling and $n > d$ give the Hotelling pivot

$$\frac{n(n-d)}{d(n-1)}(r_j^{\mathrm{probe}})^{\mathsf{T}}\widehat{\Sigma}_j^{-1} r_j^{\mathrm{probe}} \overset{\mathrm{d}}{=} F_{d,n-d}.$$

Consequently $\|r_j^{\mathrm{probe}}\| \le \epsilon_j$ with conditional probability at least $1 - \delta_j$. A union bound gives all these inequalities simultaneously with conditional probability at least $1 - \delta$. On that event,

$$\|\widetilde{H}_{\cdot j} - H_{\cdot j}\| \le \frac{\epsilon_j + \epsilon_0}{\delta_H}, \qquad \|\widehat{H} - H\|_{\mathrm{op}} \le \|\widetilde{H} - H\|_{\mathrm{F}} \le \epsilon_H.$$

Symmetrization cannot increase the operator norm error because H is symmetric. Put $e = \widehat{x} - x_\star$. If $\widehat{\mu} > \epsilon_H$, then $\widehat{H} \succ 0$ and $\bar{G}_0 = He + r_0^{\mathrm{probe}}$ imply

$$\|e\| \le \|\widehat{H}^{-1}\bar{G}_0\| + \frac{\epsilon_H}{\widehat{\mu}}\|e\| + \frac{\epsilon_0}{\widehat{\mu}}.$$

Solving this inequality gives Equation (94). Otherwise the infinite radius covers . Taking expectations over the training data proves the unconditional coverage. With known μ_0, the one-group bound follows instead from $\mu_0\|e\| \le \|He\| \le \|\bar{G}_0\| + \epsilon_0$. The numbers of oracle calls are respectively $(d+1)n$ and n. □

E.3 Full Averaging, Burn-In, and Switching

We first fix the full budget and keep the randomness of the state after burn-in. This allows a correct comparison between continuing the accelerated method and switching to SGD. The following finite-dimensional formulas keep the random state after burn-in. They require only independent additive noise inputs with covariance Σ/b. Let $T = t + m$, $E_1 = [I, 0]$, $J = [I, I]^{\mathsf{T}}$, and, for a fixed method, define

$$\mathsf{L}_m = \frac{1}{m}\sum_{k=0}^{m-1} \mathsf{O}_\gamma A^k, \qquad K_{m,s} = \frac{1}{m}\sum_{k=s}^{m-1} \mathsf{O}_\gamma A^{k-s} B_\alpha \quad (1 \le s \le m),$$

where an empty sum is zero. For a deterministic initial state $\mathbf{z}_0$, write

$$a_t = A^t \mathbf{z}_0, \qquad P_t = \frac{1}{b}\sum_{i=0}^{t-1} A^i B_\alpha \Sigma B_\alpha^{\mathsf{T}} (A^{\mathsf{T}})^i.$$

Then the tail output has moments

$$\mathbb{E}\bar{u}_{t,m} = \mathsf{L}_m a_t, \qquad \mathrm{Cov}(\bar{u}_{t,m}) = \mathsf{L}_m P_t \mathsf{L}_m^{\mathsf{T}} + \frac{1}{b}\sum_{s=1}^{m} K_{m,s}\Sigma K_{m,s}^{\mathsf{T}}. \tag{95}$$

Indeed, expansion of the recursion gives $\bar{u}_{t,m} = \mathsf{L}_m \mathbf{z}_t + \sum_{s=1}^m K_{m,s}\eta_{t+s}^{(b)}$; the future noises are independent of $\mathbf{z}_t$. This proves both formulas. On the scalar slow mode, with a zero-velocity initial state, the mean is also $e_0 m^{-1}\sum_{k=t}^{T-1} r_k$.

For an accelerated first stage followed by a second method, reset velocity at the switch: the second state is $Je_t = JE_1 z_t$. Put superscript (2) on $\mathsf{L}_m, K_{m,s}$ for the second method, and compute a_t, P_t using the first method. The output then has

$$\mathbb{E}\bar{u}^{(2)} = \mathsf{L}_m^{(2)} JE_1 a_t, \qquad \mathrm{Cov}(\bar{u}^{(2)}) = \mathsf{L}_m^{(2)} JE_1 P_t E_1^{\mathsf{T}} J^{\mathsf{T}} (\mathsf{L}_m^{(2)})^{\mathsf{T}} + \frac{1}{b}\sum_{s=1}^{m} K_{m,s}^{(2)}\Sigma(K_{m,s}^{(2)})^{\mathsf{T}}. \tag{96}$$

The first covariance term is the propagated uncertainty from warm-up; discarding it would grant the second stage an artificial deterministic start. These identities also show why there is no automatic ranking between continuing accelerated tail averaging and switching to SGD.

All three protocols—full averaging, tail averaging, and two stages—are compared at T sequential rounds and $N = bT$ oracle samples. We report both the retained-window reference $V_\star/(bm)$ and the all-sample ERM reference $V_\star/(bT)$. A method calibrated to the former can still lose efficiency relative to the latter by spending too many rounds on warm-up. The paired all-sample discrepancy in the experiments is computed by jointly propagating the output and the sum of all noise inputs, retaining their cross-covariance.

Link with the Scalar Remainder. Tail averaging can reduce deterministic bias. In the same scalar model, the mean of the average starting at round t is $e_0 m^{-1} \sum_{j=t}^{t+m-1} r_j$, so its normalized bias is

$$\frac{\mu e_0}{\sigma}\sqrt{b/m}\sum_{j=t}^{t+m-1} r_j.$$

The infinite tail from Equation (26) provides a sufficient bound with the repeated-root factor included. If all burn-in and retained rounds use the same batch, the full work is $(t+m)b$ samples and $t+m$ rounds; mb counts only retained-window work.

E.4 Choosing among Specified Procedures

In the additive quadratic model, any linear procedure π fixed in advance and started with zero velocity can be written as

$$\widehat{x}_\pi - x_\star = F_\pi e_0 + \sum_{s=1}^{T_\pi} W_{\pi,s}\eta_s^{(b_{\pi,s})}, \qquad V_\pi = \sum_{s=1}^{T_\pi} W_{\pi,s}\Sigma W_{\pi,s}^\mathsf{T}/b_{\pi,s}.$$

The matrices are computed recursively, including the random state at the switching time. Let Π_N be a finite grid, specified in advance, of steps, methods, windows, batches and switching times with $\sum_s b_{\pi,s} = N$ for every candidate. All steps discarded by the averaging are paid for.

Proposition E.10 (Optimal choice within a given grid)**.** *With known H, Σ and the bound $\|e_0\| \le R_0$, put*

$$U_\pi = R_0^2\|F_\pi\|_{\mathrm{op}}^2 + \operatorname{tr} V_\pi, \qquad \pi_\star \in \arg\min_{\pi\in\Pi_N} U_\pi.$$

Then $\sup_{\|e_0\|\le R_0}\mathbb{E}\|\widehat{x}_{\pi_\star} - x_\star\|^2 = \min_{\pi\in\Pi_N}\sup_{\|e_0\|\le R_0}\mathbb{E}\|\widehat{x}_\pi - x_\star\|^2 = U_{\pi_\star}$. Under Gaussian noise, the radius $R_0\|F_\pi\|_{\mathrm{op}} + \mathrm{Quant}_{1-\delta}(0, V_\pi)$ is valid for the same procedure. Minimizing this radius over Π_N solves a separate selection problem for the confidence ball.

Proof of Theorem E.10. The representation preceding the proposition gives

$$\mathbb{E}\|\widehat{x}_\pi - x_\star\|^2 = \|F_\pi e_0\|^2 + \operatorname{tr} V_\pi.$$

The supremum of the first term over $\|e_0\| \le R_0$ equals $R_0^2\|F_\pi\|_{\mathrm{op}}^2$, by a maximizing right singular vector. Thus the worst-case risk is U_π for each candidate. Minimizing over the prescribed finite grid proves the asserted minimax identity within that grid. Under Gaussian noise, the centered random term has law $V_\pi^{1/2}Z$, so the triangle inequality gives the stated radius. The choice of π here depends only on the known geometry and the prescribed grid, not on the training observations; hence this choice does not change its coverage guarantee. □

With known e_0 one can minimize the risk; this is an *oracle* choice for the analysis, since $x_\star$ is usually unknown. The choice via R_0 uses only the stated bound on the start. Thus the burn-in length and the switching time are justified relative to an explicit grid and target metric.

If the parameters are estimated from an independent pilot sample of size N_{pilot}, the main budget equals $N - N_{\mathrm{pilot}}$. On the simultaneous event $\max_{\pi\in\Pi_N}|\widehat{U}_\pi - U_\pi| \le \epsilon$, the minimizer of $\widehat{U}$ satisfies $U_{\widehat{\pi}} \le \min_\pi U_\pi + 2\epsilon$. This follows from two applications of the error bound and the definition of the minimizer.

F Heavy Tails and Clipping

For a robust online procedure, the link between the optimization error and the distribution of the output has to be checked separately. Clipping changes the oracle, so its effect on the optimization guarantee, on the bias and on the distribution of the output should be analyzed jointly. We first note what heavy tails do inside the stochastically linear class, and then how clipping leaves it.

F.1 Finite Variance and High Confidence Levels

Remark F.1 (Heavy tails inside the stochastically linear class). In the additive quadratic model $\hat{x}^{\mathrm{ERM}}_{t,m,b} - x_\star = -H^{-1}\bar{\eta}^{(b)}_{t,m}$, so every deviation bound for the empirical noise mean—Gaussian, Fuk–Nagaev or stable—applies verbatim to the offline estimator, and by Theorem A.2 the online output adds only the bridge correction $H^{-1}\bar{g}_{t,m}$. Within this class, heavy tails therefore degrade the statistical term identically in both languages; the languages differ in the robustifications they admit (Section F.4).

A Map of Moment Regimes. The available guarantees depend on the moments of the noise. With a finite third moment, Theorem B.7 gives a quantitative independent normal approximation. With a finite variance but no third moment, Theorem B.8 still gives a qualitative CLT with centering and variance; the Nagaev bound below additionally requires a moment of some order $r_{\mathrm{tail}} > 2$, which does not follow from a finite variance alone, and uniformly over the class no sub-Gaussian radius is available for the ERM or the averaged outputs (Theorem F.2). With infinite variance, the stable limit transfers under a domain-of-attraction assumption and the correct normalization, strictly stable noise admits an stable law and calibration criteria (Theorem F.4), and the Gaussian covariance and the MSE may not exist.

F.1.1 Transfer of the Nagaev Bound

Let $a \in \mathbb{R}^d$, $Y_a = a^\top H^{-1}\xi$, and assume $\mathbb{E}Y_a = 0$, $\mathbb{E}Y_a^2 = \sigma_a^2 > 0$, and $\mathbb{E}|Y_a|^{r_{\mathrm{tail}}} < \infty$ for some $r_{\mathrm{tail}} > 2$. For independent copies $Y_{a,1}, \dots, Y_{a,N}$, a Fuk–Nagaev inequality gives constants $c_{r_{\mathrm{tail}}}, C_{r_{\mathrm{tail}}} > 0$, depending only on r_{tail}, such that

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Y_{a,i}\right| > x\right) \le 2\exp\left(-c_{r_{\mathrm{tail}}}\frac{Nx^2}{\sigma_a^2}\right) + \frac{C_{r_{\mathrm{tail}}}\mathbb{E}|Y_a|^{r_{\mathrm{tail}}}}{N^{r_{\mathrm{tail}}-1}x^{r_{\mathrm{tail}}}}, \qquad x > 0. \tag{97}$$

See Fuk and Nagaev (1971); Nagaev (1979, 1998). The zero-variance case is trivial.

The decomposition Equation (15) and a union bound imply, for every $x, r_\partial > 0$ and $N = mb$,

$$\mathbb{P}(|a^\top \bar{u}_{t,m}| > x + r_\partial) \le 2\exp\left(-c_{r_{\mathrm{tail}}}\frac{Nx^2}{\sigma_a^2}\right) + \frac{C_{r_{\mathrm{tail}}}\mathbb{E}|Y_a|^{r_{\mathrm{tail}}}}{N^{r_{\mathrm{tail}}-1}x^{r_{\mathrm{tail}}}} + \mathbb{P}\left(\frac{|a^\top H^{-1}\Delta_{t,m}|}{\alpha m} > r_\partial\right). \tag{98}$$

A sufficient empirical-mean radius at confidence $1-\delta$ is

$$C'_{r_{\mathrm{tail}}}\left[\sigma_a\sqrt{\frac{\log(4/\delta)}{N}} + \left(\frac{\mathbb{E}|Y_a|^{r_{\mathrm{tail}}}}{\delta N^{r_{\mathrm{tail}}-1}}\right)^{1/r_{\mathrm{tail}}}\right]. \tag{99}$$

In particular, the boundary can be correlated with and can partly cancel the empirical-noise term. Robust nonlinear estimators under finite covariance can also have better high-confidence behavior than an empirical mean; (Gorbunov et al., 2020; Puchkin et al., 2024).

F.1.2 The High-Confidence Limit of the Empirical Mean

The Nagaev radius Equation (99) contains a term polynomial in $1/\delta$. The construction is the classical one of Catoni (2012).

Proposition F.2 (High-confidence limit of the empirical mean and of averaged outputs)**.** *Consider the scalar additive quadratic model with curvature $\mu > 0$ and centered i.i.d. noise of variance σ^2, and let $N = mb$.*

1. *For every $N \ge 1$ and $\delta \in (0, 1/8]$ there is a symmetric noise law with variance σ^2 such that the ERM of the N observations satisfies*
$$\mathbb{P}\left\{|\hat{x}^{\mathrm{ERM}} - x_\star| \ge \frac{\sigma}{\mu\sqrt{2N\delta}}\right\} \ge \delta.$$
2. *Let $\bar{u}_{0,m}$ be the full average of SGD, HB or NAG with the standard parameters of Theorem B.2, started at $e_0 = 0$, and let $m \ge 4S_\infty$. For every $\delta \in (0, 1/16]$ there is a symmetric noise law with variance σ^2 such that*
$$\mathbb{P}\left\{|\bar{u}_{0,m}| \ge \frac{\sigma}{4\mu\sqrt{N\delta}}\right\} \ge \delta.$$

Indeed, let the noise take the values $\pm\omega_{\mathrm{spike}}$ with probability $\pi_N/2$ each and the value 0 otherwise, with $\omega_{\mathrm{spike}} = \sigma/\sqrt{\pi_N}$. The event that one of the N noises is nonzero has probability $N\pi_N(1-\pi_N)^{N-1}$, and on it the ERM

error equals $\omega_{\text{spike}}/(\mu N)$ in absolute value. With $\pi_N = 2\delta/N$ this equals $\sigma/(\mu\sqrt{2N\delta})$, while $N\pi_N(1-\pi_N)^{N-1} = 2\delta(1-2\delta/N)^{N-1} \geq 2\delta e^{-4\delta} \geq \delta$ for $\delta \leq 1/8$. For the averaged outputs, Equation (22) with $e_0 = 0$ gives $\bar{u}_{0,m} = -(\mu N)^{-1}\sum_{s,i} w_{m-s}\xi_{s,i}$ with $w_j = 1 - r_j \in [0,1]$. Since $r_j \geq 0$ and $\sum_j r_j = S_\infty$, fewer than $2S_\infty$ indices satisfy $r_j > 1/2$; for $m \geq 4S_\infty$ at least $N/2$ of the N noises therefore carry a weight of at least $1/2$. With $\pi_N = 4\delta/N$, the event that one noise is nonzero and that it carries such a weight has probability at least $(N/2)\pi_N(1-\pi_N)^{N-1} \geq 2\delta e^{-8\delta} \geq \delta$ for $\delta \leq 1/16$, and on it $|\bar{u}_{0,m}| \geq \omega_{\text{spike}}/(2\mu N) = \sigma/(4\mu\sqrt{N\delta})$.

The Gaussian radius $\Phi^{-1}(1-\delta/2)\sigma/(\mu\sqrt{N})$ is of order $\sigma\sqrt{\log(1/\delta)}/(\mu\sqrt{N})$, so the ratio of the two radii grows like $(\delta\log(1/\delta))^{-1/2}$ as $\delta \to 0$. No ball of sub-Gaussian radius centered at the ERM or at the averaged output is therefore valid uniformly over the finite-variance class, whatever Fisher or Wilks expansion is used to compute it. A one-pass clipped estimator does not have this obstruction: the radius of Theorem F.6 depends on δ only through $\sqrt{\log(2d/\delta)}$ under a finite variance alone. This is the precise sense in which heavy tails separate the offline empirical-risk description from an online procedure that is allowed to clip. The lower bound concerns uniformity over the class: for a fixed noise law and a fixed δ, the normal approximations of Section B.2 remain valid as $N \to \infty$.

F.2 Infinite Variance

A stable domain of attraction generally requires a slowly varying factor in addition to the power of the sample size (Samorodnitsky and Taqqu, 1994). For SGD with infinite noise variance, α-stable limits of Polyak–Ruppert averages were obtained by Wang et al. (2021).

Proposition F.3 (Stable-law transfer with the correct normalizer)**.** *Suppose the centered i.i.d. single-sample noises satisfy, for positive constants c_N,*

$$c_N^{-1}\sum_{i=1}^{N}\xi_i \xrightarrow{\text{d}} Z_{\alpha_0}, \qquad 1 < \alpha_0 < 2,$$

where Z_{α_0} is an α_0-stable vector. In a stable domain of attraction one may take $c_N = N^{1/\alpha_0}\mathcal{L}_{\text{sv}}(N)$ with $\mathcal{L}_{\text{sv}}$ slowly varying. For $N = mb \to \infty$, if

$$\frac{N}{\alpha m c_N}\Delta_{t,m} \xrightarrow{\mathbb{P}} 0, \tag{100}$$

then the additive quadratic algorithm satisfies

$$\frac{N}{c_N}\bar{u}_{t,m} \xrightarrow{\text{d}} -H^{-1}Z_{\alpha_0}. \tag{101}$$

Proof of Theorem F.3. Enumerate the $N = mb$ individual noises in the retained window. Multiplying Equation (15) by N/c_N gives

$$\frac{N}{c_N}\bar{u}_{t,m} = -H^{-1}\frac{1}{c_N}\sum_{i=1}^{N}\xi_i + H^{-1}\frac{N}{\alpha m c_N}\Delta_{t,m}.$$

The first term has the stable limit by the domain-of-attraction assumption. With the fixed Hessian of the additive model, the second term converges to zero in probability by Equation (100). Slutsky's theorem proves Equation (101). □

The normalization $c_N = N^{1/\alpha_0}$ is the special case of normal attraction, not every stable domain of attraction. In that case, for fixed algorithmic gain α, the boundary condition becomes $b^{1-1/\alpha_0}m^{-1/\alpha_0}\Delta_{t,m} \xrightarrow{\mathbb{P}} 0$. If the driving noise is strictly α_0-stable and the linear chain is stationary, each state scales in distribution as b^{1/α_0-1} times its single-sample counterpart. Indeed, independent batching has this scaling and the stationary state is the convergent stable linear filter $\sum_{j\geq 0} A^j B_\alpha \eta_{-j}^{(b)}$; stability makes its coefficient α_0-powers summable. Tightness of its endpoint marginals then makes the normalized stochastic boundary $O_{\mathbb{P}}(m^{-1/\alpha_0})$.

Calibration under Strictly Stable Noise. When the noise is strictly stable, the cold-start law of Theorem B.1 remains exact, and the calibration analysis of Section B.1 carries over with the exponent 2 replaced by α_0.

Proposition F.4 (Stable calibration of the averaged output)**.** *In the scalar model of Theorem B.1, let the single-sample noises be i.i.d. symmetric α_0-stable with $1 < \alpha_0 < 2$ and scale σ_{α_0}, that is, $\mathbb{E}e^{is\xi} = \exp(-\sigma_{\alpha_0}^{\alpha_0}|s|^{\alpha_0})$, and let $Z_{\alpha_0}^{\text{st}}$ be standard symmetric α_0-stable. Put $N = mb$, $w_j = 1 - r_j$ and*

$$\theta_{\alpha_0} = \frac{\mu e_0}{\sigma_{\alpha_0}}, \qquad h_m^{(\alpha_0)} = \theta_{\alpha_0} S_m\, b^{1-1/\alpha_0} m^{-1/\alpha_0}, \qquad v_m^{(\alpha_0)} = \frac{1}{m}\sum_{j=0}^{m-1}|w_j|^{\alpha_0}.$$

Then, for all $m, b \geq 1$ and any two-step parameters,

$$\frac{\mu N^{1-1/\alpha_0}}{\sigma_{\alpha_0}} \bar{u}_{0,m} \stackrel{\mathrm{d}}{=} h_m^{(\alpha_0)} + \big(v_m^{(\alpha_0)}\big)^{1/\alpha_0} Z_{\alpha_0}^{\mathrm{st}}, \tag{102}$$

whereas the ERM of the same N observations satisfies $\mu N^{1-1/\alpha_0}(\hat{x}^{\mathrm{ERM}} - x_\star)/\sigma_{\alpha_0} \stackrel{\mathrm{d}}{=} Z_{\alpha_0}^{\mathrm{st}}$. For the standard SGD, HB and NAG parameters of Theorem B.2, along any sequence with $m \to \infty$, integer $b \geq 1$, arbitrary $\kappa \geq 1$ and arbitrary signed θ_{α_0},

$$v_m^{(\alpha_0)} \to 1 \quad \Longleftrightarrow \quad S_\infty/m \to 0, \tag{103}$$

$$\big(h_m^{(\alpha_0)}, v_m^{(\alpha_0)}\big) \to (0,1) \quad \Longleftrightarrow \quad S_\infty/m \to 0 \text{ and } |\theta_{\alpha_0}| S_\infty b^{1-1/\alpha_0} m^{-1/\alpha_0} \to 0. \tag{104}$$

Indeed, for symmetric strictly stable noise and deterministic coefficients ω_i^{st}, $\sum_i \omega_i^{\mathrm{st}} \xi_i \stackrel{\mathrm{d}}{=} \sigma_{\alpha_0} (\sum_i |\omega_i^{\mathrm{st}}|^{\alpha_0})^{1/\alpha_0} Z_{\alpha_0}^{\mathrm{st}}$. Applied to Equation (22), where each weight $w_{m-s}/(\mu N)$ appears b times, this gives the scale

$$\frac{\sigma_{\alpha_0} (bm\, v_m^{(\alpha_0)})^{1/\alpha_0}}{\mu N} = \frac{\sigma_{\alpha_0}}{\mu} \big(v_m^{(\alpha_0)}\big)^{1/\alpha_0} N^{1/\alpha_0 - 1},$$

and the deterministic part $e_0 S_m/m$ becomes $h_m^{(\alpha_0)}$ after normalization; the ERM is the same computation with unit weights. For the standard parameters $0 \leq r_j \leq 1$, and for $r \in [0,1]$ one has $r \leq 1 - (1-r)^{\alpha_0} \leq \alpha_0 r$. Hence $S_m/m \leq 1 - v_m^{(\alpha_0)} \leq \alpha_0 S_m/m$, which gives Equation (103) as for Equation (29), while Equation (104) follows because $S_m = S_\infty(1 + o(1))$ when $S_\infty/m \to 0$.

Consequently, at $b = 1$ and fixed $\theta_{\alpha_0} \neq 0$, calibrated inference from the full average requires $m \gg (|\theta_{\alpha_0}| S_\infty)^{\alpha_0}$, that is, $m \gg \kappa^{\alpha_0}$ for SGD and $m \gg \kappa^{\alpha_0/2}$ for HB and NAG, up to the factor $|\theta_{\alpha_0}|^{\alpha_0}$. At a fixed budget N and with $bS_\infty/N \to 0$, the start-induced critical batch becomes

$$b_{\mathrm{crit}}^{(\alpha_0)}(N) \asymp \frac{N^{1/\alpha_0}}{|\theta_{\alpha_0}| S_\infty}, \tag{105}$$

which reduces to Equation (38) at $\alpha_0 = 2$; momentum again enlarges it by a factor of order $\sqrt{\kappa}$. Heavier tails lower the onset and raise the critical batch: the statistical scale $N^{1/\alpha_0 - 1}$ decays more slowly, so a fixed start is hidden sooner, although the accuracy is worse. The optimization description is silent in this regime: the mean squared error is infinite, so guarantees in mean square carry no information, whereas Equation (102) yields calibrated confidence intervals from stable quantiles. These intervals require the index α_0 and the scale σ_{α_0}, whose estimation is a separate problem; for noise in a domain of attraction, Theorem F.3 gives the corresponding asymptotic statement.

Conjecture F.5 (Stiff mode of HB under stable noise)**.** *On the stiffest direction, the relative scale defect of optimally tuned HB is $v_m^{(\alpha_0)} - 1 = m^{-1} \sum_{j<m} (|1 - r_j|^{\alpha_0} - 1)$ with the residual $r_j = (-\varrho)^j \{1 + (1+\varrho)j\}$ of Theorem B.9. The sum is dominated by the indices with $|r_j| \gg 1$, and replacing it by the corresponding integral suggests*

$$v_m^{(\alpha_0)} - 1 \sim C_{\alpha_0} \frac{\kappa^{(1+\alpha_0)/2}}{m}, \qquad C_{\alpha_0} = \frac{2^{\alpha_0} \Gamma(\alpha_0 + 1)}{(2\alpha_0)^{\alpha_0 + 1}},$$

when $\kappa \to \infty$ and $m/\sqrt{\kappa} \to \infty$. At $\alpha_0 = 2$ this reproduces the coefficient $\kappa^{3/2}/(8m)$ of Equation (48). The conjecture predicts persistence of the obstruction under infinite variance, with the exponent $3/2$ replaced by $(1+\alpha_0)/2$; a complete proof requires control of the oscillating lower-order terms.

A nonlinear stable PR theorem would additionally require control of its remainders at scale N/c_N.

F.3 A Fully Specified Finite-Variance Robust Benchmark

For a threshold $\tau_{\mathrm{clip}} > 0$, write $\mathrm{clip}_{\tau_{\mathrm{clip}}}(v) = v \min\{1, \tau_{\mathrm{clip}}/\|v\|\}$ for $v \neq 0$ and $\mathrm{clip}_{\tau_{\mathrm{clip}}}(0) = 0$.

We first give a fully specified clipping guarantee. Consider the additive quadratic oracle with known $H \succ 0$. Fix a query x_0 before observing the samples, and suppose known bounds satisfy $\|x_0 - x_\star\| \leq R_{\mathrm{init}}$, $\|H\| \leq L$, and $\mathbb{E}\|\xi\|^2 \leq \nu_2$. Let $M_{\mathrm{clip}} = L^2 R_{\mathrm{init}}^2 + \nu_2 > 0$. For $N = mb$ independent observations, clip each full gradient before averaging within a batch:

$$Y_{k,j} = G(x_0, \zeta_{k,j}), \qquad \widetilde{Y}_{k,j}^{\mathrm{cl}} = \mathrm{clip}_{\tau_{\mathrm{clip}}}(Y_{k,j}), \qquad \widehat{x}_{\mathrm{clip}} = x_0 - H^{-1} \frac{1}{mb} \sum_{k=1}^{m} \sum_{j=1}^{b} \widetilde{Y}_{k,j}^{\mathrm{cl}}.$$

This is a streaming estimator with one running vector sum; an offline implementation on the same observations returns the same output. It uses N oracle calls and one final linear solve.

Proposition F.6 (Finite-budget coverage after full-gradient clipping)**.** *Assume only independent centered oracle noise with the finite second moment above. For* $0 < \delta < 1$*, let* $\Lambda_\delta = \log(2d/\delta)$*. For any* $\tau_{\mathrm{clip}} > 0$*, with probability at least* $1 - \delta$*,*

$$\|H(\widehat{x}_{\mathrm{clip}} - x_\star)\| \le \frac{M_{\mathrm{clip}}}{\tau_{\mathrm{clip}}} + \sqrt{d}\left\{\sqrt{\frac{2\nu_2\Lambda_\delta}{N}} + \frac{4\tau_{\mathrm{clip}}\Lambda_\delta}{3N}\right\}. \tag{106}$$

In particular, the choice $\tau_{\mathrm{clip}} = \sqrt{M_{\mathrm{clip}}N/\Lambda_\delta}$ *gives*

$$\mathbb{P}\left\{\|\widehat{x}_{\mathrm{clip}} - x_\star\| \le \frac{4}{\mu}\sqrt{\frac{dM_{\mathrm{clip}}\log(2d/\delta)}{N}}\right\} \ge 1 - \delta, \qquad \mu = \lambda_{\min}(H). \tag{107}$$

For $M_{\mathrm{clip}} = 0$ *the oracle is deterministic and* $x_0 = x_\star$*, so the zero-radius statement is immediate.*

Proof of Theorem F.6. Write $g = H(x_0 - x_\star)$, $\tau = \tau_{\mathrm{clip}}$, and $T_\tau(y) = \mathrm{clip}_\tau(y)$. For an individual answer $Y = g + \xi$, $\mathbb{E}\|Y\|^2 = \|g\|^2 + \mathbb{E}\|\xi\|^2 \le M_{\mathrm{clip}}$, and

$$\|\mathbb{E}T_\tau(Y) - g\| \le \mathbb{E}\|Y - T_\tau(Y)\| \le \mathbb{E}[\|Y\|\mathbf{1}_{\{\|Y\|>\tau\}}] \le M_{\mathrm{clip}}/\tau.$$

The clipping map is Euclidean projection onto the closed ball of radius τ, hence it is nonexpansive. If Y' is an independent copy of Y, then for each coordinate i,

$$\mathrm{Var}((T_\tau(Y))_i) = \tfrac{1}{2}\mathbb{E}[(T_\tau(Y))_i - (T_\tau(Y'))_i]^2 \le \tfrac{1}{2}\mathbb{E}\|Y - Y'\|^2 = \mathbb{E}\|\xi\|^2 \le \nu_2.$$

The centered clipped coordinate is bounded in absolute value by 2τ. Bernstein's inequality and a union bound over both tails and all d coordinates imply, with probability at least $1 - \delta$,

$$\left\|\frac{1}{N}\sum_{\ell=1}^{N}\{T_\tau(Y_\ell) - \mathbb{E}T_\tau(Y_\ell)\}\right\| \le \sqrt{d}\left\{\sqrt{\frac{2\nu_2\Lambda_\delta}{N}} + \frac{4\tau\Lambda_\delta}{3N}\right\}.$$

Since $H(\widehat{x}_{\mathrm{clip}} - x_\star) = g - N^{-1}\sum_\ell T_\tau(Y_\ell)$, adding the bias bound proves Equation (106). Substituting the stated threshold, using $\nu_2 \le M_{\mathrm{clip}}$ and $d \ge 1$, bounds its right-hand side by

$$(1 + \sqrt{2} + 4/3)\sqrt{dM_{\mathrm{clip}}\Lambda_\delta/N} < 4\sqrt{dM_{\mathrm{clip}}\Lambda_\delta/N}.$$

Finally $\|H^{-1}\| = 1/\mu$ proves Equation (107). The case $M_{\mathrm{clip}} = 0$ is deterministic as stated. □

The benchmark closes a finite-variance, finite-budget comparison for the same target and data stream.

F.4 Clipping along a Moving Trajectory

A finite third absolute moment permits the quantitative independent-weight bounds in Section B.2. Finite variance without a third moment can still permit a qualitative CLT, but not that Berry–Esseen estimate. With infinite variance, the original $V_\star$ and MSE may not exist.

For the moving-query recursion, specify $b = 1$ and clip the *full gradient*, not an unobservable isolated noise vector. For a threshold $\tau_{\mathrm{cl}} > 0$, define $\mathrm{clip}_{\tau_{\mathrm{cl}}}(v) = v\min\{1, \tau_{\mathrm{cl}}/\|v\|\}$ when $v \ne 0$, and zero when $v = 0$. Let $\tau_{k,\mathrm{cl}} > 0$ be predictable and set

$$g_k^{\mathrm{cl}} = \mathrm{clip}_{\tau_{k,\mathrm{cl}}}(Hu_k + \xi_{k+1}), \qquad \mathfrak{b}_k^{\mathrm{cl}} = \mathbb{E}[g_k^{\mathrm{cl}} \mid \mathcal{F}_k] - Hu_k, \qquad \varepsilon_{k+1}^{\mathrm{cl}} = g_k^{\mathrm{cl}} - \mathbb{E}[g_k^{\mathrm{cl}} \mid \mathcal{F}_k]. \tag{108}$$

Thus $\mathfrak{b}_k^{\mathrm{cl}}$ is predictable and $\mathbb{E}[\varepsilon_{k+1}^{\mathrm{cl}} \mid \mathcal{F}_k] = 0$. The recursion is $e_{k+1} = e_k + \beta(e_k - e_{k-1}) - \alpha g_k^{\mathrm{cl}}$, with the same query definition as before. Summing it yields the exact pathwise identity

$$\sqrt{m}\bar{u}_{0,m} = -\frac{H^{-1}}{\sqrt{m}}\sum_{k=0}^{m-1}\varepsilon_{k+1}^{\mathrm{cl}} - \frac{H^{-1}}{\sqrt{m}}\sum_{k=0}^{m-1}\mathfrak{b}_k^{\mathrm{cl}} + \frac{H^{-1}\Delta_{0,m}}{\alpha\sqrt{m}}. \tag{109}$$

It exposes the accumulated clipping bias as well as the algorithmic boundary.

Theorem F.7 (Conditional clipped-gradient CLT)**.** *Fix $H \succ 0$ and the parameters α, β, γ. Consider clipped recursions of length m, allowing thresholds, initial states, and hence trajectories to depend on m; the row index is suppressed. Assume the clipped gradients are integrable and the martingale differences in Equation* (108) *are square-integrable, and*

$$\frac{1}{\sqrt{m}} \sum_{k=0}^{m-1} \mathfrak{b}_k^{\mathrm{cl}} \xrightarrow{\mathbb{P}} 0, \qquad \frac{\Delta_{0,m}}{\alpha\sqrt{m}} \xrightarrow{\mathbb{P}} 0, \tag{110}$$

$$\frac{1}{m} \sum_{k=0}^{m-1} \mathbb{E}[\varepsilon_{k+1}^{\mathrm{cl}} (\varepsilon_{k+1}^{\mathrm{cl}})^{\mathsf{T}} \mid \mathcal{F}_k] \xrightarrow{\mathbb{P}} \Sigma_{\mathrm{cl}}, \tag{111}$$

where Σ_{cl} is deterministic and positive semidefinite. For every $\varepsilon_{\mathrm{Lin}} > 0$, assume conditional Lindeberg:

$$\frac{1}{m} \sum_{k=0}^{m-1} \mathbb{E}\Big[\|\varepsilon_{k+1}^{\mathrm{cl}}\|^2 \mathbf{1}_{\{\|\varepsilon_{k+1}^{\mathrm{cl}}\| > \varepsilon_{\mathrm{Lin}}\sqrt{m}\}} \mid \mathcal{F}_k\Big] \xrightarrow{\mathbb{P}} 0. \tag{112}$$

Then

$$\sqrt{m}\bar{u}_{0,m} \xrightarrow{\mathrm{d}} \mathcal{N}(0, H^{-1}\Sigma_{\mathrm{cl}}H^{-1}).$$

The original PR covariance is recovered only if $\Sigma_{\mathrm{cl}} = \Sigma$ with finite Σ.

Proof of Theorem F.7. For any fixed vector a, the variables $a^{\mathsf{T}}\varepsilon_{k+1}^{\mathrm{cl}}/\sqrt{m}$ form a martingale triangular array. Its predictable quadratic variation converges to $a^{\mathsf{T}}\Sigma_{\mathrm{cl}}a$ by Equation (111). The scalar conditional Lindeberg condition follows from Equation (112) and $|a^{\mathsf{T}}\varepsilon| \le \|a\|\|\varepsilon\|$. The martingale CLT and Cramér–Wold therefore give

$$m^{-1/2} \sum_{k=0}^{m-1} \varepsilon_{k+1}^{\mathrm{cl}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_{\mathrm{cl}}).$$

Both other terms in Equation (109) vanish in probability by Equation (110), since H and the method parameters are fixed. Multiplication by $-H^{-1}$ and Slutsky's theorem prove the conclusion, including the possibly singular Gaussian case. □

The conditions are checkable targets. Two elementary bounds illustrate their distinct roles. With centered raw oracle noise and moment order $r_{\mathrm{mom}} > 1$,

$$\|\mathfrak{b}_k^{\mathrm{cl}}\| \le \mathbb{E}[\|Hu_k + \xi_{k+1}\| \mathbf{1}_{\{\|Hu_k+\xi_{k+1}\| > \tau_{k,\mathrm{cl}}\}} \mid \mathcal{F}_k] \le \frac{\mathbb{E}[\|Hu_k + \xi_{k+1}\|^{r_{\mathrm{mom}}} \mid \mathcal{F}_k]}{\tau_{k,\mathrm{cl}}^{r_{\mathrm{mom}}-1}}. \tag{113}$$

Indeed, clipping changes a vector by norm $(\|v\| - \tau_{\mathrm{cl}})_+$. Also $\|\varepsilon_{k+1}^{\mathrm{cl}}\| \le 2\tau_{k,\mathrm{cl}}$; hence the envelope $\tau_{\max,m} := \max_{k<m} \tau_{k,\mathrm{cl}} = o(\sqrt{m})$ ensures Lindeberg. For example, if the raw gradients have uniformly bounded r_{mom}th moments for $r_{\mathrm{mom}} > 2$, using $\tau_{k,\mathrm{cl}} = m^{\omega_{\mathrm{cl}}}$ with $1/[2(r_{\mathrm{mom}} - 1)] < \omega_{\mathrm{cl}} < 1/2$ makes the accumulated bias vanish in L_1 and ensures Lindeberg.

Full-gradient clipping depends on the current state, so the independent deterministic-weight representation from Section B.2 cannot be imported into this theorem. For $b > 1$ one distinguishes

$$\frac{1}{b}\sum_{j=1}^{b} \mathrm{clip}_{\tau_{\mathrm{cl}}}(G_{k,j}) \quad \text{and} \quad \mathrm{clip}_{\tau_{\mathrm{cl}}}\left(\frac{1}{b}\sum_{j=1}^{b} G_{k,j}\right).$$

They have different predictable biases and conditional covariances. At a fixed threshold, asymmetric noise can give $\mathfrak{b}_k^{\mathrm{cl}} \neq 0$ even at $u_k = 0$, changing the target. In the scalar location model, clipping an individual quadratic-loss gradient is the derivative of a Huber loss, so an offline robust estimating equation provides a natural comparison. Theorem F.8 makes precise when such a potential exists in general. Finite-time coverage, a verified accelerated boundary bound, and comparison of online and offline procedures at the same target and observation budget require further analysis.

Remark F.8. The clipped martingale difference $\varepsilon_{k+1}^{\mathrm{cl}}$ in Equation (108) depends on the query point whenever clipping is active with positive probability. A potential exists in two important cases. With the Huber function $\mathcal{H}_{\tau_{\mathrm{H}}}(r) = r^2/2$ for $|r| \le \tau_{\mathrm{H}}$ and $\mathcal{H}_{\tau_{\mathrm{H}}}(r) = \tau_{\mathrm{H}}|r| - \tau_{\mathrm{H}}^2/2$ otherwise, the vector location loss $\ell(x, z) = \frac{1}{2}\|x - z\|^2$ gives $\mathrm{clip}_{\tau_{\mathrm{H}}}(x-z) = \nabla_x \mathcal{H}_{\tau_{\mathrm{H}}}(\|x-z\|)$, and least squares, $\ell(x; a_{\mathrm{feat}}, y) = \frac{1}{2}(a_{\mathrm{feat}}^{\mathsf{T}}x - y)^2$, gives $\mathrm{clip}_{\tau_{\mathrm{H}}}\{a_{\mathrm{feat}}(a_{\mathrm{feat}}^{\mathsf{T}}x - y)\} =$

$\nabla_x \mathcal{H}_{\tau_{\mathrm{H}}/\|a_{\mathrm{feat}}\|}(a_{\mathrm{feat}}^{\mathsf{T}} x - y)$, a Huber loss with a design-dependent threshold. In both cases, the offline counterpart is a robust M-estimator. For the adaptive Huber estimator, whose threshold grows with the sample size, sub-Gaussian deviation bounds and a nonasymptotic Bahadur representation hold under finite second moments (Zhou et al., 2018); regression versions are available under weaker moment assumptions (Sun et al., 2020). In general, for $g = \nabla_x \ell$, $r = \|g\|$ and $\omega_{\tau_{\mathrm{H}}}(r) = \min\{1, \tau_{\mathrm{H}}/r\}$, the clipped field $\omega_{\tau_{\mathrm{H}}}(r) g$ has Jacobian $\omega_{\tau_{\mathrm{H}}} \nabla_x^2 \ell + \omega'_{\tau_{\mathrm{H}}}(r) r^{-1} g g^{\mathsf{T}} \nabla_x^2 \ell$; where clipping is active, this matrix is symmetric if and only if g is an eigenvector of $\nabla_x^2 \ell$. Otherwise the clipped field has no potential, and the natural offline counterpart is the estimating equation $N^{-1} \sum_{i=1}^{N} \mathrm{clip}_{\tau_{\mathrm{H}}}(\nabla_x \ell(x, Z_i)) = 0$, i.e., a Z-estimator (van der Vaart, 1998), rather than a clipped empirical risk.

G Additional Model Extensions

This section delineates the scope of the main finite-budget theory under broader stochastic dynamics. It treats multiplicative noise, decreasing steps and changing momentum, and Richardson–Romberg extrapolation, while stating explicitly which conclusions are unconditional and which require additional moment or stability assumptions.

G.1 Multiplicative Noise

Remark G.1 (Multiplicative linear noise). For general LSA, random matrices multiply the current state. Then a fixed step can contribute a step-dependent covariance correction, as in Mou et al. (2020b). Our universality theorem is for additive gradient noise and for decreasing gains that localize the multiplicative component. This boundary is important: it prevents the false conclusion that momentum is covariance-neutral in every linear stochastic approximation problem.

The Oracle and Covariance Correction. The additive model has the same oracle covariance at the minimizer and at every query point. To separate these quantities, consider $b = 1$ and

$$G(x_\star + u, \zeta_{k+1}) = Hu + \xi_{k+1} + \mathcal{M}_{k+1} u, \qquad \mathbb{E}\xi_{k+1} = 0, \quad \mathbb{E}\mathcal{M}_{k+1} = 0, \tag{114}$$

where the pairs $(\xi_{k+1}, \mathcal{M}_{k+1})$ are i.i.d. and independent of the past $\mathcal{F}_k$, and u_k is $\mathcal{F}_k$-measurable. Within a pair ξ and $\mathcal{M}$ need not be independent. Assume their second moments are finite. For the two-step recursion driven by this oracle, summation gives

$$\sqrt{m}\bar{u}_{0,m} = -\frac{H^{-1}}{\sqrt{m}} \sum_{k=0}^{m-1} \xi_{k+1} - \frac{H^{-1}}{\sqrt{m}} \sum_{k=0}^{m-1} \mathcal{M}_{k+1} u_k + \frac{H^{-1}\Delta_{0,m}}{\alpha\sqrt{m}}. \tag{115}$$

The martingale difference is $\xi_{k+1} + \mathcal{M}_{k+1} u_k$, whose conditional covariance is

$$\Sigma_{\mathrm{traj}}(u_k) = \Sigma + \mathbb{E}[\xi u_k^{\mathsf{T}} \mathcal{M}^{\mathsf{T}} + \mathcal{M} u_k \xi^{\mathsf{T}} \mid \mathcal{F}_k] + \mathbb{E}[\mathcal{M} u_k u_k^{\mathsf{T}} \mathcal{M}^{\mathsf{T}} \mid \mathcal{F}_k]. \tag{116}$$

Here the conditional expectations integrate a pair $(\xi, \mathcal{M})$. If the average of $\Sigma_{\mathrm{traj}}(u_k)$ converges in probability to a deterministic Σ_{eff}, conditional Lindeberg holds, and the scaled boundary vanishes, the martingale CLT gives covariance $H^{-1}\Sigma_{\mathrm{eff}} H^{-1}$.

Writing $\|\mathcal{M}\|$ for the operator norm, the sufficient localization condition

$$\frac{1}{m} \sum_{k=0}^{m-1} \mathbb{E}\|u_k\|^2 \longrightarrow 0$$

makes the extra term in Equation (115) negligible in L^2: martingale orthogonality bounds its squared norm before the factor H^{-1} by $\mathbb{E}\|\mathcal{M}\|^2 \, m^{-1} \sum_k \mathbb{E}\|u_k\|^2$. At a fixed step with nonzero additive noise, such localization need not hold. Consequently there is no general fixed-step covariance neutrality for momentum in this model.

Proposition G.2 (A scalar fixed-step covariance correction)**.** *Consider stationary SGD*

$$e_{k+1} = (1 - \alpha(\lambda_0 + \mathcal{M}_{k+1}))e_k - \alpha\xi_{k+1}, \qquad \lambda_0 > 0,$$

where the i.i.d. centered pairs have independent components, with $\mathbb{E}\mathcal{M}^2 = \sigma_{\mathcal{M}}^2$ *and* $\mathbb{E}\xi^2 = \sigma^2 > 0$*. Suppose*

$$(1 - \alpha\lambda_0)^2 + \alpha^2 \sigma_{\mathcal{M}}^2 < 1.$$

There is a causal stationary solution with finite second moment

$$P = \mathbb{E}e_k^2 = \frac{\alpha\sigma^2}{2\lambda_0 - \alpha(\lambda_0^2 + \sigma_{\mathcal{M}}^2)}.$$

For this solution,

$$\sqrt{m}\bar{e}_{0,m} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\sigma^2 + \sigma_{\mathcal{M}}^2 P}{\lambda_0^2}\right). \tag{117}$$

In particular the variance exceeds σ^2/λ_0^2 when $\sigma_{\mathcal{M}}^2 > 0$.

Proof of Theorem G.2. Put $a_k = 1 - \alpha(\lambda_0 + \mathcal{M}_k)$ and $q_{\mathrm{ms}}^2 = \mathbb{E}a_k^2 < 1$. Extend the iid pairs to all integer times. The series

$$e_k = -\alpha \sum_{j=0}^{\infty} \left(\prod_{r=0}^{j-1} a_{k-r} \right) \xi_{k-j}$$

converges in L^2: the L^2 norm of its jth term is $\alpha\sigma q_{\mathrm{ms}}^j$. Empty products equal one. It defines a causal stationary solution. As a measurable function of the shifted iid sequence, this solution is ergodic. pairs are independent of e_k and have independent centered components; taking first and second moments of the recursion gives $\mathbb{E}e_k = 0$ and

$$P = q_{\mathrm{ms}}^2 P + \alpha^2\sigma^2, \qquad P = \frac{\alpha\sigma^2}{2\lambda_0 - \alpha(\lambda_0^2 + \sigma_{\mathcal{M}}^2)}.$$

Define the effective martingale difference $\zeta_{k+1}^{\mathrm{eff}} = \xi_{k+1} + \mathcal{M}_{k+1}e_k$. Its conditional variance is $\sigma^2 + \sigma_{\mathcal{M}}^2 e_k^2$, whose average converges almost surely to $\sigma^2 + \sigma_{\mathcal{M}}^2 P$ by ergodicity. For every $\epsilon > 0$, stationarity and its finite second moment give

$$\mathbb{E}\left[\frac{1}{m}\sum_{k=0}^{m-1} \mathbb{E}\Big[(\zeta_{k+1}^{\mathrm{eff}})^2 \mathbf{1}_{\{|\zeta_{k+1}^{\mathrm{eff}}| > \epsilon\sqrt{m}\}} \mid \mathcal{F}_k\Big]\right] = \mathbb{E}\Big[(\zeta_1^{\mathrm{eff}})^2 \mathbf{1}_{\{|\zeta_1^{\mathrm{eff}}| > \epsilon\sqrt{m}\}}\Big] \to 0.$$

Thus conditional Lindeberg holds in probability. Summing the scalar recursion gives

$$\lambda_0\sqrt{m}\,\bar{e}_{0,m} = -\frac{1}{\sqrt{m}}\sum_{k=0}^{m-1}\zeta_{k+1}^{\mathrm{eff}} + \frac{e_0 - e_m}{\alpha\sqrt{m}}.$$

The endpoint tends to zero in L^2 by stationarity. The martingale CLT and Slutsky's theorem prove Equation (117). Since $\sigma^2 > 0$, one has $P > 0$, so the asserted variance inflation is strict when $\sigma_{\mathcal{M}}^2 > 0$. □

Entry Cost and Limits of Local Information.

Proposition G.3 (Second moments of the accelerated multiplicative oracle)**.** *In the scalar model Equation* (114)*, let $\mathcal{M}, \xi$ be independent within a pair, $\mathbb{E}\mathcal{M}^2 = \sigma_{\mathcal{M}}^2$ and $\mathbb{E}\xi^2 = \sigma^2$. For any fixed parameters of the two-step procedure, set $\mathsf{M}_\alpha = B_\alpha \mathsf{O}_\gamma$ and $\mathcal{K}(P) = APA^\mathsf{T} + \sigma_{\mathcal{M}}^2 \mathsf{M}_\alpha P \mathsf{M}_\alpha^\mathsf{T}$. For a deterministic $\mathbf{z}_0$ the uncentered second moment equals*

$$U_k = \mathbb{E}\mathbf{z}_k\mathbf{z}_k^\mathsf{T} = \mathcal{K}^k(\mathbf{z}_0\mathbf{z}_0^\mathsf{T}) + \sigma^2 \sum_{j=0}^{k-1} \mathcal{K}^j(B_\alpha B_\alpha^\mathsf{T}). \tag{118}$$

A sufficient condition for the stability of second moments is $\rho(A \otimes A + \sigma_{\mathcal{M}}^2 \mathsf{M}_\alpha \otimes \mathsf{M}_\alpha) < 1$. For $E_1 = [1, 0]$, the condition $E_1 U_t E_1^\mathsf{T} \le \varepsilon r^2$ guarantees $\mathbb{P}\{|e_t| > r\} \le \varepsilon$. For staying in the region over the whole finite interval $t \le k \le T$ it suffices that $\sum_{k=t}^{T} E_1 U_k E_1^\mathsf{T} \le \varepsilon r^2$.

Proof of Theorem G.3. In the scalar multiplicative model the state recursion is

$$\mathbf{z}_{k+1} = (A + \mathcal{M}_{k+1}\mathsf{M}_\alpha)\mathbf{z}_k + B_\alpha \xi_{k+1}.$$

Condition on $\mathcal{F}_k$. Centering removes the cross terms involving A and one noise, and independence within a pair removes the term involving $\mathcal{M}_{k+1}\xi_{k+1}$. Taking expectations therefore yields $U_{k+1} = \mathcal{K}(U_k) + \sigma^2 B_\alpha B_\alpha^\mathsf{T}$. Iteration proves Equation (118). Under column vectorization, $\mathcal{K}$ is represented by $A \otimes A + \sigma_{\mathcal{M}}^2 \mathsf{M}_\alpha \otimes \mathsf{M}_\alpha$. Spectral radius below one makes its powers decay geometrically and its geometric series converge, proving bounded second moments. Finally, $\mathbb{E}e_k^2 = E_1 U_k E_1^\mathsf{T}$; Markov's inequality proves the single-time certificate, and a union bound proves the certificate for the whole finite interval. □

This is a finite-dimensional check for SGD, HB and NAG; it accounts for the acceleration parameters and for the global intensity $\sigma_{\mathcal{M}}$. The smallest index t satisfying the chosen inequality is a certified burn-in in this class and enters the budget $N = b(t + m)$. At a high stationary noise level such an index may not exist. For independent averaging of b pairs, $\sigma_{\mathcal{M}}^2$ and σ^2 are replaced by $\sigma_{\mathcal{M}}^2/b$ and σ^2/b in the formulas.

Proposition G.4 (Local covariance does not bound the risk of this family)**.** *Fix $\lambda_0 > 0$, σ^2, $e_{-1} = e_0 \neq 0$, $\alpha > 0$, $\beta \in [0, 1)$ and $\gamma \geq 0$. Over the class of centered independent pairs $(\mathcal{M}, \xi)$ with arbitrary finite $\sigma^2_{\mathcal{M}}$, the risk of the two-point average $(u_0 + u_1)/2$ is not bounded by any finite function of $(\lambda_0, \sigma^2, e_0, \alpha, \beta, \gamma)$.*

Proof of Theorem G.4. The zero initial velocity gives $u_0 = e_0$ and $e_1 = e_0 - \alpha(\lambda_0 e_0 + \mathcal{M}_1 e_0 + \xi_1)$. Hence

$$\frac{u_0 + u_1}{2} = \left(1 - \frac{\alpha(1+\gamma)\lambda_0}{2}\right) e_0 - \frac{\alpha(1+\gamma)}{2}(\mathcal{M}_1 e_0 + \xi_1).$$

For the independent centered components allowed in the class, its variance equals $\alpha^2(1+\gamma)^2(\sigma^2_{\mathcal{M}} e_0^2 + \sigma^2)/4$. It diverges as $\sigma^2_{\mathcal{M}} \to \infty$, whereas the specified scalar parameters and the squared bias stay fixed. The risk is at least this variance, which proves the claim. □

G.2 Decreasing Steps and Changing Momentum

G.2.1 Fixed Momentum: A Conditional Transfer

Consider the normalized recursion

$$\Delta x_{k+1} = \beta_k \Delta x_k - (1 - \beta_k)\alpha_k G(s_k, \zeta_{k+1}), \tag{119}$$

$$x_{k+1} = x_k + \Delta x_{k+1}, \qquad s_k = x_k + \lambda_k^{\mathrm{qry}} \Delta x_k, \tag{120}$$

where $\lambda_k^{\mathrm{qry}} = 0$ for HB and $\lambda_k^{\mathrm{qry}} = \beta_k$ for NAG. We average $\bar{x}_n = n^{-1}\sum_{k=1}^n x_k$. Write $e_k = x_k - x_\star$ and $\xi_{k+1} = G(s_k, \zeta_{k+1}) - F(s_k)$, with $F = \nabla f$. These results are conditional transfer theorems: the stability and moment bounds below are hypotheses, not conclusions. Asymptotic equivalence of averaged momentum and SGD has prior results, including Tang et al. (2023); increasing-momentum averaging is also studied by An and Huo (2026).

Theorem G.5 (Fixed momentum with a Hölder linearization remainder)**.** *Let $\beta_k = \beta \in [0, 1)$ be fixed and $\alpha_k = a_{\mathrm{gain}}(k + k_0)^{-\gamma_{\mathrm{step}}}$ with $a_{\mathrm{gain}} > 0$. Assume the following.*

1. *$F(x_\star) = 0$, $H = \nabla F(x_\star) \succ 0$, and, for some $\nu \in (0, 1]$, the queried points satisfy almost surely*
$$\|F(s_k) - H(s_k - x_\star)\| \leq M_\nu \|s_k - x_\star\|^{1+\nu}.$$
2. *ξ_{k+1} is a martingale difference, its conditional covariance converges in probability to the deterministic matrix $\Sigma_\star$, and $\sup_k \mathbb{E}\|\xi_{k+1}\|^{2+\varepsilon_{\mathrm{mom}}} < \infty$ for some $\varepsilon_{\mathrm{mom}} > 0$.*
3. *For a constant C_{mom} and all k, including the finite starting indices,*
$$\mathbb{E}\|e_k\|^2 \leq C_{\mathrm{mom}}\alpha_k, \qquad \mathbb{E}\|\Delta x_k\|^2 \leq C_{\mathrm{mom}}\alpha_k^2.$$

If $1/(1+\nu) < \gamma_{\mathrm{step}} < 1$, then

$$\sqrt{n}(\bar{x}_n - x_\star) = -H^{-1}\frac{1}{\sqrt{n}}\sum_{k=1}^n \xi_{k+1} + R_n, \tag{121}$$

where

$$R_n = O_{\mathbb{P}}\left(n^{-(1-\gamma_{\mathrm{step}})/2} + n^{-[\gamma_{\mathrm{step}}(1+\nu)-1]/2}\right). \tag{122}$$

Consequently,

$$\sqrt{n}(\bar{x}_n - x_\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, H^{-1}\Sigma_\star H^{-1}). \tag{123}$$

If $F(x) = H(x - x_\star)$ at every queried point, the curvature term vanishes and the conclusion holds for every $0 < \gamma_{\mathrm{step}} < 1$ under the same moment assumptions.

The usual locally Lipschitz Hessian case is $\nu = 1$, giving the original $\gamma_{\mathrm{step}} \in (1/2, 1)$ range. Balancing the two displayed remainder powers gives $\gamma_{\mathrm{step}} = 2/(2+\nu)$ and rate $n^{-\nu/[2(2+\nu)]}$.

For fixed momentum, the familiar corrected state is

$$x_k^{\mathrm{corr}} = x_k + \frac{\beta}{1-\beta}\Delta x_k, \qquad x_{k+1}^{\mathrm{corr}} = x_k^{\mathrm{corr}} - \alpha_k G(s_k, \zeta_{k+1}). \tag{124}$$

The important improvement over bounding velocities separately is

$$\sum_{k=1}^{n} \Delta x_k = x_n - x_0.$$

Thus the linear query mismatch is an endpoint

Proof of Theorem G.5. For this proof put $\gamma = \gamma_{\text{step}}$ and $\|Y\|_{L^2} = (\mathbb{E}\|Y\|^2)^{1/2}$. All constants below may depend on the fixed parameters in the assumptions but not on n. Define

$$\widetilde{e}_k = x_k^{\text{corr}} - x_\star = e_k + \frac{\beta}{1-\beta}\Delta x_k, \qquad \mathfrak{r}_k = F(s_k) - H(s_k - x_\star).$$

The moment assumptions give $\|\widetilde{e}_k\|_{L^2} = O(\alpha_k^{1/2})$ and $\mathbb{E}\|s_k - x_\star\|^2 = O(\alpha_k)$. Here $\lambda^{\text{qry}} = 0$ for HB and $\lambda^{\text{qry}} = \beta$ for NAG is constant. The corrected-state recursion yields the identity

$$H\sum_{k=1}^{n} e_k = T_n - \sum_{k=1}^{n} \xi_{k+1} - \sum_{k=1}^{n} \mathfrak{r}_k - H\lambda^{\text{qry}}(e_n - e_0), \qquad T_n = \sum_{k=1}^{n} \frac{\widetilde{e}_k - \widetilde{e}_{k+1}}{\alpha_k}. \tag{125}$$

In particular, no sum of the norms of the increments is needed for the linear query mismatch.

Summation by parts gives

$$T_n = \frac{\widetilde{e}_1}{\alpha_1} - \frac{\widetilde{e}_{n+1}}{\alpha_n} + \sum_{k=2}^{n} (\alpha_k^{-1} - \alpha_{k-1}^{-1})\widetilde{e}_k. \tag{126}$$

Since $\alpha_k^{-1} - \alpha_{k-1}^{-1} = O((k+k_0)^{\gamma-1})$, Minkowski's inequality shows

$$\|T_n\|_{L^2} \le C\left(1 + n^{\gamma/2} + \sum_{k=2}^{n} (k+k_0)^{\gamma/2-1}\right) = O(n^{\gamma/2}).$$

Also $\|e_n - e_0\|_{L^2} = O(1)$. By the Taylor bound and $1+\nu \le 2$,

$$\mathbb{E}\|\mathfrak{r}_k\| \le M_\nu(\mathbb{E}\|s_k - x_\star\|^2)^{(1+\nu)/2} \le C\alpha_k^{(1+\nu)/2}.$$

For every $0 < \gamma < 1$ and $0 < \nu \le 1$ the exponent $\gamma(1+\nu)/2$ is strictly below one. Hence

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E}\|\mathfrak{r}_k\| = O\big(n^{[1-\gamma(1+\nu)]/2}\big).$$

Multiplying Equation (125) by $H^{-1}/\sqrt{n}$, and using Markov's inequality for the curvature sum and the L^2 bounds for the other terms, proves Equations (121) and (122).

For completeness, let $C_k = \mathbb{E}[\xi_{k+1}\xi_{k+1}^{\mathsf{T}} \mid \mathcal{F}_k]$, with $\mathcal{F}_k$ the history before that noise is observed. The uniform $(2+\varepsilon_{\text{mom}})$ moment implies uniform integrability of C_k. Its assumed convergence in probability therefore implies L^1 convergence and $n^{-1}\sum_{k=1}^{n} C_k \to \Sigma_\star$ in probability. For every $\epsilon > 0$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\big[\|\xi_{k+1}\|^2 \mathbf{1}_{\{\|\xi_{k+1}\| > \epsilon\sqrt{n}\}}\big] \le C\epsilon^{-\varepsilon_{\text{mom}}} n^{-\varepsilon_{\text{mom}}/2} \longrightarrow 0.$$

Thus the conditional Lindeberg sum tends to zero in probability. The martingale CLT, applied to every fixed scalar projection, gives $n^{-1/2}\sum_k \xi_{k+1} \xrightarrow{\text{d}} \mathcal{N}(0, \Sigma_\star)$. When $\gamma(1+\nu) > 1$, both remainder powers vanish, and Slutsky's theorem proves Equation (123). If F is linear, $\mathfrak{r}_k = 0$ identically, so every $0 < \gamma < 1$ is allowed under the same moment assumptions. □

G.2.2 Batching and the Cost of Localization

Corollary G.6 (Batch-dependent moments remove an artificial batch cap)**.** *For each horizon n, run the fixed-momentum recursion with a batch $b = b_n \geq 1$. Assume the Taylor bound in Theorem G.5 holds uniformly, and that for a constant C_{mom} independent of n, b, and k,*

$$\mathbb{E}\,\|e_{k,b}\|^2 \leq C_{\mathrm{mom}}\alpha_k/b, \qquad \mathbb{E}\,\|\Delta x_{k,b}\|^2 \leq C_{\mathrm{mom}}\alpha_k^2/b \tag{127}$$

for all starting and subsequent indices, uniformly in n. The individual oracle noises in each batch are conditionally independent martingale differences with uniformly bounded $(2+\varepsilon_{\mathrm{mom}})$ moments; their averaged conditional covariance converges in probability to $\Sigma_\star$. Then, for $0 < \gamma_{\mathrm{step}} < 1$,

$$\begin{aligned}\sqrt{nb}(\bar{x}_{n,b} - x_\star) &= -\frac{H^{-1}}{\sqrt{nb}}\sum_{k=1}^{n}\sum_{j=1}^{b}\xi_{k+1,j} + R_{n,b},\\ R_{n,b} &= O_{\mathbb{P}}\Big(n^{-(1-\gamma_{\mathrm{step}})/2} + b^{-\nu/2}n^{[1-\gamma_{\mathrm{step}}(1+\nu)]/2}\Big).\end{aligned} \tag{128}$$

Hence the same sandwich CLT holds whenever $b^{-\nu/2}n^{[1-\gamma_{\mathrm{step}}(1+\nu)]/2} \to 0$.

In the triangular-array assumption of Theorem G.6, *averaged conditional covariance* means the row-wise condition

$$\frac{1}{nb_n}\sum_{k=1}^{n}\sum_{j=1}^{b_n}\mathbb{E}\Big[\xi_{k+1,j}^{(n)}(\xi_{k+1,j}^{(n)})^{\mathsf{T}} \mid \mathcal{F}_{k,n}\Big] \xrightarrow{\mathbb{P}} \Sigma_\star,$$

where $\mathcal{F}_{k,n}$ is the history before the kth batch in row n. Conditional independence within a batch preserves these conditional moments when its individual noises are revealed successively.

Proof of Theorem G.6. Put $\gamma = \gamma_{\mathrm{step}}$ and $\bar{\xi}_{k+1,b} = b^{-1}\sum_{j=1}^{b}\xi_{k+1,j}$. The pathwise identity Equation (125) applies with this batch noise. Under Equation (127), its corrected errors and query errors satisfy, uniformly in the array,

$$\|\widetilde{e}_{k,b}\|_{L^2} \leq C\sqrt{\alpha_k/b}, \qquad \mathbb{E}\|s_{k,b} - x_\star\|^2 \leq C\alpha_k/b.$$

The same Abel sum and endpoint cancellation consequently give

$$\sqrt{\frac{b}{n}}\,\|T_{n,b}\|_{L^2} = O\big(n^{-(1-\gamma)/2}\big), \qquad \sqrt{\frac{b}{n}}\,\|e_{n,b} - e_{0,b}\|_{L^2} = O(n^{-1/2}).$$

The uniform moment bound at the starting indices is used in both estimates. For the curvature remainder,

$$\sqrt{\frac{b}{n}}\sum_{k=1}^{n}\mathbb{E}\|\mathfrak{r}_{k,b}\| \leq C\sqrt{\frac{b}{n}}\sum_{k=1}^{n}(\alpha_k/b)^{(1+\nu)/2} = O\big(b^{-\nu/2}n^{[1-\gamma(1+\nu)]/2}\big).$$

Multiplication of the pathwise identity by $H^{-1}\sqrt{b/n}$ now proves the expansion and remainder in Equation (128).

To obtain the CLT, reveal the conditionally independent individual noises within each batch one at a time. They form a martingale triangular array whose normalized sum is $(nb)^{-1/2}\sum_{k,j}\xi_{k+1,j}$. The assumed averaged conditional-covariance limit gives its predictable covariance convergence to $\Sigma_\star$. The uniform individual moment bound implies, for every $\epsilon > 0$,

$$\frac{1}{nb}\sum_{k=1}^{n}\sum_{j=1}^{b}\mathbb{E}\big[\|\xi_{k+1,j}\|^2\mathbf{1}_{\{\|\xi_{k+1,j}\|>\epsilon\sqrt{nb}\}}\big] \leq C\epsilon^{-\varepsilon_{\mathrm{mom}}}(nb)^{-\varepsilon_{\mathrm{mom}}/2} \longrightarrow 0.$$

This verifies conditional Lindeberg in probability. The martingale CLT and Slutsky's theorem prove the sandwich limit whenever the displayed curvature term vanishes; the endpoint term already tends to zero for $0 < \gamma < 1$. □

For $\nu = 1$ and $\gamma_{\mathrm{step}} > 1/2$, this condition imposes no upper bound on batch growth. The condition $b = o(n^{1/3})$ at $\gamma_{\mathrm{step}} = 2/3$, which follows from unscaled uniform moment bounds, is a sufficient condition. The stronger Equation (127) excludes a fixed nonzero cold start as $b \to \infty$. A warm start or a separately proved burn-in should supply those bounds. The statistical gain does not pay for that initialization for free.

Limits of the Statement. A local Taylor bound can replace the bound at all query points for a single infinite trajectory if an almost surely finite time exists after which every query stays in the neighborhood; the finite-prefix remainder is then $O_{\mathbb{P}}(n^{-1/2})$. Triangular batch arrays need a corresponding uniform argument. At $\alpha_k \asymp 1/k$, arithmetic averaging is generally inefficient even in one dimension: a factor-two variance can already occur in a scalar linear example. For a fixed positive gain on a nonlinear problem, stationary averaging generally targets a biased stationary mean.

G.2.3 Verified Moment Bounds for SGD

Proposition G.7 (Verification of the moment assumptions for SGD)**.** *Let $f \in C^2$, $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all x, let the Hessian be globally Lipschitz, and let the noise be independent, additive, centered, with a finite moment of order $2+\varepsilon_{\mathrm{mom}}$ and covariance Σ. Let $\sigma^2 = \operatorname{tr}\Sigma$, $b \geq 1$, $\beta = 0$, $\Delta x_0 = 0$, and let the nonincreasing steps satisfy*

$$0 < \alpha_k \leq 1/L, \qquad \alpha_{k+1} \geq \alpha_k - \tfrac{\mu}{2}\alpha_k^2. \tag{129}$$

Then, for $C_{\mathrm{sgd}} \geq 2\sigma^2/\mu$ and any deterministic start,

$$\mathbb{E}\|e_k\|^2 \leq \|e_0\|^2 \prod_{i=0}^{k-1}(1-\mu\alpha_i)^2 + C_{\mathrm{sgd}}\alpha_k/b. \tag{130}$$

If $\alpha_{\mathrm{first}} := \alpha_k|_{k=0}$ and $\mathbb{E}\|e_0\|^2 \leq C_{\mathrm{sgd}}\alpha_{\mathrm{first}}/b$, the first term can be dropped. With such a start, $\mathbb{E}\|\Delta x_{k+1}\|^2 \leq \alpha_k^2(L^2 C_{\mathrm{sgd}}\alpha_k + \sigma^2)/b$. For $\alpha_k = a_{\mathrm{sgd}}(k+k_0)^{-\gamma_{\mathrm{step}}}$ with $a_{\mathrm{sgd}} > 0$, $1/2 < \gamma_{\mathrm{step}} < 1$, the conditions Equation (129) *are ensured by a sufficiently large k_0; hence the assumptions of Theorem G.5 with $\nu = 1$ hold for every fixed b, including a cold start. For an array of growing b, the warm-start condition gives the uniform moment bounds Equation* (127)*.*

Proof of Theorem G.7. Write $\nabla f(x_k) = H_k e_k$ with $H_k = \int_0^1 \nabla^2 f(x_\star + se_k)\,ds$. Its eigenvalues lie in $[\mu, L]$, and $0 < \alpha_k \leq 1/L$ gives $\|I - \alpha_k H_k\| \leq 1 - \mu\alpha_k$. Thus, for $p_k = \mathbb{E}\|e_k\|^2$,

$$p_{k+1} \leq (1-\mu\alpha_k)^2 p_k + \alpha_k^2\sigma^2/b.$$

The slow-step condition and $\mu\alpha_k \leq 1$ imply

$$\begin{aligned} C_{\mathrm{sgd}}\alpha_{k+1} - C_{\mathrm{sgd}}\alpha_k(1-\mu\alpha_k)^2 &\geq C_{\mathrm{sgd}}\mu\alpha_k^2(3/2 - \mu\alpha_k) \\ &\geq C_{\mathrm{sgd}}\mu\alpha_k^2/2 \geq \sigma^2\alpha_k^2. \end{aligned}$$

Induction, separating the homogeneous recursion from its noise term, proves Equation (130). If $p_0 \leq C_{\mathrm{sgd}}\alpha_{\mathrm{first}}/b$, the same induction starts directly with this envelope and removes the homogeneous term. Independence and centering of the noise also give

$$\mathbb{E}\|\Delta x_{k+1}\|^2 = \alpha_k^2\{\mathbb{E}\|\nabla f(x_k)\|^2 + \sigma^2/b\} \leq \alpha_k^2(L^2 p_k + \sigma^2/b),$$

which is the claimed velocity bound.

For $\alpha_k = a_{\mathrm{sgd}}(k+k_0)^{-\gamma_{\mathrm{step}}}$, the mean value theorem gives

$$\frac{\alpha_k - \alpha_{k+1}}{\alpha_k^2} \leq \frac{\gamma_{\mathrm{step}}}{a_{\mathrm{sgd}}}(k+k_0)^{\gamma_{\mathrm{step}}-1}.$$

A sufficiently large k_0 ensures both conditions in Equation (129). Moreover, $\prod_{i<k}(1-\mu\alpha_i)^2 \leq \exp(-2\mu\sum_{i<k}\alpha_i)$ decays faster than any inverse power of k when $\gamma_{\mathrm{step}} < 1$. For fixed b, it can therefore be absorbed into an $O(\alpha_k)$ moment bound even from a cold start. The velocity bound, the bounded adjacent step ratios, and the globally Lipschitz Hessian verify the remaining moment and Taylor assumptions of Theorem G.5 with $\nu = 1$. For growing batches, the warm-start envelope gives the same bounds uniformly with factor $1/b$, as required by Equation (127). □

With a fixed cold start and a growing batch, the remainder in Equation (130) has to be paid for by burn-in. It suffices to ensure $2\mu\sum_{i<t}\alpha_i \geq \log(\|e_0\|^2 b/(C_{\mathrm{sgd}}\alpha_t))$ in order to suppress the initial contribution to $C_{\mathrm{sgd}}\alpha_t/b$; the constant of the final moment bound can then be doubled. For example, the gradient of the nonlinear experiment $F(x) = x + 0.4x^2/(1+x^2)$ satisfies $1-3\sqrt{3}/20 \leq F'(x) \leq 1+3\sqrt{3}/20$ and $|F''(x)| \leq 0.8$. For $\alpha_k = 1.2(k+40)^{-2/3}$ both conditions Equation (129) hold: the maximal step is below $1/L$, and $(\alpha_k - \alpha_{k+1})/\alpha_k^2 \leq (5/9)40^{-1/3} < \mu/2$. Thus the moment assumptions of the nonlinear experiment for SGD with this policy are verified globally.

G.2.4 Increasing Momentum: A Conditional Transfer

We use the assumptions of Theorem G.5. The exponent ϑ_{vel} below controls the velocity moment and is unrelated to the initial signal-to-noise ratio $\theta = \mu e_0/\sigma$.

Theorem G.8 (Direct Abel transfer for increasing momentum)**.** *Retain assumptions (i)–(ii) of Theorem G.5 and* $\mathbb{E}\|e_k\|^2 \le C_{\mathrm{mom}}\alpha_k$. *Let*

$$\alpha_k = a_{\mathrm{mom}}(k+k_0)^{-\gamma_{\mathrm{step}}}, \qquad \varpi_k = 1-\beta_k = \varpi_0(k+k_0)^{-\rho_{\mathrm{mom}}} \in (0,1],$$

with $a_{\mathrm{mom}}, \varpi_0 > 0$, $0 < \gamma_{\mathrm{step}} < 1$, *and* $\rho_{\mathrm{mom}} \ge 0$, *and suppose, for some* $\vartheta_{\mathrm{vel}} \in [0,1]$,

$$\mathbb{E}\|\Delta x_k\|^2 \le C_{\mathrm{mom}}\alpha_k^2\varpi_k^{\vartheta_{\mathrm{vel}}}. \tag{131}$$

For HB or NAG, Equation (121) *holds with*

$$R_n = O_{\mathbb{P}}\Big(n^{-(1-\gamma_{\mathrm{step}})/2} + n^{-[\gamma_{\mathrm{step}}(1+\nu)-1]/2} + n^{\rho_{\mathrm{mom}}(1-\vartheta_{\mathrm{vel}}/2)-1/2}\Big). \tag{132}$$

In particular, the CLT Equation (123) *holds if*

$$\frac{1}{1+\nu} < \gamma_{\mathrm{step}} < 1, \qquad \rho_{\mathrm{mom}} < \frac{1}{2-\vartheta_{\mathrm{vel}}}.$$

The curvature term and its restriction on γ_{step} *are omitted for an linear* F.

Proof of Theorem G.8. Use the local abbreviations $\gamma = \gamma_{\mathrm{step}}$, $r = \rho_{\mathrm{mom}}$, and $\vartheta = \vartheta_{\mathrm{vel}}$. If $r = 0$, the momentum is fixed and the estimates in the proof of Theorem G.5 apply; those remainder estimates hold for every $0 < \gamma < 1$, before requiring their decay. Suppose now that $r > 0$. Set

$$c_k = \alpha_k^{-1}, \qquad d_k = (\alpha_k\varpi_k)^{-1}, \qquad \mathfrak{r}_k = F(s_k) - H(s_k - x_\star).$$

Rearrangement of Equation (119) gives, ,

$$F(s_k) + \xi_{k+1} = -c_k\Delta x_k - d_k(\Delta x_{k+1} - \Delta x_k).$$

Consequently

$$\begin{aligned} H\sum_{k=1}^n e_k &= -\sum_{k=1}^n \xi_{k+1} - A_n - D_n - \sum_{k=1}^n \mathfrak{r}_k - HQ_n, \\ A_n &= \sum_{k=1}^n c_k\Delta x_k, \qquad D_n = \sum_{k=1}^n d_k(\Delta x_{k+1} - \Delta x_k), \qquad Q_n = \sum_{k=1}^n \lambda_k^{\mathrm{qry}}\Delta x_k. \end{aligned} \tag{133}$$

Apply summation by parts first to the errors and then to the increments:

$$A_n = c_n e_n - c_1 e_0 - \sum_{k=1}^{n-1}(c_{k+1} - c_k)e_k,$$

$$D_n = d_n\Delta x_{n+1} - d_1\Delta x_1 - \sum_{k=2}^n (d_k - d_{k-1})\Delta x_k.$$

Since $c_{k+1} - c_k = O((k+k_0)^{\gamma-1})$ and $\|e_k\|_{L^2} = O((k+k_0)^{-\gamma/2})$, the first line yields $\|A_n\|_{L^2} = O(n^{\gamma/2})$. For the second line use

$$d_k - d_{k-1} = O((k+k_0)^{\gamma+r-1}), \qquad \|\Delta x_k\|_{L^2} = O((k+k_0)^{-\gamma-r\vartheta/2}).$$

Writing $a = r(1-\vartheta/2) > 0$, this gives

$$\|D_n\|_{L^2} \le C\left(1 + n^a + \sum_{k=2}^n (k+k_0)^{a-1}\right) = O(n^a).$$

For HB, $Q_n = 0$. For NAG, $\lambda_k^{\mathrm{qry}} = 1 - \varpi_k$, so a further telescoping identity gives

$$Q_n = e_n - e_0 - \left\{\varpi_n e_n - \varpi_1 e_0 + \sum_{k=1}^{n-1}(\varpi_k - \varpi_{k+1})e_k\right\}.$$

The sequence ϖ_k is nonincreasing and bounded, its successive differences have a bounded sum, and $\sup_k \|e_k\|_{L^2} < \infty$. Thus $\|Q_n\|_{L^2} = O(1)$, retaining the cancellation of the linear query mismatch even when momentum varies.

Finally $\varpi_k^{\vartheta} \le 1$ implies $\mathbb{E}\|s_k - x_\star\|^2 = O(\alpha_k)$. The same Taylor-remainder estimate as in the fixed-momentum proof gives

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E}\|\mathfrak{r}_k\| = O\big(n^{[1-\gamma(1+\nu)]/2}\big).$$

Multiply Equation (133) by $H^{-1}/\sqrt{n}$. The bounds for A_n, D_n, Q_n, and the curvature sum prove Equation (132); the $O_{\mathbb{P}}(n^{-1/2})$ query endpoint is absorbed by the first term. The noise assumptions give the same martingale CLT as in Theorem G.5. All remainders vanish when $\gamma(1+\nu) > 1$ and $r(1-\vartheta/2) < 1/2$, which is the stated range. For an linear F, $\mathfrak{r}_k = 0$ and its restriction on γ disappears. □

With the basic velocity estimate ($\vartheta_{\text{vel}} = 0$), this gives the conditional range $\rho_{\text{mom}} < 1/2$, with no additional restriction coupling ρ_{mom} and γ_{step}. With the stronger estimate $\mathbb{E}\,\|\Delta x_k\|^2 = O(\alpha_k^2 \varpi_k)$, the conditional range is $\rho_{\text{mom}} < 1$. The latter estimate is a substantial assumption when momentum tends to one. This theorem does not establish it for general nonlinear dynamics, and it does not prove that the boundary $\rho_{\text{mom}} = 1$ is necessary.

G.3 Richardson–Romberg: A Different Correction Mechanism

Sheshukova et al. (2025, Theorem 6, Corollary 7, equations (36)–(41)) analyze two coupled constant-step SGD trajectories with tail averaging. With step $\alpha_{\text{RR}} = n^{-1/2}$, their bound on $\{\mathbb{E}\|H_\star(\widehat{x}^{\text{RR}} - x_\star)\|^2\}^{1/2}$ has a leading $n^{-1/2}$ term, an $O(n^{-3/4})$ remainder, and an initialization term; its hypotheses include sixth-moment controls and fourth-order smoothness. The exponent $3/4$ is described as best known, not as a universal lower bound. Squaring an upper bound $K_1 n^{-1/2} + K_2 n^{-3/4}$ also produces $2K_1K_2 n^{-5/4}$, so it does not identify an $n^{-3/2}$ first correction to the MSE. Their shared data stream preserves the leading noise term in $2\widehat{x}_{\alpha_{\text{RR}}} - \widehat{x}_{2\alpha_{\text{RR}}}$. For independent copies with leading covariance Σ_{lead}/n, elementary covariance algebra instead gives $5\Sigma_{\text{lead}}/n$. Their convention uses $2n$ observations including burn-in and two gradient evaluations per observation. This correction targets nonlinear stationary bias; that bias is absent in our additive quadratic model, whose initialization and covariance transients remain relevant.

H Details of the Census Experiments

This section gives the complete protocol for the Census experiments summarized in the main text. It specifies the data processing, stochastic oracles, algorithmic parameters, interval construction, and additional numerical results needed for reproduction.

H.1 Data, Oracle, and Parameters

Data. The file `adult_reconstruction.csv` of Ding et al. (2021) contains the 49,531 persons of the 1994 Current Population Survey from which the UCI Adult data were drawn, with their total annual income (reported up to \$99,999). Age, usual weekly hours, and years of education are standardized; class of worker, marital status, relationship, race, sex, occupation, and region of birth enter as indicators, with levels rarer than 0.1% merged. Write (z_i, y_i) for the encoded covariate–response pair of record i, with $z_i \in \mathbb{R}^{44}$ including the intercept. The linear model uses log income as response; the logistic model uses the indicator that income exceeds \$50,000. All coefficients are penalized with a ridge level λ; besides the cross-validated level we use two larger levels that reduce the condition number by factors 10 and 100 (Table 5). The reconstruction is governed by the IPUMS terms of use; the accompanying Folktables code is released under the MIT license. The supplementary code downloads the data from its cited source and does not redistribute it. The file contains demographic attributes, including race and sex. We use them only as covariates in aggregate algorithmic simulations, report no individual-level results.

Prediction-Powered Oracles. A pilot sample of 10^4 persons, drawn once and independently of all replicates, is used to fit gradient-boosting predictors $\hat{f}$ of log income and $\hat{p}$ of the income indicator (`scikit-learn`, default hyperparameters). Let $\Gamma_z := \mathbb{E}[zz^\top]$. With $r = y - \hat{f}(z)$ and the identity $(\Gamma_z + \lambda I)x_\star = \mathbb{E}[zy]$, the linear oracle of Section 6 equals

$$g_i(x) = (\Gamma_z + \lambda I)(x - x_\star) - \big(z_i r_i - \mathbb{E}[zr]\big),$$

so its noise has covariance $\Sigma = \mathrm{Cov}(zr)$ and does not depend on x; a batch averages b independent draws with replacement, and $V_\star = H^{-1}\Sigma H^{-1}$ with $H = \Gamma_z + \lambda I$. Define $\mathrm{sigm}(t) := (1 + e^{-t})^{-1}$. The logistic oracle

is $\mathbb{E}[z\,\mathrm{sigm}(z^\mathsf{T}x)] + \lambda x - \mathbb{E}[z\hat{p}(z)] - z_i\{y_i - \hat{p}(z_i)\}$; its global constants in Theorem C.2 are $\mu = \lambda$ and $L = \lambda + \lambda_{\max}(\Gamma_z)/4$. All expectations are finite-population averages over the file.

Table 5: Constants of the Census Experiments

Quantity	Cross-validated λ	Second λ	Third λ
κ of H, linear model	1.5×10^4	1532	153
κ of H at $x_\star$, logistic model	4.1×10^4	209	23
L/μ with global constants, logistic model	1.3×10^5	425	39
Variance of log income explained by $\hat{f}$	66%		
Reduction of $\mathrm{tr}\,\Sigma$ by the control variate	factor 168		

H.2 Methods, Intervals, and Protocol

Methods. SGD, HB, and NAG use the policies of Table 2 with $\mu = \lambda_{\min}(H)$ and $L = \lambda_{\max}(H)$, zero initial velocity, and the start x_0 with intercept equal to the mean pilot response and all other coefficients zero. The output is the full average of the query points. In the logistic experiment, which uses the third ridge level and $x_0 = 0$, NAG uses the burn-in Equation (57) with $\mathcal{V}_0$ computed on the file and $\sigma^2 = \mathrm{tr}\,\Sigma$, followed by tail averaging; constant-step SGD uses the analogous burn-in with $\kappa^{-1/2}$ replaced by κ^{-1}. The batch schedules for NAG are $b = m^2$ ($m = 4, \dots, 128$), $b = m$ ($m = 8, \dots, 1024$), and $b = 64$ ($m = 16, \dots, 1024$), and SGD uses $b = m^2$ ($m = 4, \dots, 64$); every acquired response, including those spent in burn-in, is counted in N and used by the ERM.

Intervals. All intervals have level 95% and are centered at the output. The oracle interval uses $V_\star$. The plug-in interval replaces Σ by the empirical covariance of the acquired terms $z_i r_i$ (the matrix H is known in this setting). The batch-means interval splits the m query points into $\lfloor\sqrt{m}\rfloor$ consecutive blocks of equal length and uses Student quantiles; equal blocks are the constant-step counterpart of the construction of Chen et al. (2020). Random scaling (Lee et al., 2022) uses the partial averages $\bar{u}_k := k^{-1}\sum_{\ell=0}^{k-1} u_\ell$ through $\hat{V} = m^{-2}\sum_{k\le m} k^2(\bar{u}_k - \bar{u}_m)(\bar{u}_k - \bar{u}_m)^\mathsf{T}$ and the critical value 6.747. Batch means and random scaling rely on a functional central limit theorem for the partial sums of the query points. Applying Theorem A.1 separately to each prefix gives only pointwise decompositions; a functional result also requires uniform endpoint control and tightness. We therefore include these two intervals as empirical comparators and do not claim their asymptotic validity from Theorem A.1 alone.

Gaussian Coverage Predictions from Finite-Sample Moments. The mean of the output is obtained by running the noiseless recursion from x_0. For its variance, the recursion decouples in the eigenbasis (λ_i, φ_i) of H; with $r_n(\lambda)$ the response of Theorem B.1, for a fixed contrast $c \in \mathbb{R}^d$,

$$\mathrm{Var}(c^\mathsf{T}\bar{u}_{0,m}) = \frac{1}{bm^2}\sum_{s=1}^{m} w_s^\mathsf{T}\Sigma w_s, \qquad w_s = \sum_{i=1}^{d}\frac{1 - r_{m-s}(\lambda_i)}{\lambda_i}(\varphi_i^\mathsf{T}c)\,\varphi_i,$$

which reduces to Equation (23) in one dimension. The predicted coverage is $\Phi((1.96 - h_c)/\sqrt{v_c}) - \Phi((-1.96 - h_c)/\sqrt{v_c})$, where

$$h_c := \frac{\mathbb{E}[c^\mathsf{T}\bar{u}_{0,m}]}{\sqrt{c^\mathsf{T}V_\star c/N}}, \qquad v_c := \frac{\mathrm{Var}(c^\mathsf{T}\bar{u}_{0,m})}{c^\mathsf{T}V_\star c/N}.$$

We call a batch size the breakdown batch if coverage stays below 90% from it on. This coverage formula is under Gaussian oracle noise; for Census resampling it is a Gaussian approximation using the finite-sample mean and variance.

Protocol. Each configuration of the linear experiments is repeated 500 times and each configuration of the logistic experiment 150 times. Within a replicate of the linear experiment, all methods acquire the same responses, and for a fixed batch size the outputs at smaller budgets are prefixes of the same runs. The logistic schedules are run independently across methods and values of m.

H.3 Additional Results

Further Results. The normalized mean squared error has the same thresholds (Figure 3). Variance estimates do not correct a shifted center: beyond the breakdown of SGD, plug-in intervals cover 0%, and trajectory-based intervals require five to eight times the oracle width.

Accuracy of the Gaussian-Moment Predictions. Figure 2(a) compares Gaussian-moment predictions with observed coverage over all 1026 configurations of the linear experiment: the median difference is 0.7 percentage points, the 90% quantile 1.9, and the maximum 5.0, while the Monte Carlo standard error is about one point near the nominal level and two points near 50%. Along the two eigendirections of Figure 1 the differences do not exceed 0.04. The predicted breakdown batch matches the observed grid point in 85% of the 81 combinations of method, ridge level, functional, and budget, and is within one grid step in 98%. Figure 2(b) shows that it grows with the budget, whereas the gradient noise scale $\operatorname{tr}\Sigma / \|\nabla f(x_0)\|^2$, which governs the optimization speedup from batching (McCandlish et al., 2018), does not depend on N (Theorem B.6).

The Spectrum and the Start. Figure 2(c) shows the relative variance excess $v - 1$ along every eigenvector of H at $N = 2^{16}$ and $b = 4$. The observations follow the values up to Monte Carlo fluctuations of order 0.1. The only large excess is that of HB along the top eigenvector, 16 against the value 14 and the value $\kappa^{3/2}/(8m) = 14.5$ of Statement 4.1; the largest excess of NAG is 0.17, within the bound $30\sqrt{\kappa}/m = 0.23$ of Statement 4.2. Figure 2(d) repeats Figure 1(c) from $x_0 = 0$. The larger start lowers the breakdown batch of SGD from 2^{10} to 2^6 and that of NAG from 2^{12} to 2^{11}, as the factor $|\theta|$ in Statement 3.2 predicts, while that of HB stays at 2^3: HB is limited by its covariance, not by its center.

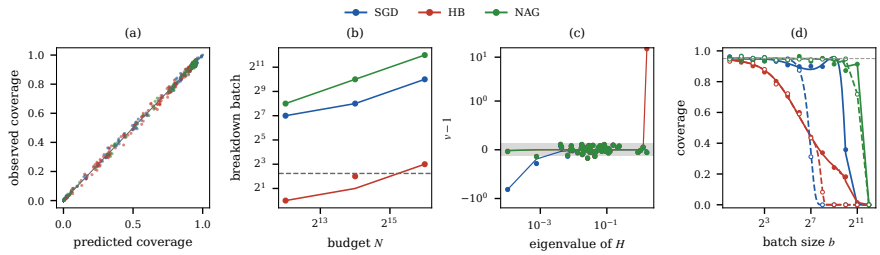

Figure 2: (a) Predicted against observed coverage over all budgets, batch sizes, ridge levels, functionals, and methods. (b) Breakdown batch for the age coefficient against the budget: observed (markers) and predicted (lines); the dashed line is the gradient noise scale. (c) Relative variance excess along the eigenvectors of H at $N = 2^{16}$ and $b = 4$ (markers: Monte Carlo; lines: exact; grey band: ± 2 Monte Carlo standard errors). (d) Coverage for the age coefficient starting from the pilot mean (solid lines, filled markers) and from $x_0 = 0$ (dashed lines, open markers); lines are predictions and markers Monte Carlo estimates.

Mean Squared Error. Figure 3 reports the same runs in the optimization metric

$$R_c := \frac{N\,\mathbb{E}\big[(c^\top(\hat{x} - x_\star))^2\big]}{c^\top V_\star c} = h_c^2 + v_c,$$

which tends to one in the PR regime. The batch at which this ratio first exceeds two closely tracks the coverage breakdown in Figure 1 and gives the same method ordering. Along the slowest eigendirection it occurs at $b = 4$ for SGD, $b = 128$ for NAG, and $b = 256$ for HB. Along the stiffest eigendirection the HB ratio is already 4.2 at $b = 1$, whereas NAG remains near one throughout the tested range. For the age coefficient, HB exceeds two from $b = 16$, SGD from $b = 2^{10}$, and NAG only at $b = 2^{12}$. The exact finite-moment predictions track the observed curves. Across the displayed points, the Monte Carlo estimates differ from the values by at most 14%, on the scale of the shown ± 2-standard-error bands.

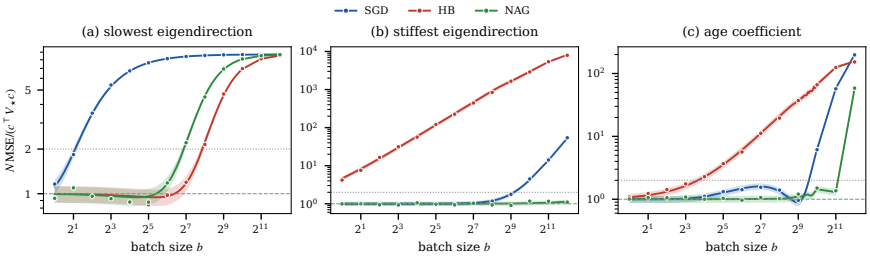

Figure 3: Normalized mean squared error against the batch size at $N = 2^{16}$ and $\kappa = 1.5 \times 10^4$, for the functionals of Figure 1. Lines show the values $h_c^2 + v_c$, shaded bands ± 2 Monte Carlo standard errors for 500 replicates around them, and markers the Monte Carlo estimates. Gray horizontal dashed and dotted lines mark the PR level and twice that level.

Gaussian Shape. To probe the shape of the output under skewed, bounded non-Gaussian noise, we repeat the linear experiment with untransformed income (in units of \$10,000) and 4000 replicates; the standardized third moment of the influence values of the age coefficient is 3.7. After centering and scaling, the Kolmogorov distance to the standard normal law never exceeds 0.018, below the 5% one-sample KS critical value 0.022, for all methods, $b \in \{1, 16, 256\}$, and $N \in \{2^8, \ldots, 2^{14}\}$, although the variance of these outputs ranges from 0.25 to 14 times the PR variance (Figure 4a). Normalization by the PR variance gives distances up to 0.96 (Figure 4b). Across this tested grid, Gaussian shape is close after normalization even when PR calibration is poor, consistent with the separation in Statement 3.3. A distance below the one-sample KS critical value indicates non-rejection in that test, not a bound on the distance between the underlying distributions. Exact standardization requires $m = N/b \geq 2$; when $N = b = 256$, the output is the deterministic x_0 and its variance is zero.

Logistic Regression. Here $N = (t + m)b$ includes burn-in observations, $\widehat{x}_N^{\mathrm{ERM}}$ is computed from all N observations, and $V_\star = H^{-1}\Sigma H^{-1}$ uses $H = \nabla^2 f(x_\star)$. Figure 4(c) shows the mean of $\sqrt{N}\left\|V_\star^{-1/2}(\hat{x} - \widehat{x}_N^{\mathrm{ERM}})\right\|$, the normalized gap between the online output and the ERM on the same N responses. With $b = m^2$, as in Theorem C.3, the gap of NAG decreases from 5.7 to 3.4 at $N \approx 2.8 \times 10^6$ after 168 rounds; with $b = m$ it reaches 2.6 after 1047 rounds. Both are well below 6.6, the normalized error of the ERM, so most of the sampling error of the online output is shared with the ERM. With a constant batch the gap grows from 3.6 to 9.5, the signature of the stationary bias of constant-gain averaging. Constant-step SGD with $b = m^2$ needs a three to four times longer burn-in than NAG and stays near 6.6, because its boundary term scales with κ rather than $\sqrt{\kappa}$.

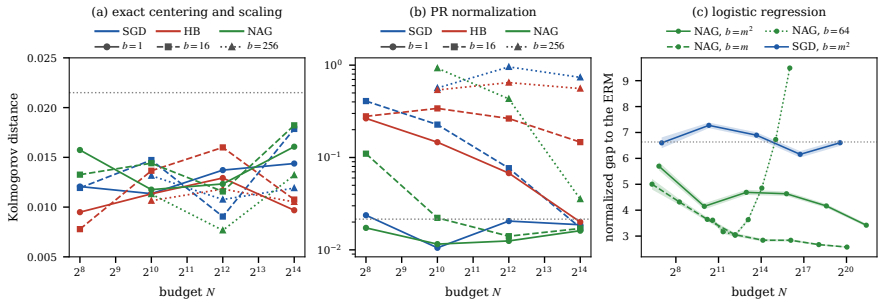

Figure 4: (a, b) Kolmogorov distance to the standard normal law for untransformed income after (a) and PR (b) normalization; colors mark the methods, line styles the batch sizes, and the dotted line is the 5% critical value for 4000 draws. (c) Normalized gap to the ERM in the logistic model; the dotted line is the normalized error of the ERM, and bands are ± 2 Monte Carlo standard errors.

Practical Intervals. Table 6 reports, for the age coefficient at $N = 2^{16}$ and $\kappa = 1.5 \times 10^4$, the coverage of nominal 95% intervals and, in parentheses, their median width relative to the oracle interval, at a small batch and past the breakdown of SGD. When the center is calibrated, the plug-in interval matches the oracle interval. When it is not, no variance estimate restores coverage at a comparable width: the plug-in interval undercovers, and the trajectory-based intervals cover only by becoming several times wider. HB inflates them already at $b = 16$, because its variance is not the PR variance.

Table 6: Coverage and Relative Width of Practical Intervals

Batch	Method	Oracle	Plug-in	Batch means	Random scaling
$b = 16$	SGD	0.93	0.93 (1.0)	0.91 (0.95)	0.92 (1.1)
	HB	0.80	0.81 (1.0)	1.00 (6.9)	1.00 (4.4)
	NAG	0.94	0.94 (1.0)	0.96 (1.1)	0.95 (1.2)
$b = 2^{11}$	SGD	0.00	0.00 (1.0)	0.91 (4.8)	1.00 (7.8)
	HB	0.01	0.01 (1.0)	0.95 (9.6)	1.00 (33)
	NAG	0.91	0.91 (1.0)	1.00 (12)	1.00 (7.6)