A SPECTRAL ANALYSIS OF AVERAGED STOCHASTIC GRADIENT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Averaging is central to the statistical behavior of stochastic gradient methods, yet existing analyses often couple the optimization dynamics, the averaging rule, and the stochastic covariance recursion. Moreover, label noise is commonly summarized by a uniform relative-noise condition, which can obscure its distribution across spectral directions. We develop a modular spectral-readout theory for constant-stepsize stochastic gradient dynamics in least-squares regression. The framework separates an algorithm-dependent transfer kernel, a deterministic readout filter, and a second-moment feedback closure. For averaged SGD, we derive finite-time spectral bounds under a general positive trace-class label-noise covariance $\mathbf{\Sigma}$, without requiring $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, with matching feedback lower bounds after a noise-dependent warmup. The bounds retain the directional noise profile and identify convergent power-law regimes with no finite relative-noise constant. For stochastic heavy-ball, a direction-wise block closure exposes a self-damping mechanism and controls trace-class noise over the reference stepsize region. For full averaging, we further obtain matching risk bounds after warmup and under a fixed stability margin. Synthetic experiments and frozen-feature linear prediction corroborate the predicted filters and rates.

1 INTRODUCTION

Stochastic gradient methods can generalize in highly overparameterized problems (Polyak & Juditsky, 1992; Bottou et al., 2018), even when the learned predictor fits the training data nearly perfectly. Understanding this form of benign overfitting (Belkin et al., 2019; Bartlett et al., 2020; Hastie et al., 2022; Koren et al., 2022) requires more than comparing the number of parameters with the sample size: the covariance spectrum determines which directions are learned and how noise in those directions affects prediction risk. Population least squares with a decaying covariance spectrum provides a canonical setting for studying this question. Although its objective is quadratic, the dynamics of stochastic gradient descent (SGD) remain nontrivial (Bach & Moulines, 2013; Défossez & Bach, 2015; Dieuleveut et al., 2020). Each update contains additive label noise as well as multiplicative randomness from the sampled covariates, and the two sources interact differently across spectral directions.

Averaging adds a second layer to these spectral dynamics. The final iterate, full average, and tail average assign different weights to errors generated at different times (Dieuleveut & Bach, 2016). Combined with the SGD recursion, these temporal weights produce a direction-dependent readout filter. The filter determines how much of the initial error and stochastic noise in each direction reaches the reported predictor. Its behavior across the covariance spectrum gives rise to the effective dimensions and spectral cutoffs in finite-time risk bounds.

Existing work has obtained sharp bias–variance bounds for averaged SGD and related momentum methods (Défossez & Bach, 2015; Zou et al., 2021; Wu et al., 2022; Li et al., 2024; Meterez et al., 2025; Li & Gu, 2025). Spectral operator arguments analyze constant-stepsize SGD in overparameterized linear regression, while augmented-state analyses handle second-order dynamics. These proofs, however, often treat the stochastic covariance recursion, temporal averaging, and spectral summation together. It is therefore difficult to distinguish the contributions of the optimization dynamics, the readout rule, and stochastic feedback.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The treatment of label noise presents a further limitation. Let $\mathbf{H} := \mathbb{E}[\mathbf{x} \otimes \mathbf{x}]$, $\mathbf{\Sigma} := \mathbb{E}[\varepsilon^2 \mathbf{x} \otimes \mathbf{x}]$ denote the feature covariance and label-noise covariance. A common assumption (Zou et al., 2021; Wu et al., 2022; Li et al., 2024; Li & Gu, 2025) is the uniform relative-noise condition $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$. It yields clean effective-dimension bounds but can exclude trace-class noise: $\|\mathbf{H}^{-1/2}\mathbf{\Sigma}\mathbf{H}^{-1/2}\|$ may be infinite even when $\mathrm{tr}(\mathbf{\Sigma}) < \infty$ and the averaged prediction risk converges. Replacing directional noise by a single relative-noise constant can thus conceal convergent regimes.

These observations motivate the following question:

> *Can a common spectral framework separate the effects of stochastic dynamics, the readout rule, and directional noise covariance?*

We answer this question for full and tail averaging of SGD and extend the analysis to stochastic heavy-ball (SHB). Our starting point is an exact filtered-innovation identity. The mean recursion determines an algorithm-dependent transfer kernel, the averaging rule determines a readout filter, and a positive feedback closure controls the state-dependent covariance. This separates changes in the averaging window from changes in the underlying optimization dynamics.

For constant-stepsize SGD, the resulting directional filters recover the classical effective-dimension structure for full averaging and the two-cutoff structure for tail averaging. Under the lower fourth-moment comparison and after a noise-dependent warmup, the variance satisfies $V_{S,M} \asymp \mathcal{G}_{T,M}(\mathbf{\Sigma}) + \eta\,\mathrm{tr}(\mathbf{\Sigma})\mathcal{G}_{T,M}(\mathbf{H})$. The first term measures label noise passed directly through the readout, while the second captures covariance generated by stochastic Hessian fluctuations. The corresponding upper bound holds without the lower comparison or warmup. Under $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, this reduces to a familiar effective-dimension bound (Zou et al., 2021; Li & Gu, 2025; Li et al., 2024); otherwise, the rate depends on both $\mathbf{H}$ and $\mathbf{\Sigma}$. For power-law spectra $\lambda_i \asymp i^{-p}$ and $\Sigma_{ii} \asymp i^{-q}$ with $1 < q < p$, no finite relative-noise constant exists, yet our bounds give an explicit polynomial decay rate. This identifies convergent regimes beyond uniform relative-noise analysis.

For SHB, each spectral direction follows a second-order recursion whose roots can change type across the spectrum. We lift each mode to a two-dimensional state and use a block Lyapunov equation to control its covariance. Retaining the directional negative term yields a self-damped closure under trace-class noise. For full averaging, we further obtain matching risk bounds after a noise-dependent warmup and under a fixed stability margin.

Our main contributions are summarized as follows.

- **A modular spectral analysis of stochastic gradient readouts.** We derive an exact representation separating the algorithm-dependent transfer kernel, the averaging-dependent readout filter, and the second-moment feedback closure. It provides a common route to finite-time risk bounds for SGD and SHB.
- **Direction-sensitive risk bounds beyond uniform relative noise.** For full and tail averaging of SGD, we retain a general positive trace-class label-noise covariance and distinguish directly filtered noise from covariance generated by stochastic Hessian fluctuations. Matching feedback bounds hold after warmup, and a power-law specialization gives convergent regimes without a finite relative-noise constant.
- **Sharp closure for second-order momentum dynamics.** For SHB, retaining the directional negative term yields a self-damped feedback closure under general trace-class noise. For full averaging, we further obtain matching direct-noise and feedback bounds after warmup and under a fixed stability margin.
- **Empirical validation on synthetic and frozen-feature models.** Synthetic experiments examine the finite-time filters, directional noise dependence, power-law decay rates, and SHB predictions. A frozen-feature linear-prediction experiment tests whether the spectral predictions remain informative on real data.

2 Preliminaries

We first specify the population least-squares model and the moment conditions used throughout the paper. We then describe SGD through general linear readouts and separate the effects of initialization

and label noise. These definitions allow the next section to analyze the readout filter and stochastic covariance feedback separately.

2.1 Problem setting and assumptions

We study population least squares with fresh samples. Let $(\mathbf{x}, y)$ be a random pair with x taking values in a finite-dimensional Euclidean space or a separable Hilbert space, and define $L(\mathbf{w}) := \frac{1}{2}\mathbb{E}\big(y - \langle \mathbf{x}, \mathbf{w}\rangle\big)^2$. Let $\mathbf{w}^*$ be a population minimizer and write $\varepsilon := y - \langle x, \mathbf{w}^*\rangle$, $\mathbf{H} := \mathbb{E}[\mathbf{x} \otimes \mathbf{x}]$, $\mathbf{\Sigma} := \mathbb{E}[\varepsilon^2 \mathbf{x} \otimes \mathbf{x}]$. The first-order optimality condition is the vector identity $\mathbb{E}[\varepsilon \mathbf{x}] = 0$. We do not assume that ε is independent of x, nor do we assume $\mathbb{E}[\varepsilon \mid \mathbf{x}] = 0$. Both $\mathbf{H}$ and $\mathbf{\Sigma}$ are positive semidefinite because, for every deterministic $\mathbf{u}$, $\langle \mathbf{u}, \mathbf{H}\mathbf{u}\rangle = \mathbb{E}\langle \mathbf{x}, \mathbf{u}\rangle^2 \geq 0$, $\langle \mathbf{u}, \Sigma \mathbf{u}\rangle = \mathbb{E}\big[\varepsilon^2 \langle \mathbf{x}, \mathbf{u}\rangle^2\big] \geq 0$. We assume that $\mathbf{H}$ is positive and trace class. Let $\mathbf{H}\mathbf{u}_i = \lambda_i \mathbf{u}_i$, $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, $\sum_i \lambda_i < \infty$. No positive lower bound on the spectrum is imposed. For a positive semidefinite trace-class operator $\mathbf{C}$, define the fourth-moment and centered fourth-moment operators

$$\mathcal{M}(\mathbf{C}) := \mathbb{E}[(\mathbf{x} \otimes \mathbf{x})\mathbf{C}(\mathbf{x} \otimes \mathbf{x})], \qquad \mathcal{N}(\mathbf{C}) := \mathcal{M}(\mathbf{C}) - \mathbf{H}\mathbf{C}\mathbf{H}.$$

Equivalently,

$$\mathcal{N}(\mathbf{C}) = \mathbb{E}[(\mathbf{x} \otimes \mathbf{x} - \mathbf{H})\mathbf{C}(\mathbf{x} \otimes \mathbf{x} - \mathbf{H})] \succeq 0.$$

Assumption 2.1 (Fourth-moment comparison)**.** There exist dimension-free constants $\psi < \infty$ and $\beta > 0$ such that, for every positive semidefinite trace-class operator C,

$$\mathcal{M}(\mathbf{C}) \preceq \psi \operatorname{tr}(\mathbf{H}\mathbf{C})\mathbf{H}, \qquad \mathcal{N}(\mathbf{C}) \succeq \beta \operatorname{tr}(\mathbf{H}\mathbf{C})\mathbf{H}.$$

The upper comparison is used for upper bounds and stochastic feedback closure. The lower comparison is an additional nondegeneracy condition used only for matching lower bounds.

Assumption 2.2 (Trace-class label noise and SGD stability)**.** The label-noise covariance satisfies $0 < \operatorname{tr}(\Sigma) < \infty$. For constant-stepsize SGD, $0 < \eta\lambda_1 < 1$ and $m := \psi \sum_i \frac{\eta\lambda_i}{2-\eta\lambda_i} < 1$.

Remark 2.3 (Comparison with the standard trace condition). Under $0 < \eta\lambda_1 < 1$, $\frac{\eta\psi}{2} \operatorname{tr}(\mathbf{H}) \leq m = \psi \sum_i \frac{\eta\lambda_i}{2-\eta\lambda_i} \leq \eta\psi \operatorname{tr}(\mathbf{H})$. Consequently, the sufficient condition $\eta\psi \operatorname{tr}(\mathbf{H}) < 1$ used in earlier analyses implies $m < 1$, whereas the converse need not hold. The renewal-mass condition therefore certifies a strictly larger stepsize region, although the enlargement is at most a factor of two. In particular, $m < 1$ still requires $\eta\psi \operatorname{tr}(\mathbf{H}) < 2$ and should not be interpreted as allowing an unrestricted large-$\eta \operatorname{tr}(\mathbf{H})$ regime.

2.2 SGD and linear readouts

Let $(\mathbf{x}_t, y_t)_{t \geq 0}$ be independent copies of $(\mathbf{x}, y)$. With $\mathbf{e}_t := \mathbf{w}_t - \mathbf{w}^*$, constant-stepsize SGD obeys $\mathbf{e}_{t+1} = (\mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{e}_t + \eta\varepsilon_t \mathbf{x}_t$. For deterministic weights $\alpha_0, \ldots, \alpha_{T-1}$, define the linear readout $\widehat{\mathbf{e}}_\alpha := \sum_{t=0}^{T-1} \alpha_t \mathbf{e}_t$, $\alpha = (\alpha_0, \ldots, \alpha_{T-1})$. The full average (Polyak & Juditsky, 1992), a tail average (Jain et al., 2017; Dieuleveut et al., 2020; Jain et al., 2018), and a recursively initialized exponential moving average (Mandt et al., 2017; Yaida, 2018) are obtained from different choices of $\{\alpha_t\}_t$. For the tail average, write $T = S + M$ and set $\bar{\mathbf{e}}_{S,M} := \frac{1}{M} \sum_{t=S}^{T-1} \mathbf{e}_t$. Full averaging is the case $S = 0$. The normal equation $\mathbb{E}[\varepsilon \mathbf{x}] = 0$ implies that prediction risk has a simple quadratic form: $L(\mathbf{w}^* + \mathbf{u}) - L(\mathbf{w}^*) = \frac{1}{2}\|\mathbf{u}\|_H^2$, $\|\mathbf{u}\|_{\mathbf{H}}^2 := \langle \mathbf{u}, \mathbf{H}\mathbf{u}\rangle$. We will use this identity to measure the error of each readout.

2.3 Bias and noise processes

To distinguish initialization error from label-noise error, split the SGD recursion into two processes driven by the same covariate samples:

$$\mathbf{b}_{t+1} = (\mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{b}_t, \qquad \mathbf{b}_0 = \mathbf{e}_0,$$
$$\mathbf{v}_{t+1} = (\mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{v}_t + \eta\varepsilon_t \mathbf{x}_t, \ \mathbf{v}_0 = 0.$$

Then $\mathbf{e}_t = \mathbf{b}_t + \mathbf{v}_t$ for every realization of the samples. Apply the same readout weights to each process and define

$$B_\alpha := \frac{1}{2}\mathbb{E}\|\widehat{\mathbf{b}}_\alpha\|_{\mathbf{H}}^2, \quad V_\alpha := \frac{1}{2}\mathbb{E}\|\widehat{\mathbf{v}}_\alpha\|_{\mathbf{H}}^2, \quad R_\alpha := \frac{1}{2}\mathbb{E}\|\widehat{\mathbf{e}}_\alpha\|_{\mathbf{H}}^2. \tag{2.1}$$

For the weights in tail-average, we write these quantities as $B_{S,M}$, $V_{S,M}$, and $R_{S,M}$. The two processes use the same covariates, so their cross term cannot generally be discarded under the normal equation alone. Cauchy–Schwarz instead gives

$$(\sqrt{V_\alpha} - \sqrt{B_\alpha})^2 \le R_\alpha \le (\sqrt{V_\alpha} + \sqrt{B_\alpha})^2. \tag{2.2}$$

In particular, if $B_\alpha = o(V_\alpha)$, then $R_\alpha/V_\alpha \to 1$, without a conditional mean-zero assumption. We next focus on V_α. Its analysis requires tracking both the noise that passes directly through the readout and the additional covariance created as stochastic updates act on earlier errors.

3 Spectral Readouts and Feedback Closure

Let $\mathbf{A} := \mathbf{I} - \eta\mathbf{H}$ and $\mathbf{C}_t := \mathbb{E}[\mathbf{v}_t \otimes \mathbf{v}_t]$. A direct conditional expectation gives the exact covariance recursion

$$\mathbf{C}_{t+1} = \mathbf{A}\mathbf{C}_t\mathbf{A} + \eta^2\mathcal{N}(\mathbf{C}_t) + \eta^2\mathbf{\Sigma}. \tag{3.1}$$

The last term injects label noise at each step, while $\mathcal{N}(\mathbf{C}_t)$ accounts for the randomness of the sampled Hessian. Define the martingale innovation $\boldsymbol{\xi}_{t+1} := (\mathbf{H} - \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{v}_t + \varepsilon_t\mathbf{x}_t$. Then $\mathbf{v}_{t+1} = \mathbf{A}\mathbf{v}_t + \eta\boldsymbol{\xi}_{t+1}$, $\mathbb{E}[\boldsymbol{\xi}_{t+1} \otimes \boldsymbol{\xi}_{t+1}] = \mathcal{N}(\mathbf{C}_t) + \mathbf{\Sigma}$, and innovations at distinct times are orthogonal in L_2. For a deterministic readout α, the effect of the innovation at time r on the reported predictor is determined by

$$K_r^\alpha(\mathbf{A}) := \sum_{t=r+1}^{T-1} \alpha_t \mathbf{A}^{t-1-r}, \quad 0 \le r \le T-2, \tag{3.2}$$

This kernel combines the mean SGD dynamics with the chosen readout weights. For a positive semidefinite trace-class operator $\mathbf{Q}$, write $\mathcal{F}_\alpha(\mathbf{Q}) := \frac{\eta^2}{2}\sum_{r=0}^{T-2} \operatorname{tr}\big(\mathbf{H}[K_r^\alpha(\mathbf{A})]^2\mathbf{Q}\big)$. This functional depends on the deterministic dynamics and the readout, but not on the random feedback trajectory. Unlike $\mathbf{C}_t$, the functional $\mathcal{F}_\alpha$ is determined entirely by the population covariance, stepsize, and readout.

Proposition 3.1 (Exact readout identity)**.** *For every deterministic linear readout,*

$$V_\alpha = \mathcal{F}_\alpha(\mathbf{\Sigma}) + \frac{\eta^2}{2}\sum_{r=0}^{T-2} \operatorname{tr}\big(\mathbf{H}[K_r^\alpha(\mathbf{A})]^2\mathcal{N}(\mathbf{C}_r)\big). \tag{3.3}$$

Moreover, in the eigenbasis of $\mathbf{H}$, $\mathcal{F}_\alpha(\mathbf{Q}) = \frac{\eta^2}{2}\sum_i \lambda_i Q_{ii} \sum_{r=0}^{T-2}[K_{i,r}^\alpha]^2$, *where* $Q_{ii} := \langle \mathbf{u}_i, \mathbf{Q}\mathbf{u}_i\rangle$ *and* $K_{i,r}^\alpha := \sum_{t=r+1}^{T-1}\alpha_t(1-\eta\lambda_i)^{t-1-r}$.

Remark 3.2. The first term of Eq. (3.3) passes the original label-noise covariance through the readout filter. The second passes through the same filter, but its input $\mathcal{N}(\mathbf{C}_r)$ depends on the earlier stochastic trajectory. Notice that Proposition 3.1 does not require $\mathbf{Q}$ to commute with $\mathbf{H}$: the filter uses its diagonal entries in the eigenbasis of $\mathbf{H}$.

We now control the feedback term. Let $r_t := \operatorname{tr}(\mathbf{H}\mathbf{C}_t)$ and $h_j := \eta^2\sum_i \lambda_i^2(1-\eta\lambda_i)^{2j}$. The fourth-moment comparisons in Assumption 2.1 turn Eq. (3.1) into upper and lower positive renewal bounds for r_t. In the upper bound, the condition $m < 1$ from Assumption 2.2 ensures that the feedback kernel has total mass below one. In particular, $\sup_{t\ge 0} r_t \le \eta\operatorname{tr}(\mathbf{\Sigma})/(1-m)$. This controls the feedback from above at every time. A lower bound for the filtered output also requires enough time for the injected noise to build up. To state a matching feedback lower bound, define $k_0 = k_0(\mathbf{\Sigma}, \eta)$ as the smallest nonnegative integer such that

$$\sum_i \frac{\Sigma_{ii}}{2-\eta\lambda_i}(1-\eta\lambda_i)^{2k_0} \le \frac{1}{2}\sum_i \frac{\Sigma_{ii}}{2-\eta\lambda_i}. \tag{3.4}$$

Such an index exists because $\mathbf{\Sigma}$ is trace class. It measures when a fixed fraction of the stationary contribution of the label-noise covariance has accumulated; it need not be uniform over all admissible noise covariances. For tail-averaging weights, denote the functional in Section 3 by $\mathcal{F}_{S,M}$.

Theorem 3.3 (Effective covariance bounds)**.** *Suppose the fourth-moment upper comparison* $\mathcal{M}(\mathbf{C}) \preceq \psi\operatorname{tr}(\mathbf{H}\mathbf{C})\mathbf{H}$ *and Assumption 2.2 hold. Then, for every* $S \ge 0$ *and* $M \ge 1$,

$$\mathcal{F}_{S,M}(\mathbf{\Sigma}) \le V_{S,M} \le \mathcal{F}_{S,M}\left(\mathbf{\Sigma} + \psi/(1-m)\eta\operatorname{tr}(\mathbf{\Sigma})\mathbf{H}\right).$$

If, in addition, the lower comparison $\mathcal{N}(\mathbf{C}) \succeq \beta\operatorname{tr}(\mathbf{H}\mathbf{C})\mathbf{H}$ *holds and* $S \ge 2k_0$, *then*

$$V_{S,M} \ge \mathcal{F}_{S,M}\left(\mathbf{\Sigma} + \beta/8\eta\operatorname{tr}(\mathbf{\Sigma})\mathbf{H}\right).$$

Remark 3.4. The upper bound in Theorem 3.3 holds from the start of the trajectory: the upper fourth-moment comparison and $m < 1$ keep the covariance feedback uniformly controlled. The direct-noise lower bound $\mathcal{F}_{S,M}(\mathbf{\Sigma}) \le V_{S,M}$ also holds without warmup because the feedback contribution is nonnegative. The matching feedback lower bound requires two additional ingredients. By the definition of k_0, the direct label-noise contribution ensures $r_t = \operatorname{tr}(\mathbf{H}\mathbf{C}_t) \ge \eta \operatorname{tr}(\mathbf{\Sigma})/4$ for every $t \ge k_0$. When $S \ge 2k_0$, at least half of the tail readout kernel mass is assigned to innovations generated at these later times. The lower fourth-moment comparison then produces the additional feedback term in the lower bound of Theorem 3.3. The theorem therefore applies the same deterministic readout functional to effective covariance operators; it does not assert that either operator is the covariance of an individual innovation. The warmup condition affects only the matching feedback lower bound and may depend on the spectral location of $\mathbf{\Sigma}$.

4 Main Results

4.1 Tail averaging and the two spectral cutoffs

We begin with the deterministic filter left by Theorem 3.3. For $\delta = \eta\lambda \in (0,1)$ and $a = 1-\delta$, define the scalar tail energy $\Gamma_{S,M}(\delta) := \sum_{r=0}^{T-2} K_r(a)^2$, $T = S + M$. Innovations arriving before the averaging window and those arriving within it contribute differently. Splitting the sum accordingly gives

$$\Gamma_{S,M}(\delta) = \frac{(1-a^M)^2(1-a^{2S})}{M^2(1-a)^2(1-a^2)} + \frac{1}{M^2(1-a)^2}\sum_{\ell=1}^{M-1}(1-a^\ell)^2. \tag{4.1}$$

The two pieces are useful for the proof; their combined behavior has a simpler description.

Lemma 4.1 (Three regimes from two time scales)**.** *Uniformly over $S \ge 0$, $M \ge 1$, $T \ge 2$, and $0 < \delta < 1$,*

$$\Gamma_{S,M}(\delta) \asymp \begin{cases} T, & 0 < \delta \le T^{-1}, \\ \delta^{-1}, & T^{-1} < \delta \le M^{-1}, \\ (M\delta^2)^{-1}, & M^{-1} < \delta < 1. \end{cases} \tag{4.2}$$

The comparison constants are universal.

The first threshold asks whether an eigen-direction has had enough time to respond over the total horizon. The second asks whether its fluctuations have been averaged over a sufficiently long window. Accordingly, define

$$k^\dagger := \max\{i : \lambda_i \ge 1/\eta T\}, \quad k^* := \max\{i : \lambda_i \ge 1/\eta M\}. \tag{4.3}$$

An empty maximum is understood as zero. Because $M \le T$, the two cutoffs satisfy $k^* \le k^\dagger$. Applying the three regimes to each spectral direction gives the filter below. We index it by (T, M): the total horizon $T = S + M$ determines how long each direction has evolved, while M determines how long its fluctuations are averaged.

Theorem 4.2 (Two-cutoff spectral formula)**.** *For every positive semidefinite trace-class operator $\mathbf{Q} \succeq 0$, $\mathcal{F}_{S,M}(\mathbf{Q}) \asymp \mathcal{G}_{T,M}(\mathbf{Q})$, where*

$$\mathcal{G}_{T,M}(\mathbf{Q}) := \frac{1}{M}\sum_{i \le k^*}\frac{Q_{ii}}{\lambda_i} + \eta \sum_{k^* < i \le k^\dagger} Q_{ii} + \eta^2 T \sum_{i > k^\dagger} \lambda_i Q_{ii}.$$

Remark 4.3. The three sums distinguish directions whose fluctuations are averaged, directions that respond during the run but are not fully averaged over the output window, and directions that barely respond. The formula concerns the readout alone: $\mathbf{Q}$ may be the actual noise covariance or the covariance operator appearing in the feedback bound. Applying it to both sides of Theorem 3.3 yields our SGD risk result.

Theorem 4.4 (Tail-averaged SGD under trace-class label noise)**.** *Under the fourth-moment upper comparison and Assumption 2.2, for every $S \ge 0$ and $M \ge 1$,*

$$V_{S,M} \lesssim \mathcal{G}_{T,M}(\mathbf{\Sigma}) + \eta \operatorname{tr}(\mathbf{\Sigma})\mathcal{G}_{T,M}(\mathbf{H}), \quad V_{S,M} \gtrsim \mathcal{G}_{T,M}(\mathbf{\Sigma}).$$

If the lower fourth-moment comparison also holds and $S \geq 2k_0$, then

$$V_{S,M} \asymp \mathcal{G}_{T,M}(\mathbf{\Sigma}) + \eta \operatorname{tr}(\mathbf{\Sigma})\mathcal{G}_{T,M}(\mathbf{H}). \tag{4.4}$$

The constants depend only on β, ψ, and the stability margin $1 - m$ whenever the corresponding quantities are used.

Remark 4.5 (Two sources of variance). The two terms in Eq. (4.4) need not have the same spectral behavior. The first filters the label-noise covariance direction by direction. The second describes noise created when random sample Hessians act on errors accumulated earlier in the trajectory. Even if $\mathbf{\Sigma}$ is concentrated in a few directions, this feedback can contribute across the spectrum of $\mathbf{H}$. If $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, the first term admits the usual effective-dimension upper bound. This comparison is useful for connecting the theorem to existing bounds, but is not required by the theorem: its two terms can instead be evaluated using the actual noise spectrum.

4.2 Power-law spectra beyond bounded relative noise

The difference between the two variance terms is visible in a power-law model. Suppose $\mathbf{H}$ and $\mathbf{\Sigma}$ are diagonal in the same basis, with $\lambda_i = i^{-p}$, $\Sigma_{ii} = i^{-q}$, $p, q > 1$. Fix p, q and an admissible stepsize η, and consider the balanced tail $S = M$, so that $T = 2M$. The two spectral cutoffs satisfy $k^* \asymp (\eta M)^{1/p}$ and $k^\dagger \asymp (2\eta M)^{1/p}$, and are therefore of the same order for fixed η. Summing the direct filtered-noise contribution over the resulting spectral regions gives

$$\mathcal{G}_{2M,M}(\mathbf{\Sigma}) \asymp \begin{cases} M^{-(q-1)/p}, & 1 < q < p+1, \\ (\log M)/M, & q = p+1, \\ M^{-1}, & q > p+1, \end{cases} \tag{4.5}$$

whereas $\mathcal{G}_{2M,M}(\mathbf{H}) \asymp M^{-(1-1/p)}$. The upper fourth-moment comparison therefore gives $V_{M,M} \lesssim \mathcal{G}_{2M,M}(\mathbf{\Sigma}) + M^{-(1-1/p)}$. If the lower fourth-moment comparison also holds, then $k_0(\mathbf{\Sigma}, \eta)$ is finite and independent of M. Hence, for all sufficiently large M such that $M \geq 2k_0$, the feedback lower bound matches the preceding upper bound, and

$$V_{M,M} = \Theta\left(M^{-\alpha}\right), \quad \alpha := \min\left\{(q-1)/p, 1 - 1/p\right\}. \tag{4.6}$$

This calculation shows where retaining $\mathbf{\Sigma}$ matters. When $1 < q < p$, the directly filtered noise is the slower term and determines the rate. When $q \geq p$, the stochastic feedback term determines the asymptotic rate in this model. In particular, the logarithmic factor at $q = p + 1$ belongs to the *direct-noise contribution*; it does not determine the total variance. For the initialization $\mathbf{e}_0 = \mathbf{u}_1$, the bias risk satisfies

$$B_{M,M} = \Theta\left(M^{-(2-1/p)}\right) = o(V_{M,M}).$$

Although the deterministic response of the leading direction decays rapidly, sampled Hessians spread its error across spectral directions; the proof of this bias rate is given in Appendix G. The risk sandwich in Eq. (2.2) therefore gives $R_{M,M} = \Theta(M^{-\alpha})$. Finally, for $1 < q < p$, we have $\Sigma_{ii}/\lambda_i = i^{p-q} \to \infty$. No finite constant σ_{rel}^2 can then satisfy $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, although the prediction risk converges at the rate above. This is the regime that a uniform relative-noise reduction cannot describe. The asymptotic statement is pointwise in fixed p and q: for fixed $p > 1$ and admissible η, Lemma G.1 gives $\log(\eta k_0) \sim p \log 2/(q-1)$ as $q \downarrow 1$. Thus k_0 is independent of M for fixed q, but the condition $M \geq 2k_0$ is not uniform near $q = 1$.

Remark 4.6 (Instance-wise sharpness versus minimax optimality). The rate in Eq. (4.6) is instance-wise sharp for the constant-stepsize tail-averaged SGD dynamics considered here. In particular, under the lower fourth-moment comparison, both $\mathcal{G}_{T,M}(\mathbf{\Sigma})$ and the feedback contribution $\eta \operatorname{tr}(\mathbf{\Sigma})\mathcal{G}_{T,M}(\mathbf{H})$ appear in the lower bound. Thus the branch $1 - 1/p$ is not an artifact of the upper-bound proof. We do not claim minimax optimality over a statistical model class. Such a comparison would require specifying a source class, its radius, the admissible noise distributions, and the estimator class. Other estimators or stepsize schedules may avoid the limitation specific to constant-stepsize tail-averaged SGD.

4.3 Other averaging rules

The readout calculation also applies to outputs other than a tail average. For full averaging, the directional noise energy satisfies $\delta^2 \sum_{r=0}^{T-2} K_r (1-\delta)^2 \asymp \min\{T^{-1}, T\delta^2\}$. The transition occurs

at $\lambda \asymp (\eta T)^{-1}$: full averaging has one time scale, whereas a tail average can have two. For completeness, a recursively initialized exponential moving average (EMA) also has an explicit readout kernel. Let $\theta = 1 - \beta_{\mathrm{ema}}$. Its mean filter and innovation kernel are

$$W_T^{\mathrm{ema}}(a) = \beta_{\mathrm{ema}}^T + \theta \frac{a^T - \beta_{\mathrm{ema}}^T}{a - \beta_{\mathrm{ema}}}, \quad K_{T,r}^{\mathrm{ema}}(a) = \theta \frac{a^{T-1-r} - \beta_{\mathrm{ema}}^{T-1-r}}{a - \beta_{\mathrm{ema}}}. \tag{4.7}$$

At $a = \beta_{\mathrm{ema}}$, these expressions are understood by continuity. When $T\theta \geq 1$, the corresponding directional noise energy satisfies $E_{T,\theta}^{\mathrm{ema}}(\delta) \asymp \min\{\theta, \delta, T\delta^2\}$. The two transitions are therefore $\lambda = \theta/\eta$ and $\lambda = (\eta T)^{-1}$. These corollaries illustrate how the deterministic readout changes the spectral filter.

5 Heavy-Ball SGD

The section changed the readout while keeping the SGD recursion fixed. We now keep the same filtered-innovation viewpoint and change the recursion. Momentum gives each spectral direction a second-order response, so its impulse energy and stochastic feedback must be controlled differently.

5.1 Dynamics and directional response

Stochastic heavy-ball (SHB) follows $\mathbf{e}_{t+1} = ((1+\mu)\mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{e}_t - \mu \mathbf{e}_{t-1} + \eta \varepsilon_t \mathbf{x}_t$, $0 \leq \mu < 1$. As for SGD, we separate the process driven by label noise from the process driven by initialization. Let $\mathbf{v}_t^{\mathrm{hb}}$ denote the former, with $\mathbf{v}_{-1}^{\mathrm{hb}} = \mathbf{v}_0^{\mathrm{hb}} = 0$, and define V_α^{hb} as the prediction risk of its linear readout: $V_\alpha^{\mathrm{hb}} := \frac{1}{2}\mathbb{E}\left\| \sum_{t=0}^{T-1} \alpha_t \mathbf{v}_t^{\mathrm{hb}} \right\|_{\mathbf{H}}^2$. Set $\tau := 1 - \mu$ and $\nu := \eta/\tau$. In direction i, the deterministic impulse response has generating function $G_i(z) := [1 - (1 + \mu - \eta\lambda_i)z + \mu z^2]^{-1} = \sum_{t \geq 0} g_{i,t} z^t$. Unlike the geometric response of SGD, this sequence may decay with or without oscillation. Its squared mass nevertheless has one expression throughout the stable range $0 < \eta\lambda_i < 2(1+\mu)$:

$$L_i := \sum_{t \geq 0} g_{i,t}^2 = \frac{1 + \mu}{\tau \eta \lambda_i [2(1+\mu) - \eta\lambda_i]}.$$

The impulse response replaces the powers of $1 - \eta\lambda_i$ in the SGD readout kernel. The remaining difficulty is to control how its stochastic covariance feeds back into later updates.

5.2 Feedback closure under trace-class noise

A coarse feedback bound based on L_i becomes large as a characteristic root approaches -1. The covariance recursion, however, contains a negative directional term that offsets part of this growth. The block Lyapunov analysis in Appendix I retains that term. To state the result, let $\ell_i := (\eta\lambda_i)^2 L_i$, $h_* := \sum_i \ell_i/(1+\ell_i)$, $\Delta_* := 1 - \psi h_*$, and, for a positive semidefinite trace-class operator $\mathbf{Q}$, set $q_*(\mathbf{Q}) := \sum_i [\ell_i/(1+\ell_i)] Q_{ii}/\lambda_i$. The factor $(1+\ell_i)^{-1}$ is the self-damping produced by the directional covariance term. Indeed, $\ell_i/(1+\ell_i) = \nu\lambda_i/[2 + \frac{2\mu}{1+\mu}\nu\lambda_i]$, so $q_*(\mathbf{\Sigma}) \leq (\nu/2)\operatorname{tr}(\mathbf{\Sigma})$ even when the relative-noise constant is infinite. The condition $\psi\nu \operatorname{tr}(\mathbf{H}) < 2$ also ensures $\Delta_* > 0$. For readout weights α, define the SHB directional kernel and filter by

$$K_{i,r}^{\alpha,\mathrm{hb}} := \sum_{t=r+1}^{T-1} \alpha_t g_{i,t-1-r}, \quad \mathcal{F}_\alpha^{\mathrm{hb}}(\mathbf{Q}) := \frac{\eta^2}{2} \sum_i \lambda_i Q_{ii} \sum_{r=0}^{T-2} \left(K_{i,r}^{\alpha,\mathrm{hb}} \right)^2.$$

As in the SGD case, this functional describes how a given covariance passes through the deterministic dynamics and the readout. The following theorem controls the covariance generated along the stochastic trajectory.

Theorem 5.1 (SHB feedback closure under trace-class noise)**.** *Suppose $\psi\nu \operatorname{tr}(\mathbf{H}) < 2$. Then, for every deterministic linear readout,*

$$\mathcal{F}_\alpha^{\mathrm{hb}}(\mathbf{\Sigma}) \leq V_\alpha^{\mathrm{hb}} \leq \mathcal{F}_\alpha^{\mathrm{hb}}\left(\mathbf{\Sigma} + \frac{\psi q_*(\mathbf{\Sigma})}{\Delta_*} \mathbf{H} \right). \tag{5.1}$$

If, in addition, $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, then $V_\alpha^{\mathrm{hb}} \leq \frac{\sigma_{\mathrm{rel}}^2}{\Delta_*} \mathcal{F}_\alpha^{\mathrm{hb}}(\mathbf{H})$.

Remark 5.2 (Role of the directional correction). The theorem retains the actual label-noise covariance in the directly filtered term. Its feedback upper bound depends on $q_*(\mathbf{\Sigma})$, which weights each noise direction by $\ell_i/(1+\ell_i)$ rather than by the raw impulse mass ℓ_i. This distinction is particularly consequential near the negative unit-root boundary, where the raw mass can be large. The lower bound in Eq. (5.1) contains the directly filtered noise only. Unlike Theorem 3.3 for tail-averaged SGD, this result does not give a matching lower bound for the SHB feedback contribution. The relative-noise statement is an upper-bound specialization of the trace-class result.

We next use the filter estimate to match the stochastic feedback contribution. Define the direct-noise state energy $d_t^{\mathrm{hb}} := \eta^2 \sum_i \lambda_i \Sigma_{ii} \sum_{s=0}^{t-1} g_{i,s}^2$, $d_\infty^{\mathrm{hb}} := \eta^2 \sum_i \lambda_i \Sigma_{ii} L_i$. Let $k_0^{\mathrm{hb}} = k_0^{\mathrm{hb}}(\mathbf{\Sigma}, \eta, \mu)$ be the smallest nonnegative integer such that $\eta^2 \sum_i \lambda_i \Sigma_{ii} \sum_{s=k_0^{\mathrm{hb}}}^{\infty} g_{i,s}^2 \le \frac{1}{2} d_\infty^{\mathrm{hb}}$. This index is finite whenever $0 < \mathrm{tr}(\mathbf{\Sigma}) < \infty$.

Corollary 5.3 (Matching full-SHB risk after warmup)**.** *Suppose that the lower fourth-moment comparison also holds and that $\Delta_* \ge \delta$ for some fixed $\delta > 0$. If $T \ge 4k_0^{\mathrm{hb}}$, $T(1-\mu) \ge 256$, then*

$$V_{\mathrm{full}}^{\mathrm{hb}} \asymp \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{\Sigma}) + \nu\, \mathrm{tr}(\mathbf{\Sigma}) \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}),$$

where the comparison constants depend only on β, ψ, and δ.

5.3 Full averaging

The closure applies to any deterministic readout. We now evaluate its filter for full averaging, where the second-order response must be controlled across different characteristic-root regimes. Let $g_{i,-1} := 0$ and $\phi_{i,t} := g_{i,t} - \mu g_{i,t-1}$. For a scalar direction with $d = \eta\lambda$, write ϕ_t for the corresponding sequence. Its full-averaging energy is exactly

$$\mathcal{E}_T^{\mathrm{hb}}(d,\mu) = \frac{1}{T^2} \sum_{L=1}^{T-1} (1-\phi_L)^2. \tag{5.2}$$

Theorem 5.4 (Matching full-SHB filter)**.** *If $0 < d < 2\tau$ and $T\tau \ge 64$, then*

$$\mathcal{E}_T^{\mathrm{hb}}(d,\mu) \asymp \min\left\{T^{-1}, T(d/\tau)^2\right\}, \tag{5.3}$$

with universal comparison constants. Consequently, defining $k_{\mathrm{hb}}(T) := \max\{i : \lambda_i \gtrsim (\nu T)^{-1}\}$, one has $\mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}) \asymp \frac{k_{\mathrm{hb}}(T)}{T} + T\nu^2 \sum_{i > k_{\mathrm{hb}}(T)} \lambda_i^2$.

Remark 5.5. The cutoff $\lambda \asymp (\nu T)^{-1}$ separates directions that are resolved over the finite run from those that are not. It is different from the characteristic-root boundary $\eta\lambda = (1-\sqrt{\mu})^2$: the former concerns the risk of a finite-time readout, while the latter describes the form of the deterministic response. The filter estimate holds across these response regimes under the theorem's conditions.

6 Experiments

We conduct experiments to examine two aspects of our theory. First, we test whether the exact stochastic second-moment recursion describes tail-averaged SGD and whether the predicted power-law decay exponent captures its dependence on the noise spectrum. Second, we investigate whether the averaging effect predicted for SHB remains visible on real neural representations. We present these two main experiments below. Additional results, including the validation of spectral cutoffs, detailed SGD and SHB dynamics, experiments with real covariates and controlled noise, and a CIFAR-10 batch-size ablation, are deferred to Section B.

Power-law synthetic model. We consider a realizable mixture model with covariance spectrum $\lambda_i = i^{-2}$ and noise spectrum $\Sigma_{ii} = i^{-q}$. Constant-stepsize SGD is run for $T = 2M$ iterations, and the final M iterates are averaged. As shown in Figure 1, the exact recursion closely matches Monte Carlo simulations. Moreover, the fitted decay exponents track the finite-window spectral prediction and recover the transition in the theoretical rate $\alpha(q) = \min\{(q-1)/2, 1/2\}$, while remaining clearly separated from the naive $1/M$ rate.

CIFAR-10 Frozen Features. We optimize a linear least-squares head on frozen ImageNet-pretrained ResNet-18 (He et al., 2016) features of CIFAR-10 (Krizhevsky, 2009) using single-sample stochastic gradients. In Figure 2, the last SHB iterate initially reduces the empirical excess

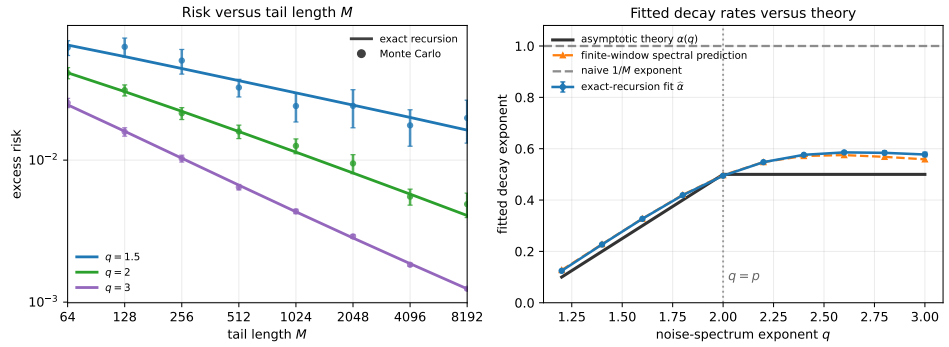

Figure 1: Tail-averaged SGD under power-law covariance and noise spectra. We set $\lambda_i = i^{-2}$ and $\Sigma_{ii} = i^{-q}$. The left panel varies the tail length M and compares the exact second-moment recursion with Monte Carlo estimates over 512 runs. The right panel compares the fitted decay exponents with the asymptotic rate $\alpha(q)$, the finite-window spectral prediction, and the naive $1/M$ exponent.

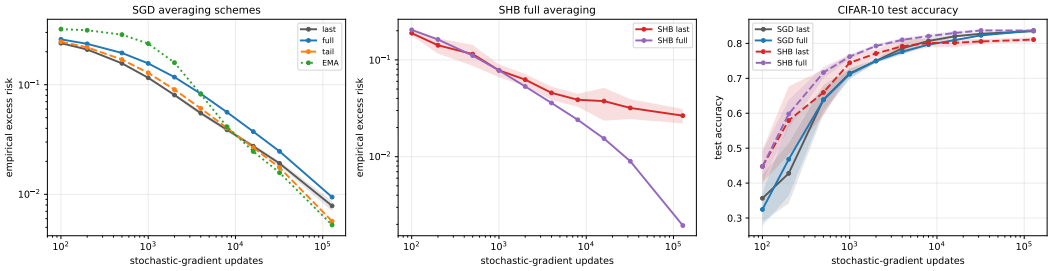

Figure 2: Linear probing on frozen ImageNet-pretrained ResNet-18 features of CIFAR-10. Only the linear head is optimized using single-sample stochastic gradients. The first two panels report empirical excess risk relative to the training-set least-squares solution, and the right panel reports test accuracy. Curves and shaded regions show the mean and one standard deviation over five sampling trajectories.

risk rapidly but eventually reaches a noise-dominated plateau, whereas full averaging continues to reduce it. Full averaging is also associated with higher test accuracy in this experiment. The SGD panel additionally illustrates the distinct finite-time behavior of different averaging schemes. This experiment provides a qualitative test of averaging on real neural features, rather than a direct measurement of the population prediction risk in our theory. Curves and shaded regions report the mean and one standard deviation over five sampling trajectories.

7 Conclusion

We studied how stochastic dynamics and averaging jointly shape prediction risk in overparameterized least squares. The filtered-innovation identity separates the response of a chosen readout from the covariance feedback generated by stochastic updates. For tail-averaged SGD, this yields two spectral cutoffs and risk bounds that retain the directional label-noise covariance. In the power-law model, prediction risk can vanish even when the noise covariance cannot be bounded by any finite multiple of the feature covariance. The bounds also distinguish directly filtered noise from variance generated by stochastic Hessian fluctuations. For stochastic heavy-ball, a block Lyapunov analysis retains a directional correction lost in coarse covariance bounds. This correction controls the feedback mass even when the raw impulse energy becomes large near the negative unit-root boundary. We obtain a trace-class closure and matching full-averaging risk after warmup and away from the stability boundary. In the synthetic experiment, exact second-moment risks agree with Monte Carlo estimates, while the fitted decay exponents track the finite-window spectral prediction and recover the theoretically predicted transition across noise regimes. A frozen ResNet-18 feature experiment compares SGD and heavy-ball readouts on a CIFAR-10 linear prediction task. Extending these results to data reuse and time-varying parameters remains open.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation for language editing and as auxiliary tools for discussing derivation steps, checking proof arguments, and improving the presentation of the theoretical results. All theoretical ideas, assumptions, mathematical claims, and final proofs were developed and independently verified by the authors. All AI-assisted content was carefully reviewed and validated by the authors, who take full responsibility for the correctness and content of the final manuscript.

ETHICS STATEMENT

This work develops a theoretical framework for understanding the spectral behavior of averaging schemes in stochastic optimization. The study uses synthetic linear-regression models and standard public datasets, and does not involve human participants, private information, or sensitive attributes.

REPRODUCIBILITY STATEMENT

We provide the assumptions, algorithmic definitions, spectral quantities, and complete proofs required to verify the theoretical results in the main paper and appendix. In particular, the appendix contains detailed derivations for the generating-function framework and its applications to full averaging, tail averaging, exponential moving averaging, and stochastic heavy-ball dynamics.

The experimental sections report the covariance and noise constructions, averaging parameters, stepsizes, momentum coefficients, iteration budgets, numbers of independent trajectories, and evaluation metrics. Additional experiments use real covariates from the public diabetes dataset and frozen ImageNet-pretrained ResNet-18 features on CIFAR-10. The appendix further specifies the feature preprocessing, least-squares reference solution, mini-batch settings, random-seed protocol, uncertainty estimates, and all relevant hyperparameters. Numerical results are saved together with the plotting data for each experiment. We intend to publicly release the complete implementation and configuration files upon acceptance.

REFERENCES

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

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

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Chi-Tsong Chen. *Linear system theory and design*. Saunders college publishing, 1984.

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pp. 205–213. PMLR, 2015.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

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.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning (ICML)*, pp. 9965–10040. PMLR, 2022.

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:19605–19617, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Xuheng Li and Quanquan Gu. Understanding sgd with exponential moving average: A case study in linear regression. *arXiv preprint arXiv:2502.14123*, 2025.

Xuheng Li, Yihe Deng, Jingfeng Wu, Dongruo Zhou, and Quanquan Gu. Risk bounds of accelerated sgd for overparameterized linear regression. In *International Conference on Learning Representations*, 2024.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18261–18271, 2020.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Alexandru Meterez, Depen Morwani, Costin-Andrei Oncescu, Jingfeng Wu, Cengiz Pehlevan, and Sham Kakade. A simplified analysis of sgd for linear regression with weight averaging. *arXiv preprint arXiv:2506.15535*, 2025.

Anthony N Michel, Ling Hou, and Derong Liu. *Stability of dynamical systems*. Springer, 2008.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020.

Katsuhiko Ogata. *Discrete-time control systems*. Prentice-Hall, Inc., 1995.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

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

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning (ICML)*, pp. 24280–24314. PMLR, 2022.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv preprint arXiv:1810.00004*, 2018.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory (COLT)*, pp. 4633–4635. PMLR, 2021.

A NOTATION

Table 1 collects the notation used in the main results and proofs.

Table 1: Summary of notation.

Symbol	Meaning
Model and assumptions	
$\mathbf{w}^*, \varepsilon$	Population minimizer and residual
$\mathbf{H}, \mathbf{\Sigma}$	Feature and label-noise covariances
$\lambda_i, \mathbf{u}_i$	Eigenpairs of $\mathbf{H}$
Q_{ii}	$\langle \mathbf{u}_i, \mathbf{Q}\mathbf{u}_i \rangle$
$\mathcal{M}, \mathcal{N}$	Fourth-moment and centered fourth-moment operators
ψ, β	Upper and lower fourth-moment constants
η, m	Stepsize and SGD feedback mass
SGD and readouts	
$\mathbf{e}_t, \mathbf{b}_t, \mathbf{v}_t$	Total, initialization, and noise errors
S, M, T	Burn-in, averaging length, and total horizon; $T = S + M$
$\alpha, \widehat{\mathbf{e}}_\alpha$	Readout weights and output error
$B_\alpha, V_\alpha, R_\alpha$	Bias, noise, and total prediction risks
$\mathbf{A}, \mathbf{C}_t, r_t$	Mean operator, noise covariance, and its risk trace
$\boldsymbol{\xi}_{t+1}$	Martingale innovation
$K_r^\alpha, \mathcal{F}_\alpha$	Readout kernel and filter functional
k_0	Noise-dependent warmup index
$k^*, k^\dagger$	Window and horizon spectral cutoffs
$\mathcal{G}_{T,M}$	Two-cutoff spectral functional
Stochastic heavy-ball	
μ, τ, ν	Momentum, $1-\mu$, and η/τ
$g_{i,t}, G_i$	Directional impulse response and generating function
L_i, ℓ_i	Impulse mass and scaled impulse mass
h_*, Δ_*	Self-damped feedback mass and margin
$q_*(\mathbf{Q})$	Direction-weighted source mass
$K_{i,r}^{\alpha,\mathrm{hb}}, \mathcal{F}_\alpha^{\mathrm{hb}}$	SHB readout kernel and filter
$\mathcal{E}_T^{\mathrm{hb}}$	SHB full-averaging directional energy

B ADDITIONAL EXPERIMENTAL RESULTS

This appendix presents additional experiments complementing the two main studies in Section 6. We begin with common experimental and implementation details. We then validate the direction-wise spectral cutoffs predicted by the deterministic filters, examine the stochastic averaging dynamics of SGD and SHB, and extend the evaluation to real covariates with controlled directional noise. Finally, we study how the mini-batch size affects the finite-time bias–variance tradeoff in the CIFAR-10 frozen-feature experiment.

B.1 EXPERIMENTAL DETAILS

Risk metrics. For the synthetic and real-covariate linear-regression experiments, we report the population excess risk $\frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{H}}^2$. In the CIFAR-10 experiments, we freeze an ImageNet-pretrained ResNet-18 and optimize only a linear least-squares head. We report the empirical quadratic excess risk relative to the exact least-squares solution, $\frac{1}{2n}\|\mathbf{X}(\mathbf{W} - \mathbf{W}^*)\|_{\mathrm{F}}^2$, together with test classification accuracy where appropriate. The CIFAR-10 experiments use $10{,}000$ training images and $2{,}000$ test images. The extracted 512-dimensional features are centered, normalized using the training set, and augmented with an intercept coordinate.

Statistical reporting. All stochastic simulations use independent sampling trajectories. The directional-filter experiment uses 4000 Monte Carlo trajectories; the power-law experiment uses

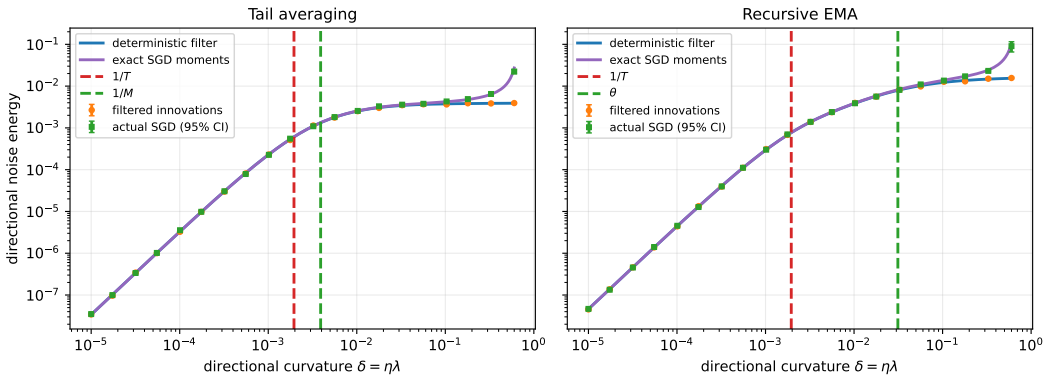

Figure 3: **Directional filter cutoffs.** Directional noise energy as a function of the normalized curvature $\delta = \eta\lambda$ for tail averaging (left) and recursive averaging (right). Solid blue curves show the deterministic-filter calculation, and purple curves show the exact second moments of the random-coefficient SGD recursion. Orange circles and green squares are the corresponding Monte Carlo estimates, with 95% confidence intervals. The vertical lines mark the predicted cutoffs: $\delta = 1/T$ and $\delta = 1/M$ for tail averaging, and $\delta = 1/T$ and $\delta = \theta$ for recursive averaging.

128 trajectories; and the synthetic SGD/SHB and real-covariate experiments use 256 trajectories. Their error bars show twice the standard error, providing approximate 95% Monte Carlo confidence intervals. The CIFAR-10 curves report the mean over five independent sampling trajectories, with shaded regions indicating one standard deviation. The frozen ResNet-18 features and the least-squares reference solution are held fixed across these trajectories, so the reported variability arises from stochastic mini-batch sampling.

Implementation. All experiments use constant stepsizes and sampling with replacement. Exact quadratic excess risks are evaluated from the corresponding population or empirical covariance matrices rather than estimated from additional test samples. For the CIFAR-10 mini-batch ablation, different batch sizes are compared at equal numbers of linear-head updates in order to isolate the effect of gradient-noise reduction; the associated numbers of processed examples are also recorded. Numerical results and plotting data are saved alongside every figure to facilitate reproduction.

B.2 Validation of Directional Filter Cutoffs

We use a one-dimensional Gaussian-design model to examine how averaging acts on eigendirections of different curvature. Let $\delta = \eta\lambda$ denote the dimensionless directional curvature. We set $\eta = 1$, draw $\mathbf{x}_t \sim \mathcal{N}(0, \delta)$ and $\varepsilon_t \sim \mathcal{N}(0, \delta)$, and simulate the random-coefficient recursion

$$\mathbf{e}_{t+1} = (1 - \mathbf{x}_t^2)\mathbf{e}_t + \varepsilon_t \mathbf{x}_t, \quad \mathbf{e}_0 = 0.$$

We compare tail averaging over the final M iterates with the recursive average

$$\widetilde{\mathbf{e}}_t = (1 - \theta)\widetilde{\mathbf{e}}_{t-1} + \theta \mathbf{e}_t.$$

The experiment uses $T = 512$, $M = 256$, $\theta = 1/32$, and 4000 Monte Carlo trials.

Figure 3 compares the deterministic-filter prediction with the exact second-moment recursion for the stochastic SGD dynamics. The corresponding Monte Carlo estimates closely match their analytical curves. For tail averaging, the changes in directional behavior occur around $\delta = 1/T$ and $\delta = 1/M$: directions below $1/T$ are barely learned within the available horizon, whereas directions above $1/M$ are effectively smoothed by the averaging window. Recursive averaging exhibits the analogous cutoffs $\delta = 1/T$ and $\delta = \theta$, since its effective averaging horizon is of order $1/\theta$. These results numerically validate the directional cutoff structure used in the spectral analysis. The gap between the deterministic-filter and exact-SGD curves at larger curvature reflects the multiplicative randomness of the Gaussian design, which is retained by the exact moment recursion.

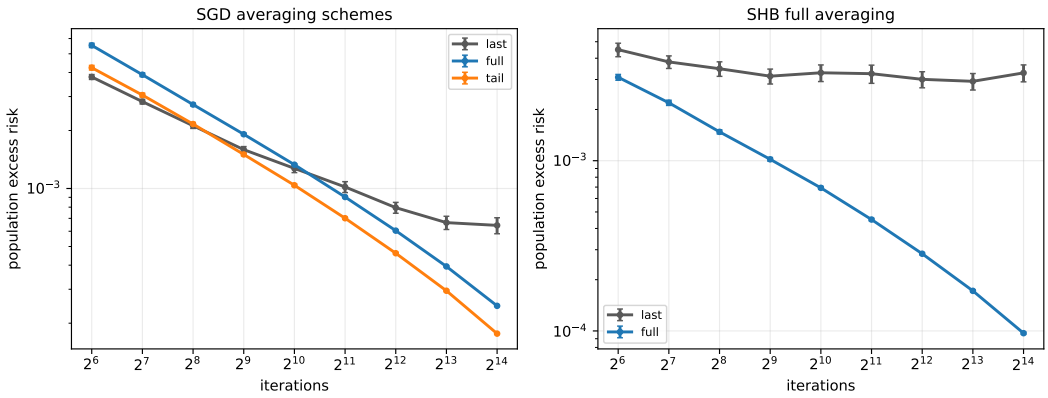

Figure 4: **Averaging dynamics of SGD and SHB.** Population excess risk versus the iteration budget in a synthetic linear-regression problem. Left: last iterate, full average, and tail average of SGD, where the tail length is $M = T/2$. Right: last iterate and full average of SHB. The last iterates approach noise-dominated plateaus, while the averaged estimators continue to reduce the excess risk. Error bars show twice the standard error over 256 independent Monte Carlo trajectories.

B.3 Averaging Dynamics of SGD and SHB

We next examine how different readouts affect the finite-time dynamics of SGD and SHB. We consider a $d = 64$ dimensional linear-regression problem with covariance eigenvalues $\lambda_i = i^{-2}$. Both algorithms use the constant stepsize $\eta = 0.03$, while SHB uses momentum parameter $\mu = 0.8$. The label-noise standard deviation is 0.2. For an iteration budget T, full averaging uses all T iterates, whereas tail averaging uses the final $M = T/2$ iterates. We simulate 256 independent trajectories and report their mean population excess risk; the error bars represent twice the standard error.

Figure 4 shows that the last iterates of both SGD and SHB eventually approach noise-dominated plateaus. For SGD, full and tail averaging continue to reduce the excess risk beyond this plateau, with tail averaging performing better at the displayed horizons because it discards the early transient iterates. The effect is more pronounced for SHB: its last iterate maintains a persistent variance floor, whereas the full average continues to decrease as the trajectory length grows. These observations are consistent with the variance-reduction mechanism captured by our averaging analysis. The two panels are intended as within-algorithm comparisons; their absolute risks should not be interpreted as a direct comparison between SGD and SHB.

B.4 Real Covariates with Directional Noise

We examine whether the averaging behavior persists beyond synthetically prescribed covariance spectra. We standardize the 442 covariate vectors from the diabetes dataset and map them to $d = 128$ fixed random ReLU features. After centering and coordinate-wise normalization, the feature vectors are rescaled so that $\max_j |x_j|^2 = 1$. Their empirical covariance $\widehat{\mathbf{H}} = \frac{1}{442}\sum_{j=1}^{442} x_j x_j^\top$ therefore provides a non-ideal spectrum induced by real covariates.

We consider two heteroscedastic label-noise models. In the bounded model, every sample has noise variance σ^2. In the second model, the noise variance is correlated with the leverage of a sample in the bottom eigenspace of $\widehat{\mathbf{H}}$, thereby amplifying noise in low-curvature directions. The two models are normalized to have the same average label-noise variance, $\frac{1}{442}\sum_{j=1}^{442}\sigma_j^2 = \sigma^2$, $\sigma = 0.15$. For the induced gradient-noise covariance $\mathbf{\Sigma} = \frac{1}{442}\sum_{j=1}^{442}\sigma_j^2 x_j x_j^\top$, we measure the relative directional noise by $r_i = \frac{\mathbf{v}_i^\top \mathbf{\Sigma}\mathbf{v}_i}{\widehat{\lambda}_i}$, where $(\widehat{\lambda}_i, \mathbf{v}_i)$ is an eigenpair of $\widehat{\mathbf{H}}$.

To isolate the variance contribution, we initialize the parameter error at zero, i.e., $\mathbf{w}_0 = \mathbf{w}^*$. Consequently, the last-iterate and fixed-window EMA risks initially increase as stochastic fluctuations accumulate and then approach stationary noise floors. This is expected for constant-stepsize SGD and should not be interpreted as optimization divergence. In contrast, the averaging windows of

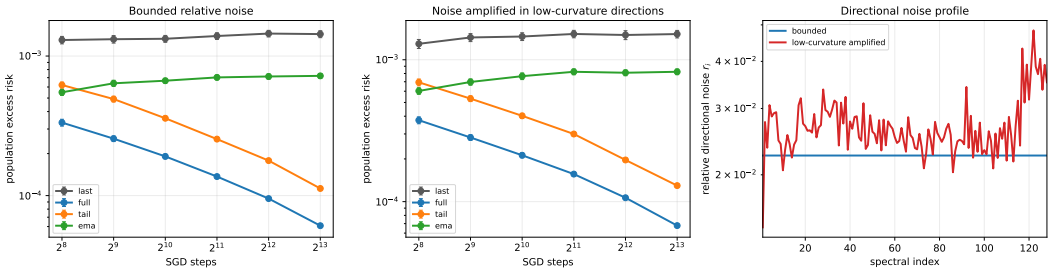

Figure 5: **SGD averaging under real covariates and directional noise.** Population excess risk under bounded relative noise (left) and noise amplified in low-curvature directions (middle). The full average uses all T iterates, the tail average uses the final $M = T/2$ iterates, and EMA uses $\theta = 1/64$. Right: relative directional noise $r_i = (v_i^\top \Sigma v_i)/\widehat{\lambda}_i$. Both noise constructions have the same average label-noise variance. Curves and error bars report the mean and twice the standard error over 256 independent Monte Carlo trajectories.

the full and tail estimators grow with the iteration budget T, allowing their variance to continue decreasing. We use stepsize $\eta = 0.8$, EMA parameter $\theta = 1/64$, and tail length $M = T/2$.

Figure 5 shows that the qualitative averaging behavior persists under the real-data covariance spectrum. Full and tail averaging continue to reduce the variance, whereas the last iterate and fixed-parameter EMA approach non-vanishing noise floors. The low-curvature construction yields moderately larger risks while preserving the same overall noise scale. The right panel verifies the intended directional structure: the bounded model has an essentially constant r_i, whereas the amplified model places more relative noise in the spectral tail. Thus, the predicted averaging behavior remains visible under a non-synthetic covariance spectrum and nonuniform directional noise.

B.5 Effect of Mini-Batch Size

We further investigate how the stochastic-gradient noise level affects the finite-time benefit of averaging on the frozen CIFAR-10 features. We consider mini-batch sizes $B \in \{1, 8, 32\}$ while keeping the stepsize $\eta = 0.03$, the SHB momentum parameter $\mu = 0.8$, and the number of linear-head updates fixed across batch sizes. This comparison is designed to isolate the effect of mini-batch noise reduction rather than computational or sample efficiency. For each configuration, we compare the last iterate with the full trajectory average over 128,000 updates.

Figure 6 illustrates the resulting bias–variance tradeoff. For SGD, the full average remains slightly worse than the last iterate over the displayed horizon because it retains the relatively large bias of early iterates. Increasing the batch size reduces the stochastic fluctuations of the last iterate and therefore further delays the point at which full averaging becomes advantageous. The effect is clearer for SHB. With $B = 1$, the last iterate reaches a pronounced noise-dominated plateau, whereas full averaging continues to reduce the excess risk. With $B = 8$, the last iterate reaches a lower and later plateau, and the full average becomes preferable only near the end of the trajectory. With $B = 32$, the last iterate has not yet reached a comparable variance floor and remains slightly better than the full average at the displayed horizon. Overall, larger mini-batches lower the last-iterate variance floor and delay the averaging crossover, consistently with the finite-time bias–variance decomposition.

C Related Work

Benign overfitting in linear regression. A growing literature studies why interpolating linear predictors can achieve small population risk in overparameterized regimes. Early work characterizes benign overfitting through the spectrum of the population covariance and the geometry of minimum-norm interpolation (Belkin et al., 2019; 2020; Bartlett et al., 2020; Hastie et al., 2022). Related results investigate harmless interpolation and the implicit regularization induced by optimization algorithms (Muthukumar et al., 2020; Liang & Rakhlin, 2020; Koren et al., 2022). These works

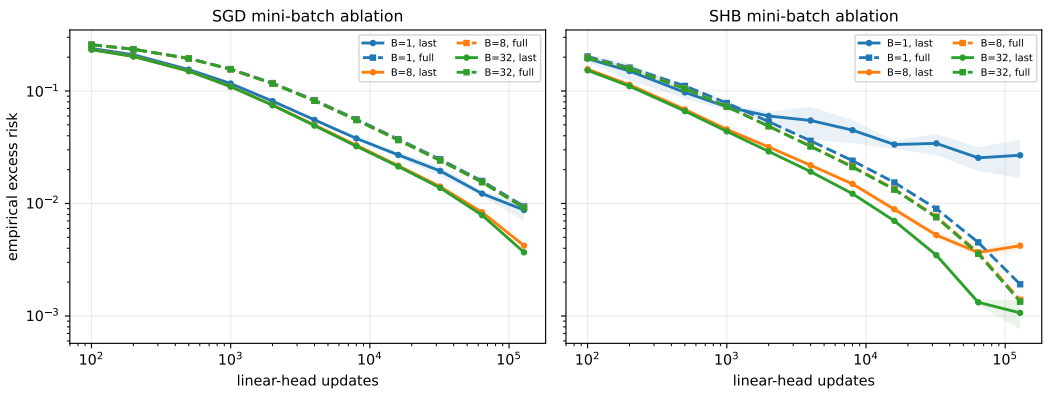

Figure 6: **Effect of mini-batch size on frozen CIFAR-10 features.** Empirical excess risk of the last iterate and full average for SGD (left) and SHB (right), using $B \in \{1, 8, 32\}$. All configurations use the same number of linear-head updates, stepsize $\eta = 0.03$, and, for SHB, momentum parameter $\mu = 0.8$. Increasing the batch size lowers the last-iterate variance floor and delays the point at which full averaging becomes beneficial. Curves and shaded regions report the mean and one standard deviation over five independent sampling trajectories.

establish that parameter dimension alone is not the appropriate complexity measure; rather, generalization depends on how the signal and noise interact with the covariance spectrum. Our analysis follows this spectral perspective but focuses on the finite-time risk of stochastic iterative algorithms and their output rules.

Constant-stepsize SGD and iterate averaging. Polyak–Ruppert averaging is a classical mechanism for improving the statistical efficiency of stochastic approximation (Polyak & Juditsky, 1992). For least-squares regression, non-asymptotic analyses of constant-stepsize SGD reveal detailed bias–variance tradeoffs determined by the spectrum of the covariance operator (Bach & Moulines, 2013; Défossez & Bach, 2015; Dieuleveut et al., 2020). Most directly related to our work, Zou et al. (2021) establish sharp upper and lower risk bounds for averaged SGD in overparameterized linear regression, including the effective head–tail spectral structure. Subsequent work studies last-iterate guarantees, alternative stepsize schedules, and simplified derivations of related bounds (Wu et al., 2022; Meterez et al., 2025). Our analysis builds on these results by expressing the effect of an averaging rule through its directional readout filter. This lets us track the label-noise covariance through the filter and treat the stochastic covariance feedback separately. The distinction matters when the noise cannot be bounded by a fixed multiple of the feature covariance, as the results below illustrate.

Averaging rules as spectral readouts. Different output rules for stochastic optimization include full averaging, tail averaging, last-iterate output, and exponential moving averaging. Tail averaging and related finite-horizon schemes have been studied through direct expansions of the SGD recursion (Zou et al., 2021; Wu et al., 2022; Meterez et al., 2025), while recent work analyzes EMA in overparameterized linear regression (Li & Gu, 2025). Although their final guarantees often exhibit spectral cutoffs, the analyses are typically specialized to the particular output rule.

We instead separate the underlying stochastic dynamics from the deterministic readout applied to the iterate trajectory. For a fixed algorithm, each readout induces its own impulse-response kernel and hence its own directional energy profile. This representation treats full averaging, tail averaging, and EMA as different readouts of the same stochastic system. The resulting cutoffs agree with the effective-dimension structure in earlier work; our contribution is the common finite-horizon representation and the preservation of the directional noise covariance, rather than the existence of spectral cutoffs by itself.

Momentum and accelerated stochastic methods. Momentum methods, including heavy-ball and accelerated SGD, generate second-order recursions whose statistical behavior differs from that of first-order SGD (Yan et al., 2018; Liu et al., 2020; Jelassi & Li, 2022). Existing analyses derive convergence and risk guarantees through Lyapunov arguments, augmented state-space recursions, or direct control of the roots of the characteristic polynomial. Recent work on accelerated SGD in

overparameterized linear regression also establishes bias–variance bounds for averaged momentum iterates (Li et al., 2024). These results show that acceleration and generalization depend delicately on the stepsize, momentum parameter, and spectral regime.

Generating functions and dynamical-systems viewpoints. Generating functions, Z-transforms, and transfer functions are classical tools for linear recurrences and discrete-time dynamical systems (Chen, 1984; Ogata, 1995; Oppenheim, 1999; Box et al., 2015). Dynamical-systems and control-theoretic viewpoints have also been used to study optimization algorithms, including spectral mode analysis and integral quadratic constraints (Lessard et al., 2016; Michel et al., 2008; Dieuleveut et al., 2020). In our work, generating functions are a technical device for deriving finite-horizon impulse responses and convolution identities. The main conceptual object is the spectral readout kernel, which separates the algorithmic recurrence from the output rule and interfaces with the stochastic covariance recursion.

D Exact Identities for SGD

D.1 Covariance recursion

Let $\mathbf{P}_t := \mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t$. Conditional on the past, $\mathbf{v}_t$ is deterministic and the current sample is independent of it. Expanding the noise-process gives

$$\mathbb{E}[\mathbf{P}_t \mathbf{C}_t \mathbf{P}_t] = \mathbf{C}_t - \eta(\mathbf{H}\mathbf{C}_t + \mathbf{C}_t\mathbf{H}) + \eta^2 \mathcal{M}(\mathbf{C}_t) = \mathbf{A}\mathbf{C}_t\mathbf{A} + \eta^2 \mathcal{N}(\mathbf{C}_t).$$

The two cross terms between $\mathbf{P}_t\mathbf{v}_t$ and $\eta\varepsilon_t\mathbf{x}_t$ vanish after unconditional expectation. Indeed, the current sample is independent of $\mathbf{v}_t$, the cross term is linear in $\mathbf{v}_t$, and $\mathbb{E}\mathbf{v}_t = 0$ by induction from Subsection 2.1. The remaining additive term is $\eta^2\mathbf{\Sigma}$, proving Eq. (3.1). For the innovation $\boldsymbol{\xi}$,

$$\mathbb{E}[\boldsymbol{\xi}_{t+1} \mid \mathcal{F}_t] = 0, \quad \mathbb{E}[\boldsymbol{\xi}_{t+1} \otimes \boldsymbol{\xi}_{t+1}] = \mathcal{N}(\mathbf{C}_t) + \mathbf{\Sigma}. \tag{D.1}$$

The martingale-difference property implies orthogonality across different times.

D.2 Proof of Proposition 3.1

Iterating $\mathbf{v}_{t+1} = \mathbf{A}\mathbf{v}_t + \eta\boldsymbol{\xi}_{t+1}$ yields

$$\mathbf{v}_t = \eta \sum_{r=0}^{t-1} \mathbf{A}^{t-1-r} \boldsymbol{\xi}_{r+1}.$$

Substitution into $\widehat{\mathbf{e}}_\alpha := \sum_{t=0}^{T-1} \alpha_t \mathbf{e}_t$, followed by an exchange of the two finite sums, gives

$$\widehat{\mathbf{v}}_\alpha = \eta \sum_{r=0}^{T-2} K_r^\alpha(\mathbf{A}) \boldsymbol{\xi}_{r+1}.$$

All cross-time terms disappear after taking the squared $\mathbf{H}$-norm and expectation. Using Eq. (D.1) proves Eq. (3.3). Since $K_r^\alpha(\mathbf{A})$ is a function of $\mathbf{H}$, it is diagonal in the basis $(u_i)_i$. Therefore we get

$$\operatorname{tr}\big(\mathbf{H}[K_r^\alpha(\mathbf{A})]^2\mathbf{Q}\big) = \sum_i \lambda_i [K_{i,r}^\alpha]^2 Q_{ii}.$$

Notice that this argument never diagonalizes $\mathbf{Q}$.

E Positive Renewal Closure

Unrolling Eq. (3.1) gives

$$\mathbf{C}_t = \eta^2 \sum_{s=0}^{t-1} \mathbf{A}^{t-1-s} \big(\mathbf{\Sigma} + \mathcal{N}(\mathbf{C}_s)\big) \mathbf{A}^{t-1-s}. \tag{E.1}$$

Define

$$d_t := \eta^2 \sum_i \lambda_i \Sigma_{ii} \sum_{j=0}^{t-1} (1-\eta\lambda_i)^{2j}, \; h_j := \eta^2 \sum_i \lambda_i^2 (1-\eta\lambda_i)^{2j}. \tag{E.2}$$

Taking $\mathrm{tr}(\mathbf{H}\cdot)$ in Eq. (E.1) and applying Assumption 2.1 yields

$$d_t + \beta \sum_{s=0}^{t-1} h_{t-1-s} r_s \le r_t \le d_t + \psi \sum_{s=0}^{t-1} h_{t-1-s} r_s. \tag{E.3}$$

The total mass of the upper kernel is

$$\psi \sum_{j\ge 0} h_j = \psi \sum_i \frac{\eta\lambda_i}{2-\eta\lambda_i} = m.$$

Moreover, $d_t \le \eta\,\mathrm{tr}(\mathbf{\Sigma})$. Induction in the upper inequality of Eq. (E.3) proves

$$r_t \le \frac{\eta\,\mathrm{tr}(\mathbf{\Sigma})}{1-m}. \tag{E.4}$$

To obtain a lower bound after warmup, write $a_i := 1-\eta\lambda_i$. From Eq. (E.2),

$$d_t = \eta \sum_i \frac{\Sigma_{ii}(1-a_i^{2t})}{2-\eta\lambda_i}.$$

The definition Eq. (3.4) implies, for every $t \ge k_0$,

$$d_t \ge \frac{\eta}{2} \sum_i \frac{\Sigma_{ii}}{2-\eta\lambda_i} \ge \frac{\eta}{4}\,\mathrm{tr}(\mathbf{\Sigma}). \tag{E.5}$$

Indeed, the discarded geometric tail is at most one half of the weighted total by Eq. (3.4), and $1 \le 2-\eta\lambda_i \le 2$.

E.1 Proof of Theorem 3.3

For the tail readout define

$$g_r := \eta^2\,\mathrm{tr}\big(\mathbf{H}^2 K_r(\mathbf{A})^2\big), \quad G := \sum_{r=0}^{T-2} g_r = 2\mathcal{F}_{S,M}(\mathbf{H}).$$

Eq. (3.3), the upper moment comparison, and Eq. (E.4) give

$$V_{S,M} \le \mathcal{F}_{S,M}(\mathbf{\Sigma}) + \frac{\psi}{2} \sum_{r=0}^{T-2} r_r g_r \le \mathcal{F}_{S,M}(\mathbf{\Sigma}) + \frac{\psi\eta\,\mathrm{tr}(\mathbf{\Sigma})}{1-m} \mathcal{F}_{S,M}(\mathbf{H}).$$

For $r < S$, the scalar tail kernel equals

$$K_r(a) = a^{S-1-r}\,\frac{1-a^M}{M(1-a)}.$$

Hence g_r is nondecreasing on $0 \le r < S$. If $S \ge 2k_0$, shifting the first k_0 indices by k_0 stays inside the pre-window range and gives

$$g_r \le g_{r+k_0}, \quad 0 \le r \le k_0 - 1.$$

Therefore

$$\sum_{r=0}^{k_0-1} g_r \le \sum_{r=k_0}^{2k_0-1} g_r \le \sum_{r=k_0}^{T-2} g_r.$$

Adding the last quantity to both sides and using $G = \sum_{r<k_0} g_r + \sum_{r\ge k_0} g_r$ gives

$$\sum_{r=k_0}^{T-2} g_r \ge \frac{1}{2} G.$$

Using only these indices, Eq. (E.5), and the lower moment comparison yields

$$V_{S,M} \ge \mathcal{F}_{S,M}(\mathbf{\Sigma}) + \frac{\beta}{2} \sum_{r=k_0}^{T-2} r_r g_r \ge \mathcal{F}_{S,M}(\mathbf{\Sigma}) + \frac{\beta\eta\,\mathrm{tr}(\Sigma)}{8} \mathcal{F}_{S,M}(\mathbf{H}).$$

Linearity of $\mathcal{F}_{S,M}$ completes the proof.

F TAIL-FILTER ESTIMATES

F.1 EXACT SCALAR KERNEL

For $a = 1 - \delta$, innovations before the averaging window and innovations inside it have kernels

$$K_r(a) = \begin{cases} a^{S-1-r}\dfrac{1-a^M}{M(1-a)}, & 0 \le r < S, \\ \dfrac{1-a^{T-1-r}}{M(1-a)}, & S \le r \le T-2. \end{cases}$$

Squaring and summing the two pieces proves Eq. (4.1).

F.2 PROOF OF LEMMA 4.1

We repeatedly use

$$1-(1-\delta)^n \asymp \min\{1, n\delta\}, \quad 1-(1-\delta)^{2n} \asymp \min\{1, n\delta\}. \tag{F.1}$$

Write Eq. (4.1) as $P_{S,M}(\delta) + W_M(\delta)$. Then

$$P_{S,M}(\delta) \asymp \frac{\min\{1, M\delta\}^2}{M^2\delta^2}\frac{\min\{1, S\delta\}}{\delta}, \tag{F.2}$$

and

$$W_M(\delta) \asymp M\min\{1, (M\delta)^{-2}\}. \tag{F.3}$$

We first justify these comparisons. Since $1 - a^2 = \delta(2-\delta) \asymp \delta$, the first term of Eq. (4.1) gives Eq. (F.2). For the second term, put

$$Z_M(\delta) := \sum_{\ell=1}^{M-1} \left(1-(1-\delta)^\ell\right)^2.$$

By Eq. (F.1),

$$Z_M(\delta) \lesssim \sum_{\ell=1}^{M-1} \min\{1, \ell^2\delta^2\} \lesssim M\min\{1, M^2\delta^2\}.$$

For the reverse inequality, retain $\ell = \lceil M/2\rceil, \dots, M-1$. There are at least $M/3$ such indices for $M \ge 2$, and every retained term is bounded below by a constant multiple of $\min\{1, M^2\delta^2\}$. Hence $Z_M(\delta) \asymp M\min\{1, M^2\delta^2\}$. Dividing by $M^2\delta^2$ proves Eq. (F.3). The case $M = 1$ has $W_M = 0$ and is read directly from $P_{S,1}$. We now prove each regime. If $\delta T \le 1$, then $\delta S, \delta M \le 1$ and

$$P_{S,M}(\delta) \asymp S, \quad W_M(\delta) \asymp M,$$

so their sum is of order T. Suppose $\delta T > 1$ and $\delta M \le 1$. If $S \ge M$, then $S \ge T/2$ and $\delta S > 1/2$, whence

$$P_{S,M}(\delta) \asymp \delta^{-1}, \quad W_M(\delta) \lesssim \delta^{-1}.$$

If $S < M$, then $M < T < 2M$, so $1/2 < \delta M \le 1$. Consequently $W_M(\delta) \asymp M \asymp \delta^{-1}$ and $P_{S,M}(\delta) \lesssim \delta^{-1}$. Finally, if $\delta M > 1$,

$$P_{S,M}(\delta) \lesssim \frac{1}{M^2\delta^3} \le \frac{1}{M\delta^2} \asymp W_M(\delta).$$

These are precisely the three cases in Eq. (4.2).

F.3 PROOF OF THEOREMS 4.2 AND 4.4

By the diagonal filter,

$$\mathcal{F}_{S,M}(\mathbf{Q}) = \frac{\eta^2}{2}\sum_i \lambda_i Q_{ii}\Gamma_{S,M}(\eta\lambda_i).$$

In the three regimes of Lemma 4.1, the summand is respectively

$$T\eta^2\lambda_i Q_{ii}, \quad \eta Q_{ii}, \quad \frac{Q_{ii}}{M\lambda_i}.$$

Splitting the spectrum at Eq. (4.3) proves Theorem 4.2. Substituting this equivalence into Theorem 3.3 proves Theorem 4.4.

G Power-Law Phase Diagram

Lemma G.1 (Warmup scale in the power-law model)**.** *Let $\lambda_i = i^{-p}$ and $\Sigma_{ii} = i^{-q}$, where $p > 1$, and let $\eta \in (0, 1)$ be fixed. Define k_0 as the smallest nonnegative integer satisfying*

$$\sum_{i\geq 1} \frac{i^{-q}}{2 - \eta i^{-p}} (1 - \eta i^{-p})^{2k_0} \leq \frac{1}{2} \sum_{i\geq 1} \frac{i^{-q}}{2 - \eta i^{-p}}.$$

Then, as $q \downarrow 1$,

$$\log(\eta k_0) = \frac{p \log 2}{q - 1} + o\left(\frac{1}{q-1}\right).$$

In particular,

$$k_0 = \eta^{-1} \exp\left(\Theta\left(\frac{p}{q-1}\right)\right).$$

Proof. Write $\epsilon := q - 1$ and

$$S_\epsilon(k) := \sum_{i\geq 1} \frac{i^{-1-\epsilon}}{2 - \eta i^{-p}} (1 - \eta i^{-p})^{2k}.$$

Since $(2 - \eta i^{-p})^{-1} \to 1/2$ as $i \to \infty$,

$$S_\epsilon(0) = \frac{1 + o(1)}{2\epsilon}, \quad \epsilon \downarrow 0.$$

Consider $k = k_\epsilon$ such that $\epsilon \log(\eta k_\epsilon) = O(1)$. The indices contributing to $S_\epsilon(k_\epsilon)$ are of order $(\eta k_\epsilon)^{1/p}$ and hence tend to infinity. On this range, $(2 - \eta i^{-p})^{-1} = 1/2 + o(1)$ and

$$(1 - \eta i^{-p})^{2k_\epsilon} = \exp(-2\eta k_\epsilon i^{-p})(1 + o(1)).$$

A sum–integral comparison therefore gives

$$S_\epsilon(k_\epsilon) = \frac{1 + o(1)}{2} \int_0^\infty x^{-1-\epsilon} e^{-2\eta k_\epsilon x^{-p}}\, dx.$$

With $z = 2\eta k_\epsilon x^{-p}$, the integral equals

$$\frac{1}{p} (2\eta k_\epsilon)^{-\epsilon/p} \Gamma\left(\frac{\epsilon}{p}\right).$$

Hence

$$\frac{S_\epsilon(k_\epsilon)}{S_\epsilon(0)} = \frac{\epsilon}{p} \Gamma\left(\frac{\epsilon}{p}\right) (2\eta k_\epsilon)^{-\epsilon/p} (1 + o(1)).$$

Because $z\Gamma(z) \to 1$ as $z \downarrow 0$, if $\epsilon \log(\eta k_\epsilon)/p \to a$, then

$$\frac{S_\epsilon(k_\epsilon)}{S_\epsilon(0)} \longrightarrow e^{-a}.$$

The minimal warmup index is determined by the level $S_\epsilon(k_0)/S_\epsilon(0) = 1/2$ up to the integer rounding. Therefore $a = \log 2$, which proves $\frac{\epsilon}{p} \log(\eta k_0) \to \log 2$. □

The balanced-tail cutoff solves $\eta M \lambda_{k_M} \asymp 1$ and therefore $k_M \asymp (\eta M)^{1/p} \asymp M^{1/p}$ for fixed η. From Theorem 4.2,

$$\mathcal{G}_{2M,M}(\mathbf{\Sigma}) \asymp \frac{1}{M} \sum_{i \leq k_M} i^{p-q} + \eta^2 M \sum_{i > k_M} i^{-(p+q)}.$$

For $a > -1$, integral comparison gives $\sum_{i\leq k} i^a \asymp k^{a+1}$; for $a = -1$ it gives $\log k$; and for $a < -1$ the partial sum is bounded above and below by positive constants. Moreover, for $b > 1$,

$$\frac{(k+1)^{1-b}}{b-1} \leq \sum_{i>k} i^{-b} \leq \frac{k^{1-b}}{b-1}.$$

Apply these estimates with $a = p - q$ and $b = p + q$, and substitute $k_M \asymp M^{1/p}$. This gives

$$\mathcal{G}_{2M,M}(\mathbf{\Sigma}) \asymp \begin{cases} M^{-(q-1)/p}, & 1 < q < p+1, \\ (\log M)/M, & q = p+1, \\ M^{-1}, & q > p+1, \end{cases}$$

whereas $\mathcal{G}_{2M,M}(\mathbf{H}) \asymp M^{-(1-1/p)}$. The slower of these two contributions proves Eq. (4.6). For $\mathbf{e}_0 = \mathbf{u}_1$, the deterministic initial energy decays exponentially, but random Hessian fluctuations generate the polynomial kernel

$$h_t = \eta^2 \sum_i i^{-2p}(1 - \eta i^{-p})^{2t} \asymp t^{-(2-1/p)}. \tag{G.1}$$

We spell out the renewal step. Let nonnegative sequences satisfy

$$q_n = u_n + \sum_{k=1}^{n} f_k q_{n-k}, \quad \sum_{k \geq 1} f_k = m_f < 1,$$

and suppose $f_k \leq C_f(k+1)^{-r}$ and $u_k \leq C_u(k+1)^{-r}$ for some $r > 1$, with $\sum_k u_k < \infty$. Then $q_n \lesssim (n+1)^{-r}$. Indeed, iteration gives the coordinatewise identity

$$q = \sum_{j \geq 0} f^{*j} * u,$$

and Tonelli's theorem yields $\sum_n q_n = (\sum_n u_n)/(1 - m_f) < \infty$. Choose $\epsilon > 0$ and $K < \infty$ such that

$$\vartheta := (1+\epsilon)m_f + 2^r \sum_{k>K} f_k < 1.$$

For $n \geq 2K$, split the convolution into $k \leq K$, $K < k \leq n/2$, and $k > n/2$. If inductively $q_j \leq C(j+1)^{-r}$ for $j < n$, the first two parts are bounded by $\vartheta C(n+1)^{-r}$. In the last part, use $f_k \leq 2^r C_f (n+1)^{-r}$ and $\sum_j q_j < \infty$. Taking C sufficiently large closes the induction.

In the initialization recursion, $f_k = \psi h_{k-1}$, $m_f = m < 1$, and $u_t = (1-\eta)^{2t}$. To justify Eq. (G.1), use $(1-u)^{2t} \leq e^{-2tu}$ for the upper bound and $(1-u)^{2t} \geq e^{-4tu}$ when $u \leq 1/2$ for the lower bound. The summand is unimodal, so its sum is comparable to its integral up to its maximal term. With $z = 2\eta t x^{-p}$,

$$\int_0^\infty x^{-2p} e^{-2\eta t x^{-p}}\, dx = \frac{1}{p}(2\eta t)^{1/p-2}\Gamma\left(2 - \frac{1}{p}\right).$$

The maximal summand is $O(t^{-2})$, which is lower order than the displayed integral. Replacing 2η by 4η gives the same power for the lower integral over the range $\eta i^{-p} \leq 1/2$. Hence Eq. (G.1) holds on both sides with $r = 2 - 1/p$. The renewal lemma proves

$$\rho_t := \mathbb{E}\|b_t\|_{\mathbf{H}}^2 \lesssim (t+1)^{-(2-1/p)}.$$

Convexity of the squared norm then gives, for $S = M$,

$$B_{M,M} \leq \frac{1}{2M} \sum_{t=M}^{2M-1} \rho_t \lesssim M^{-(2-1/p)}. \tag{G.2}$$

For the lower bound, let $\mathbf{D}_r := \mathbb{E}[\mathbf{b}_r \otimes \mathbf{b}_r]$. The initialization innovation has conditional covariance $\mathcal{N}(\mathbf{D}_r)$, so the centered output variance is

$$\frac{\eta^2}{2} \sum_{r=0}^{2M-2} \operatorname{tr}\big(\mathbf{H} K_r(\mathbf{A})^2 \mathcal{N}(\mathbf{D}_r)\big).$$

Retain only $r = 0$. Since $\mathbf{D}_0 = \mathbf{u}_1 \otimes \mathbf{u}_1$ and $\operatorname{tr}(\mathbf{H}\mathbf{D}_0) = 1$, Assumption 2.1 gives

$$B_{M,M} \geq \frac{\beta\eta^2}{2} \sum_i \lambda_i^2 K_0(1 - \eta\lambda_i)^2.$$

For $\lambda_i \le (4\eta M)^{-1}$ and $t \in [M, 2M-1]$, $(1-\eta\lambda_i)^{t-1} \ge (1-1/(4M))^{2M} \ge c$. Hence $K_0(1-\eta\lambda_i) \ge c$, and therefore

$$B_{M,M} \gtrsim \sum_{\lambda_i \le (4\eta M)^{-1}} \lambda_i^2 \asymp M^{-(2-1/p)}.$$

Combining the two bounds gives

$$B_{M,M} = \Theta(M^{-(2-1/p)}).$$

Since $2 - 1/p > \min\{(q-1)/p, 1-1/p\}$, the total-risk conclusion follows from Eq. (2.2).

G.1 A REALIZABLE JOINT DISTRIBUTION

The power-law pair $(\mathbf{H}, \mathbf{\Sigma})$ can be realized by an actual joint distribution satisfying the moment assumptions with dimension-free constants. Let

$$\kappa := \frac{1}{4\sum_i \lambda_i^2}, \quad \pi_i := \kappa\lambda_i^2, \quad \mathbf{G} := \frac{4}{3}(\mathbf{H} - \kappa\mathbf{H}^2).$$

Then $\sum_i \pi_i = 1/4$, $0 < \kappa \le 1/4$, and $\mathbf{H} \preceq \mathbf{G} \preceq (4/3)\mathbf{H}$. Given any summable sequence $s_i \ge 0$, define the following mixture: with probability $3/4$, draw $\mathbf{x} \sim \mathcal{N}(0, \mathbf{G})$ and set $\varepsilon = 0$; with probability π_i, draw x uniformly from $\{e_i, -e_i\}$ and independently draw ε uniformly from $\{\sqrt{s_i/\pi_i}, -\sqrt{s_i/\pi_i}\}$. The Gaussian component contributes $(3/4)\mathbf{G} = \mathbf{H} - \kappa\mathbf{H}^2$ to the input covariance, while the atoms contribute $\sum_i \pi_i \mathbf{e}_i \otimes \mathbf{e}_i = \kappa\mathbf{H}^2$. Hence $\mathbb{E}[\mathbf{x}\otimes\mathbf{x}] = \mathbf{H}$. The independent signs give $\mathbb{E}[\varepsilon \mid \mathbf{x}] = 0$, and the ith atom contributes $s_i \mathbf{e}_i \otimes \mathbf{e}_i$ to the noise covariance. Thus

$$\mathbb{E}[\varepsilon\mathbf{x}] = 0, \quad \mathbf{\Sigma} = \mathrm{diag}(s_i), \quad \mathbb{E}\varepsilon^2 = \sum_i s_i < \infty.$$

It remains to verify Assumption 2.1. For $\mathbf{C} \succeq 0$, the Gaussian fourth-moment formula gives

$$\mathcal{M}(\mathbf{C}) = \frac{3}{4}\{2\mathbf{GCG} + \mathrm{tr}(\mathbf{GC})\mathbf{G}\} + \sum_i \pi_i C_{ii} \mathbf{e}_i \otimes \mathbf{e}_i.$$

Since $\mathbf{GCG} \preceq \mathrm{tr}(\mathbf{GC})\mathbf{G}$ and $\mathbf{H} \preceq \mathbf{G} \preceq (4/3)\mathbf{H}$, the Gaussian term is bounded above by $4\,\mathrm{tr}(\mathbf{HC})\mathbf{H}$. For the atomic term,

$$\pi_i C_{ii} = \kappa\lambda_i^2 C_{ii} \le \kappa\lambda_i \,\mathrm{tr}(\mathbf{HC}),$$

so it is at most $\kappa\,\mathrm{tr}(\mathbf{HC})\mathbf{H}$. Therefore

$$\mathcal{M}(\mathbf{C}) \preceq \frac{17}{4}\,\mathrm{tr}(\mathbf{HC})\mathbf{H}.$$

Finally, apply the operator law of total variance to $\mathbf{Z} = \mathbf{x}\otimes\mathbf{x}$ and retain only the within- Gaussian-component variance:

$$\begin{aligned}
\mathcal{N}(\mathbf{C}) &= \mathbb{E}[(\mathbf{Z}-\mathbf{H})\mathbf{C}(\mathbf{Z}-\mathbf{H})] \\
&\succeq \frac{3}{4}\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{G})}[(\mathbf{x}\otimes\mathbf{x}-\mathbf{G})\mathbf{C}(\mathbf{x}\otimes\mathbf{x}-\mathbf{G})] \\
&= \frac{3}{4}\{\mathbf{GCG} + \mathrm{tr}(\mathbf{GC})\mathbf{G}\} \succeq \frac{3}{4}\,\mathrm{tr}(\mathbf{HC})\mathbf{H}.
\end{aligned}$$

Thus one may take $\psi = 17/4$ and $\beta = 3/4$, independently of (s_i) and the dimension. Choosing $s_i = i^{-q}$ realizes the setting.

H FULL AVERAGING AND EMA

For full averaging,

$$\delta^2 \sum_{r=0}^{T-2} K_r(1-\delta)^2 = \frac{1}{T^2}\sum_{q=1}^{T-1}\big(1-(1-\delta)^q\big)^2.$$

The upper bound in Subsection 4.3 follows from $1-(1-\delta)^q \le \min\{1, q\delta\}$. For the lower bound, retain $q \in [\lceil T/2 \rceil, T-1]$ and use Eq. (F.1). For EMA, expanding the recursive initialization gives

$$\mathbf{m}_T = \beta_{\mathrm{ema}}^T \mathbf{w}_0 + \theta \sum_{t=0}^{T-1} \beta_{\mathrm{ema}}^{T-1-t} \mathbf{w}_t.$$

Substituting the recursion $\mathbf{e}_{t+1} = a\mathbf{e}_t + \eta\boldsymbol{\xi}_{t+1}$ and evaluating the finite geometric convolution proves Eq. (4.7). For the energy upper bound, regard the kernel as the convolution of $p_q = \theta\beta_{\mathrm{ema}}^q$ and $q_q = \delta(1-\delta)^q$. Young's inequality gives

$$\|p * q\|_2^2 \lesssim \min\{\theta, \delta\},$$

and the finite-horizon bound gives $T\delta^2$. If $T\theta \ge 1$, the last $\lceil 1/\theta \rceil$ EMA weights can be compared with uniform averaging weights. Here are the details. First suppose $\theta \le 1/2$ and let $M_\theta := \lceil 1/\theta \rceil \le T$. For each of the last M_θ iterates,

$$\theta\beta_{\mathrm{ema}}^{T-1-t} \ge \theta\beta_{\mathrm{ema}}^{M_\theta - 1} \ge \frac{1}{4M_\theta},$$

where we used $\log(1-\theta) \ge -2(\log 2)\theta$. All scalar impulse responses are nonnegative, so the pointwise ordering of the readout weights transfers to the innovation kernels. Consequently,

$$E_{T,\theta}^{\mathrm{ema}}(\delta) \gtrsim \min\{M_\theta^{-1}, \delta, T\delta^2\} \gtrsim \min\{\theta, \delta, T\delta^2\}.$$

If $\theta > 1/2$, the most recent iterate has weight at least $1/2$ and the same conclusion follows by comparison with a length-one tail readout.

I Heavy-Ball Proofs

I.1 Impulse-response mass

Let $b_i := 1+\mu-\eta\lambda_i$ and define $S_0 := \sum_{t\ge 0} g_{i,t}^2$ and $S_1 := \sum_{t\ge 1} g_{i,t}g_{i,t-1}$. The impulse recursion implies

$$S_1 = \frac{b_i}{1+\mu} S_0. \tag{I.1}$$

Squaring the recursion and summing gives

$$S_0 - 1 = (b_i^2 + \mu^2)S_0 - 2\mu b_i S_1.$$

Substitution of Eq. (I.1) and $(1+\mu)^2 - b_i^2 = \eta\lambda_i[2(1+\mu) - \eta\lambda_i]$ proves the result.

I.2 Proof of Theorem 5.1

For a block family $X_t^v = (\mathbf{X}_{i,t}^v)_i$, define

$$X_{i,t}^v := \mathbb{E}\begin{bmatrix} v_{i,t} \\ v_{i,t-1} \end{bmatrix}\begin{bmatrix} v_{i,t} \\ v_{i,t-1} \end{bmatrix}^\top, \ \mathscr{R}(X_t^v) := \sum_i \lambda_i (\mathbf{X}_{i,t}^v)_{11}, \ \mathbf{A}_i := \begin{bmatrix} 1+\mu-\eta\lambda_i & -\mu \\ 1 & 0 \end{bmatrix}, \ \mathbf{E}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The fourth-moment comparison, with the diagonal mean term retained, gives

$$\mathbb{E}[(\xi_{i,t+1}^v)^2] \le \psi\lambda_i \mathscr{R}(X_t^v) - \lambda_i^2 (\mathbf{X}_{i,t}^v)_{11} + \Sigma_{ii}. \tag{I.2}$$

To see the negative term explicitly, condition on the past and write $\mathbf{z} = \mathbf{v}_t$. The multiplicative part of the innovation is $(\mathbf{H} - \mathbf{x}\otimes\mathbf{x})\mathbf{z}$. Its i-th conditional second moment equals

$$\mathbb{E}\left[\langle \mathbf{e}_i, (\mathbf{H} - \mathbf{x}\otimes\mathbf{x})\mathbf{z}\rangle^2 \mid \mathcal{F}_t\right] = [\mathcal{M}(\mathbf{z}\otimes\mathbf{z})]_{ii} - \lambda_i^2 z_i^2 \le \psi\lambda_i \|\mathbf{z}\|_H^2 - \lambda_i^2 z_i^2.$$

For the residual-noise process, the cross term with $\varepsilon\mathbf{x}$ vanishes after unconditional expectation because the current sample is independent of $\mathbf{v}_t$ and $\mathbb{E}\mathbf{v}_t = 0$. Adding $\Sigma_{ii} = \mathbb{E}[\varepsilon^2 x_i^2]$ proves Eq. (I.2). The quantity in braces is nonnegative because $\psi \ge 1$ and $\mathscr{R}(X_t^v) \ge \lambda_i (\mathbf{X}_{i,t}^v)_{11}$. Hence the block operator

$$(\mathscr{K} X_t^v)_i = \mathbf{A}_i \mathbf{X}_{i,t}^v \mathbf{A}_i^\top + \eta^2\big(\psi\lambda_i \mathscr{R}(X_t^v) - \lambda_i^2 (\mathbf{X}_{i,t}^v)_{11}\big)\mathbf{E}_{11}$$

is positive. Let

$$\mathbf{S}_i := \sum_{t\geq 0} \mathbf{A}_i^t \mathbf{E}_{11} (\mathbf{A}_i^\top)^t.$$

Then $\mathbf{S}_i = \mathbf{A}_i \mathbf{S}_i \mathbf{A}_i^\top + \mathbf{E}_{11}$ and $(\eta\lambda_i)^2 (\mathbf{S}_i)_{11} = \ell_i$. Define

$$C_i^\Sigma := \frac{\eta^2}{1+\ell_i} \left(\Sigma_{ii} + \frac{\psi q_*(\mathbf{\Sigma})}{\Delta_*} \lambda_i \right) \mathbf{S}_i.$$

We now verify the fixed point without suppressing the coefficient calculation. First,

$$\mathscr{R}(C^\Sigma) = \sum_i \frac{\ell_i}{1+\ell_i} \left(\frac{\Sigma_{ii}}{\lambda_i} + \frac{\psi q_*(\mathbf{\Sigma})}{\Delta_*} \right) = q_*(\mathbf{\Sigma}) + \frac{\psi h_* q_*(\mathbf{\Sigma})}{\Delta_*} = \frac{q_*(\mathbf{\Sigma})}{\Delta_*},$$

and direction by direction,

$$\mathscr{K} C^\Sigma + \eta^2 \Lambda^\Sigma = C^\Sigma, \qquad \Lambda_i^\Sigma := \Sigma_{ii} \mathbf{E}_{11}.$$

Indeed, write $c_i := \eta^2 (\Sigma_{ii} + \psi\lambda_i q_*(\mathbf{\Sigma})/\Delta_*)/(1+\ell_i)$, so $C_i^\Sigma = c_i \mathbf{S}_i$. Since $\mathbf{A}_i \mathbf{S}_i \mathbf{A}_i^\top = \mathbf{S}_i - \mathbf{E}_{11}$, the coefficient of $\mathbf{E}_{11}$ in $(\mathscr{K} C^\Sigma)_i + \eta^2 \Sigma_{ii} E_{11} - C_i^\Sigma$ is

$$-c_i + \eta^2 \left(\psi\lambda_i \frac{q_*(\mathbf{\Sigma})}{\Delta_*} - \lambda_i^2 c_i L_i + \Sigma_{ii} \right) = -c_i(1+\ell_i) + \eta^2 \left(\Sigma_{ii} + \psi\lambda_i \frac{q_*(\mathbf{\Sigma})}{\Delta_*} \right) = 0.$$

Positivity and zero initialization imply $X_t \preceq C^\Sigma$. Substitution into Eq. (I.2), followed by the exact filtered-innovation identity, proves Eq. (5.1).

If $\mathbf{\Sigma} \preceq \sigma_{\mathrm{rel}}^2 \mathbf{H}$, then $q_*(\mathbf{\Sigma}) \leq \sigma_{\mathrm{rel}}^2 h_*$. Therefore

$$\mathbf{\Sigma} + \frac{\psi q_*(\mathbf{\Sigma})}{\Delta_*} \mathbf{H} \preceq \sigma_{\mathrm{rel}}^2 \left(1 + \frac{\psi h_*}{\Delta_*} \right) \mathbf{H} = \frac{\sigma_{\mathrm{rel}}^2}{\Delta_*} \mathbf{H},$$

which proves the result.

I.3 Proof of the full-SHB feedback lower bound

The covariance recursion is the sum of the direct label-noise contribution and the covariance generated by the random Hessians. Since the latter is positive semidefinite,

$$r_t^{\mathrm{hb}} \geq d_t^{\mathrm{hb}} := \eta^2 \sum_i \lambda_i \Sigma_{ii} \sum_{s=0}^{t-1} g_{i,s}^2.$$

The sequence d_t^{hb} is nondecreasing and converges to $d_\infty^{\mathrm{hb}} = \eta^2 \sum_i \lambda_i \Sigma_{ii} L_i$. Using $L_i := \sum_{t\geq 0} g_{i,t}^2 = \frac{1+\mu}{\tau\eta\lambda_i[2(1+\mu)-\eta\lambda_i]}$ and $\nu = \eta/\tau$ gives

$$d_\infty^{\mathrm{hb}} = \nu(1+\mu) \sum_i \frac{\Sigma_{ii}}{2(1+\mu)-\eta\lambda_i} \geq \frac{\nu}{2} \operatorname{tr}(\mathbf{\Sigma}).$$

By the definition of k_0^{hb}, $d_t^{\mathrm{hb}} \geq d_\infty^{\mathrm{hb}}/2$ for every $t \geq k_0^{\mathrm{hb}}$. Hence $r_t^{\mathrm{hb}} \geq \frac{\nu}{4} \operatorname{tr}(\mathbf{\Sigma})$, $t \geq k_0^{\mathrm{hb}}$. We next show that full averaging retains a constant fraction of its filter energy after this warmup. Write $\kappa_r^{\mathrm{hb}} := \eta^2 \sum_i \lambda_i^2 \big(K_{i,r}^{\mathrm{full,hb}}\big)^2$. The telescoping identity in Eq. (I.5) gives $\eta\lambda_i K_{i,r}^{\mathrm{full,hb}} = (1 - \phi_{i,T-1-r})/T$. Therefore

$$\sum_{r=k_0^{\mathrm{hb}}}^{T-2} \kappa_r^{\mathrm{hb}} = \frac{1}{T^2} \sum_i \sum_{L=1}^{T-1-k_0^{\mathrm{hb}}} (1-\phi_{i,L})^2.$$

When $T \geq 4k_0^{\mathrm{hb}}$, the inner sum contains the block $\lceil T/2 \rceil \leq L \leq \lfloor 3T/4 \rfloor$. Consider first the nonnegative-root region $0 < d_i \leq (1-\sqrt{\mu})^2$. The argument used in Subsection I.4 applies uniformly throughout this block and gives $1 - \phi_{i,L} \gtrsim \min\{1, T d_i/\tau\}$. Since the block contains a constant multiple of T indices,

$$\frac{1}{T^2} \sum_{L=\lceil T/2 \rceil}^{\lfloor 3T/4 \rfloor} (1-\phi_{i,L})^2 \gtrsim \min\left\{ \frac{1}{T}, T \left(\frac{d_i}{\tau} \right)^2 \right\}.$$

It remains to consider $d_i > (1-\sqrt{\mu})^2$. The partial-sum estimate established in the proof of Theorem 5.4 holds for every endpoint: $|\sum_{L=0}^{n}\phi_{i,L}| \le 14/\tau$. The sum of $\phi_{i,L}$ over the block $[T/2, 3T/4]$ is consequently bounded in absolute value by $28/\tau$. Hence

$$\sum_{L=\lceil T/2\rceil}^{\lfloor 3T/4\rfloor}(1-\phi_{i,L})^2 \ge \frac{T}{4} - \frac{56}{\tau} - 2.$$

For $T\tau \ge 256$, the right-hand side is bounded below by a universal constant times T. In this root region, $d_i > (1-\sqrt{\mu})^2 \ge \tau^2/4$, so $T\tau \ge 256$ also implies $Td_i/\tau \ge 64$. Thus the active branch of the filter is T^{-1}, and the same directional lower bound follows. Combining the two root regions with the upper half of Eq. (5.3) yields

$$\sum_{r=k_0^{\mathrm{hb}}}^{T-2}\kappa_r^{\mathrm{hb}} \gtrsim \sum_{r=0}^{T-2}\kappa_r^{\mathrm{hb}} = 2\mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}). \tag{I.3}$$

The comparison constant is universal. Notice that no monotonicity of the oscillatory impulse response is used; the argument only uses the energy contained in the middle time block. We now return to the stochastic feedback. The exact innovation decomposition and the lower fourth-moment comparison give

$$V_{\mathrm{full}}^{\mathrm{hb}} \ge \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{\Sigma}) + \frac{\beta}{2}\sum_{r=0}^{T-2} r_r^{\mathrm{hb}}\kappa_r^{\mathrm{hb}}.$$

All summands are nonnegative. Restricting the sum to $r \ge k_0^{\mathrm{hb}}$ and applying Eq. (I.3) gives

$$V_{\mathrm{full}}^{\mathrm{hb}} \gtrsim \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{\Sigma}) + \beta\nu\,\mathrm{tr}(\mathbf{\Sigma})\mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}). \tag{I.4}$$

For the reverse inequality, the sharp closure theorem gives

$$V_{\mathrm{full}}^{\mathrm{hb}} \le \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{\Sigma}) + \frac{\psi q_*(\mathbf{\Sigma})}{\Delta_*}\mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}).$$

Under $\psi\nu\,\mathrm{tr}(\mathbf{H}) < 2$, one has $0 < \nu\lambda_i < 2$. Since

$$\frac{q_*(\mathbf{\Sigma})}{\mathrm{tr}(\mathbf{\Sigma})} = \frac{\sum_i \frac{\nu\Sigma_{ii}}{2+\frac{2\mu}{1+\mu}\nu\lambda_i}}{\sum_i \Sigma_{ii}},$$

the denominator in each summand lies between 2 and 4. Thus $\nu\,\mathrm{tr}(\mathbf{\Sigma})/4 \le q_*(\mathbf{\Sigma}) \le \nu\,\mathrm{tr}(\mathbf{\Sigma})/2$. If $\Delta_* \ge \delta > 0$, the preceding upper bound becomes

$$V_{\mathrm{full}}^{\mathrm{hb}} \lesssim_{\psi,\delta} \mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{\Sigma}) + \nu\,\mathrm{tr}(\mathbf{\Sigma})\mathcal{F}_{\mathrm{full}}^{\mathrm{hb}}(\mathbf{H}).$$

Together with Eq. (I.4), this proves Theorem 5.3. The cutoff form follows from Theorem 5.4.

I.4 Proof of Theorem 5.4

The identities

$$\phi_{t+1} - \phi_t = -dg_t, \quad K_r^{\mathrm{hb}} = \frac{1-\phi_{T-1-r}}{Td} \tag{I.5}$$

prove Eq. (5.2). Let $d_- := (1-\sqrt{\mu})^2$. If $0 < d \le d_-$, both roots are nonnegative and write them as $r_+ = 1-s$ and $r_- = 1-t$, where $0 < s \le t \le 1$. Comparing the sum and product of the roots gives the exact identities

$$st = d, \quad s + t = \tau + d. \tag{I.6}$$

Moreover,

$$dg_j = st\sum_{k=0}^{j}(1-s)^{j-k}(1-t)^k.$$

The right-hand side is the probability mass function of $X+Y$, where X and Y are independent geometric random variables on $\{0,1,\dots\}$ with success probabilities s and t. Consequently,

$$1-\phi_L = d\sum_{j=0}^{L-1} g_j = \Pr(X+Y \le L-1).$$

Because $d \le d_- \le \tau$, Eq. (I.6) implies

$$\frac{\tau}{2} \le t \le 2\tau, \qquad \frac{d}{2\tau} \le s = \frac{d}{t} \le \frac{2d}{\tau}.$$

For $L \ge T/2$ and $T\tau \ge 64$,

$$\Pr(Y \le \lfloor L/2 \rfloor) = 1 - (1-t)^{\lfloor L/2 \rfloor + 1} \ge 1 - e^{-tL/2} \ge 1 - e^{-8},$$
$$\Pr(X \le \lfloor L/2 \rfloor - 1) \asymp \min\{1, Ls\} \asymp \min\{1, Ld/\tau\}.$$

Independence and the inclusion $\{X \le \lfloor L/2 \rfloor - 1, Y \le \lfloor L/2 \rfloor\} \subseteq \{X + Y \le L - 1\}$ give the lower bound

$$1 - \phi_L \gtrsim \min\{1, Ld/\tau\}.$$

For the upper bound, $X + Y \le L - 1$ implies $X \le L - 1$, so

$$1 - \phi_L \le \Pr(X \le L - 1) \lesssim \min\{1, Ld/\tau\}.$$

Thus the desired comparison holds uniformly on the final half of the time horizon, which is exactly the block needed in Eq. (5.2). Summing this block proves the matching bound in the nonnegative-root region. If $d_- < d < 2\tau$, the telescoping identity also gives

$$\sum_{L=0}^{T-1} \phi_L = \frac{\tau}{d}(1 - \phi_{T-1}) + g_{T-1}. \tag{I.7}$$

We next record the uniform root bounds used below. For distinct roots,

$$g_t = \frac{r_+^{t+1} - r_-^{t+1}}{r_+ - r_-}, \quad \phi_t = \frac{(r_+ - \mu) r_+^t - (r_- - \mu) r_-^t}{r_+ - r_-}; \tag{I.8}$$

at a repeated root these expressions are interpreted by differentiation. In the nonnegative-root region, $g_t \ge 0$ and $\phi_{t+1} - \phi_t = -d g_t$ imply $0 \le \phi_t \le 1$. In the complex-root region, write $r_\pm = \sqrt{\mu} e^{\pm \mathrm{i}\omega}$ in Eq. (I.8). The sine-difference formula and

$$(t+1)(\sqrt{\mu})^t \le \sum_{j=0}^{t} (\sqrt{\mu})^j \le \frac{1}{1 - \sqrt{\mu}} \le \frac{2}{\tau}$$

give $|g_t| \le 2/\tau$; substituting the same representation into $g_t - \mu g_{t-1}$ and using $d < 2\tau$ gives $|\phi_t| \le 2$. In the negative-real-root region, set $u_\pm = -r_\mp \in (0, 1)$. Formula Eq. (I.8) becomes an alternating divided difference of u^t. The mean-value theorem gives $|g_t| \le (t+1)\max(u_+, u_-)^t$, while the coefficients in ϕ_t have opposite signs and total absolute mass at most two under $d < 2\tau$. This again gives

$$|g_t| \le \frac{2}{\tau}, \qquad |\phi_t| \le 2. \tag{I.9}$$

The repeated-root cases follow by the differentiated formulas and the same envelope, so the constants do not deteriorate at either discriminant boundary.

Using Eq. (I.9) in Eq. (I.7), and noting that $d_- = \tau^2/(1 + \sqrt{\mu})^2 \ge \tau^2/4$,

$$\left| \sum_{L=0}^{T-1} \phi_L \right| \le \frac{14}{\tau}.$$

Consequently, we have

$$\sum_{L=1}^{T-1} (1 - \phi_L)^2 = (T-1) - 2\sum_{L=1}^{T-1} \phi_L + \sum_{L=1}^{T-1} \phi_L^2 \ge (T-1) - \frac{30}{\tau} \gtrsim T$$

when $T\tau \ge 64$. The matching upper bound follows from $|1 - \phi_L| \le 3$. Finally, $d > d_- \ge \tau^2/4$ and $T\tau \ge 64$ imply that the active branch of the minimum is T^{-1}. Combining the two root regions proves the theorem.

J Complete Initialization-Bias Analysis

This section keeps both components of the initialization error. The first is the deterministic mean response. The second is the fluctuation generated when the random Hessian acts on the initialization process. The latter is absent if one replaces $\boldsymbol{x}_t \otimes \boldsymbol{x}_t$ by $\mathbf{H}$, and therefore cannot be omitted from the bias of stochastic gradient methods.

J.1 A common identity for SGD readouts

Recall the initialization process $\mathbf{b}_{t+1} = (\mathbf{I} - \eta \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{b}_t$, with $\boldsymbol{b}_0 = \boldsymbol{e}_0$, and set $\mathbf{A} := \mathbf{I} - \eta\mathbf{H}$ and $\mathbf{D}_t := \mathbb{E}[\mathbf{b}_t \otimes \mathbf{b}_t]$. Writing $\boldsymbol{\zeta}_{t+1} := (\mathbf{H} - \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{b}_t$ gives $\mathbf{b}_{t+1} = \mathbf{A}\mathbf{b}_t + \eta\boldsymbol{\zeta}_{t+1}$. Conditional on the past, $\mathbb{E}[\boldsymbol{\zeta}_{t+1} \mid \mathcal{F}_t] = 0$, and hence innovations from different times are orthogonal in L_2. Moreover, $\mathbb{E}[\boldsymbol{\zeta}_{t+1} \otimes \boldsymbol{\zeta}_{t+1}] = \mathcal{N}(\boldsymbol{D}_t)$ and $\mathbb{E}\mathbf{b}_t = \mathbf{A}^t \boldsymbol{e}_0$. For deterministic weights $\alpha = (\alpha_0, \dots, \alpha_{T-1})$, define

$$W_\alpha(\mathbf{A}) := \sum_{t=0}^{T-1} \alpha_t \mathbf{A}^t, \qquad K_r^\alpha(\mathbf{A}) := \sum_{t=r+1}^{T-1} \alpha_t \mathbf{A}^{t-1-r}.$$

Iterating the recursion and exchanging the two finite sums yields

$$\widehat{\mathbf{b}}_\alpha = W_\alpha(\mathbf{A})\mathbf{e}_0 + \eta \sum_{r=0}^{T-2} K_r^\alpha(\mathbf{A})\boldsymbol{\zeta}_{r+1}.$$

The first term is deterministic and the second has mean zero. Orthogonality of the innovations therefore gives the exact decomposition

$$B_\alpha = \underbrace{\frac{1}{2}\|W_\alpha(\mathbf{A})\mathbf{e}_0\|_{\mathbf{H}}^2}_{B_\alpha^{\mathrm{mean}}} + \underbrace{\frac{\eta^2}{2}\sum_{r=0}^{T-2} \operatorname{tr}\big(\mathbf{H}[K_r^\alpha(\mathbf{A})]^2 \mathcal{N}(\mathbf{D}_r)\big)}_{B_\alpha^{\mathrm{fluc}}}. \tag{J.1}$$

Thus, the mean filter alone is not the SGD bias. The second term in equation J.1 is the initialization-induced Hessian fluctuation. The fluctuation term can also be characterized without hiding it in an unspecified constant. Let

$$\rho_t := \operatorname{tr}(\mathbf{H}\mathbf{D}_t), \qquad u_t := \|\mathbf{A}^t\mathbf{e}_0\|_{\mathbf{H}}^2, \quad h_j := \eta^2 \operatorname{tr}(\mathbf{H}^2\mathbf{A}^{2j}) = \eta^2 \sum_i \lambda_i^2 (1-\eta\lambda_i)^{2j}.$$

Duhamel's formula for $\mathbf{D}_t$ and the fourth-moment comparison imply

$$u_t + \beta \sum_{r=0}^{t-1} h_{t-1-r}\rho_r \le \rho_t \le u_t + \psi \sum_{r=0}^{t-1} h_{t-1-r}\rho_r.$$

For $c \in \{\beta, \psi\}$, let $q^{(c)}$ be the nonnegative renewal sequence

$$q_t^{(c)} = u_t + c \sum_{r=0}^{t-1} h_{t-1-r} q_r^{(c)}, \quad q_0^{(c)} = u_0.$$

Positivity gives $q_t^{(\beta)} \le \rho_t \le q_t^{(\psi)}$. If $j_{\alpha,r} := \eta^2 \operatorname{tr}(\mathbf{H}^2[K_r^\alpha(\mathbf{A})]^2)$, then equation J.1 and the same moment comparison yield

$$B_\alpha^{\mathrm{mean}} + \frac{\beta}{2}\sum_{r=0}^{T-2} j_{\alpha,r} q_r^{(\beta)} \le B_\alpha \le B_\alpha^{\mathrm{mean}} + \frac{\psi}{2}\sum_{r=0}^{T-2} j_{\alpha,r} q_r^{(\psi)}. \tag{J.2}$$

Equation equation J.2 is useful when no source condition is imposed on $\mathbf{e}_0$: it is a two-sided, finite-time characterization, and every quantity on its right-hand side is deterministic. Under the stability condition $m = \psi \sum_{j\ge 0} h_j < 1$, the upper renewal sequence has the convergent expansion $q^{(\psi)} = \sum_{n\ge 0} (\psi h)^{*n} * u$. Summing the renewal equation gives a useful history bound. Namely,

$$\mathfrak{H}_0^{\mathrm{sgd}} := \sum_{t\ge 0} u_t = \sum_i \frac{e_{0,i}^2}{\eta(2-\eta\lambda_i)}, \quad \sum_{t\ge 0} q_t^{(\psi)} = \frac{\mathfrak{H}_0^{\mathrm{sgd}}}{1-m}.$$

This infinite-horizon identity is useful for establishing summability and provides a valid coarse upper bound. It does not, however, retain the finite-horizon spectral transition at $\eta\lambda_i \asymp T^{-1}$. In the calculations below we instead retain the finite history

$$\sum_{t=0}^{T-1} u_t \asymp \frac{1}{\eta}\,\mathsf{C}_T(\boldsymbol{e}_0), \quad \mathsf{C}_T(\boldsymbol{e}_0) := \sum_{i\le k_T} e_{0,i}^2 + T\eta \sum_{i>k_T} \lambda_i e_{0,i}^2,$$

where the index k_T is introduced only after balancing T and $(\eta\lambda_i)^{-1}$ in the geometric history sum.

J.2 Full averaging

For $\alpha_t = 1/T$, put $\delta_i := \eta\lambda_i$ and $a_i := 1-\delta_i$. Direct summation gives

$$W_T^{\mathrm{full}}(a_i) = \frac{1-a_i^T}{T\delta_i}, \quad K_{i,r}^{\mathrm{full}} = \frac{1-a_i^{T-1-r}}{T\delta_i}.$$

Substitution into equation J.1 gives the complete full-average bias:

$$B_T^{\mathrm{full}} = \frac{1}{2}\sum_i \lambda_i e_{0,i}^2 \left(\frac{1-a_i^T}{T\delta_i}\right)^2 + \frac{\eta^2}{2}\sum_{r=0}^{T-2} \mathrm{tr}\big(\mathbf{H}[K_r^{\mathrm{full}}(\mathbf{A})]^2 \mathcal{N}(\mathbf{D}_r)\big)\,. \tag{J.3}$$

The cutoff is obtained from the mean filter rather than introduced in advance. Since $1-(1-\delta)^T \asymp \min\{1, T\delta\}$ for $0<\delta<1$, the two branches are 1 and $(T\delta)^{-1}$. Equating them gives $T\delta \asymp 1$, or $\lambda \asymp (\eta T)^{-1}$. If $k_T := \max\{i : \lambda_i \ge (\eta T)^{-1}\}$ is introduced only after this calculation, then

$$B_{T,\mathrm{mean}}^{\mathrm{full}} \asymp \frac{1}{\eta^2 T^2}\sum_{i\le k_T} \frac{e_{0,i}^2}{\lambda_i} + \sum_{i>k_T} \lambda_i e_{0,i}^2.$$

The fluctuation part is obtained from equation J.2 with

$$j_{\mathrm{full},r} = \frac{1}{T^2}\sum_i \big(1-a_i^{T-1-r}\big)^2.$$

The sequence $j_{\mathrm{full},r}$ decreases with r. Comparing the two branches 1 and $(T\eta\lambda_i)^2$ gives

$$J_T^{\mathrm{out}} := j_{\mathrm{full},0} \asymp \frac{k_T}{T^2} + \eta^2 \sum_{i>k_T} \lambda_i^2.$$

The input history has a cutoff as well. Define

$$\mathsf{J}_T(\boldsymbol{e}_0) := \sum_i \lambda_i e_{0,i}^2 \min\{1, (T\eta\lambda_i)^{-1}\} = \frac{1}{\eta T}\sum_{i\le k_T} e_{0,i}^2 + \sum_{i>k_T} \lambda_i e_{0,i}^2$$

and

$$\mathsf{S}_T := \sum_i \min\{T^{-1}, T\eta^2\lambda_i^2\} = \frac{k_T}{T} + T\eta^2 \sum_{i>k_T} \lambda_i^2.$$

Notice that $\mathsf{C}_T(\boldsymbol{e}_0) = \eta T \mathsf{J}_T(\boldsymbol{e}_0)$ and $\mathsf{S}_T = T J_T^{\mathrm{out}}$. Thus the two finite-time factors combine without losing either cutoff. More precisely, restricting the time sum to $0 \le r < \lfloor T/2 \rfloor$ gives $\sum_r u_r j_{\mathrm{full},r} \gtrsim \mathsf{J}_T(\boldsymbol{e}_0)\mathsf{S}_T$, whereas the renewal expansion and its total mass give the reverse inequality up to the stability factor. Consequently,

$$c\beta\, \mathsf{J}_T(\boldsymbol{e}_0)\mathsf{S}_T \le B_{T,\mathrm{fluc}}^{\mathrm{full}} \le \frac{C\psi}{1-m}\, \mathsf{J}_T(\boldsymbol{e}_0)\mathsf{S}_T.$$

Here $c, C > 0$ are absolute constants. This is the cutoff-bearing initialization result: the same balance $T\eta\lambda_i \asymp 1$ appears in both the finite input history and the full-readout energy.

J.3 TAIL AVERAGING

Let $T = S + M$ and average the iterates $S, \dots, T-1$. The mean response is

$$W^{\text{tail}}_{S,M}(a_i) = a_i^S \frac{1 - a_i^M}{M\delta_i}.$$

The innovation kernel has two pieces because an innovation may occur before or during the averaging window:

$$K^{\text{tail}}_{i,r} = \begin{cases} a_i^{S-1-r} \dfrac{1 - a_i^M}{M\delta_i}, & 0 \le r < S, \\ \dfrac{1 - a_i^{T-1-r}}{M\delta_i}, & S \le r \le T-2. \end{cases}$$

Consequently,

$$B^{\text{tail}}_{S,M} = \frac{1}{2}\sum_i \lambda_i e_{0,i}^2 a_i^{2S} \left(\frac{1 - a_i^M}{M\delta_i}\right)^2 + \frac{\eta^2}{2}\sum_{r=0}^{T-2} \operatorname{tr}\big(\mathbf{H}[K^{\text{tail}}_r(\mathbf{A})]^2 \mathcal{N}(\mathbf{D}_r)\big). \tag{J.4}$$

Here the averaging transition is found by comparing 1 with $(M\delta)^{-1}$, which gives $\lambda \asymp (\eta M)^{-1}$. Only after this balance do we write $k_M := \max\{i : \lambda_i \ge (\eta M)^{-1}\}$. The deterministic mean contribution then satisfies

$$B^{\text{tail}}_{S,M,\text{mean}} \asymp \frac{1}{\eta^2 M^2} \sum_{i \le k_M} \frac{a_i^{2S} e_{0,i}^2}{\lambda_i} + \sum_{i > k_M} \lambda_i a_i^{2S} e_{0,i}^2.$$

The factor a_i^{2S} is the burn-in attenuation and should not be replaced by a second effective-dimension cutoff. The Hessian-fluctuation contribution is given by equation J.2, with

$$j_{\text{tail},r} = \eta^2 \sum_i \lambda_i^2 (K^{\text{tail}}_{i,r})^2.$$

For every r, this quantity is bounded by its value at the left boundary of the averaging window. At $r = S - 1$,

$$J^{\text{tail}}_M := j_{\text{tail},S-1} = \frac{1}{M^2}\sum_i (1 - a_i^M)^2 \asymp \frac{k_M}{M^2} + \eta^2 \sum_{i > k_M} \lambda_i^2.$$

The input history up to the total horizon $T = S + M$ contributes $\eta^{-1}\mathsf{C}_T(\boldsymbol{e}_0)$, while the output factor contributes J^{tail}_M. Hence, for $S \ge 1$,

$$\frac{\beta}{2}\rho_{S-1} J^{\text{tail}}_M \le B^{\text{tail}}_{S,M,\text{fluc}} \le \frac{C\psi}{\eta(1-m)} \mathsf{C}_T(\boldsymbol{e}_0) J^{\text{tail}}_M.$$

Thus tail averaging has two distinct initialization cutoffs: the history cutoff $\lambda \asymp (\eta T)^{-1}$ and the output cutoff $\lambda \asymp (\eta M)^{-1}$. The lower bound retains ρ_{S-1} because a long burn-in may remove most of the initialization energy before averaging begins.

J.4 EXPONENTIALLY WEIGHTED AVERAGING

Let $\theta = 1 - \beta_{\text{ema}}$ and use the recursively initialized readout

$$\widehat{\mathbf{b}}^{\text{ema}}_T = \beta_{\text{ema}}^T \boldsymbol{e}_0 + \theta \sum_{t=0}^{T-1} \beta_{\text{ema}}^{T-1-t} \mathbf{b}_t.$$

For a scalar direction $a_i = 1 - \delta_i$, finite geometric summation gives

$$W^{\text{ema}}_T(a_i) = \beta_{\text{ema}}^T + \theta \frac{a_i^T - \beta_{\text{ema}}^T}{a_i - \beta_{\text{ema}}}, \qquad K^{\text{ema}}_{i,r} = \theta \frac{a_i^{T-1-r} - \beta_{\text{ema}}^{T-1-r}}{a_i - \beta_{\text{ema}}}.$$

At $a_i = \beta_{\text{ema}}$, the quotients are understood by continuity. Thus

$$B^{\text{ema}}_T = \frac{1}{2}\sum_i \lambda_i e_{0,i}^2 [W^{\text{ema}}_T(a_i)]^2 + \frac{\eta^2}{2}\sum_{r=0}^{T-2} \operatorname{tr}\big(\mathbf{H}[K^{\text{ema}}_r(\mathbf{A})]^2 \mathcal{N}(\mathbf{D}_r)\big). \tag{J.5}$$

The exact formula also shows where its transitions come from. If $\delta_i \le \theta/2$, the optimization mode is slower than the EMA memory and $W_T^{\mathrm{ema}}(a_i) \asymp a_i^T$. If $\delta_i \ge 2\theta$, the readout memory is slower and the response is governed by β_{ema}^T. Around $\delta_i = \theta$, the continuous limit is $\beta_{\mathrm{ema}}^T + \theta T \beta_{\mathrm{ema}}^{T-1}$. The finite-horizon learning transition is therefore obtained from $T\delta_i \asymp 1$, giving $\lambda \asymp (\eta T)^{-1}$, while $\delta_i \asymp \theta$ marks the change between optimization-limited and memory-limited decay. These conclusions follow from the exact filter; no cutoff is needed to state equation J.5.

The fluctuation term is again covered by equation J.2, now with $j_{\mathrm{ema},r} = \eta^2 \sum_i \lambda_i^2 (K_{i,r}^{\mathrm{ema}})^2$. Put $L = T - 1 - r$. Since $\eta\lambda_i K_{i,r}^{\mathrm{ema}} = \delta_i\theta(a_i^L - \beta_{\mathrm{ema}}^L)/(a_i - \beta_{\mathrm{ema}})$, this factor is a convolution of two geometric impulse responses. Its two Young bounds give, uniformly in r, $|\eta\lambda_i K_{i,r}^{\mathrm{ema}}| \le \min\{\delta_i, \theta\}$. We therefore define the output envelope

$$\mathsf{P}_\theta := \sum_i \min\{\delta_i^2, \theta^2\}, \quad j_{\mathrm{ema},r} \le \mathsf{P}_\theta.$$

The two branches meet at $\eta\lambda_i \asymp \theta$. Only after this balance, set $k_\theta := \max\{i : \lambda_i \ge \theta/\eta\}$, which gives

$$\mathsf{P}_\theta \asymp \theta^2 k_\theta + \eta^2 \sum_{i > k_\theta} \lambda_i^2.$$

Pairing this output factor with the finite initialization history gives the concrete upper bound

$$B_{T,\mathrm{fluc}}^{\mathrm{ema}} \le \frac{C\psi}{\eta(1-m)} \mathsf{C}_T(\boldsymbol{e}_0)\mathsf{P}_\theta.$$

The initialization fluctuation therefore contains two independently derived transitions: the input-history cutoff $\lambda \asymp (\eta T)^{-1}$ and the EMA output cutoff $\lambda \asymp \theta/\eta$. A matching product lower bound is not valid for arbitrary initial conditions: in a scalar instance the initialization can decay exponentially while this uniform upper envelope decays only polynomially. The exact renewal sandwich above remains the appropriate two-sided formula in that generality.

J.5 Full averaging for stochastic heavy-ball

For the initialization process of SHB, take zero initial momentum, $\mathbf{b}_{-1} = \mathbf{b}_0 = \boldsymbol{e}_0$, and write

$$\mathbf{b}_{t+1} = ((1+\mu)\mathbf{I} - \eta\mathbf{H})\mathbf{b}_t - \mu\mathbf{b}_{t-1} + \eta\boldsymbol{\zeta}_{t+1}, \qquad \boldsymbol{\zeta}_{t+1} = (\mathbf{H} - \mathbf{x}_t \otimes \mathbf{x}_t)\mathbf{b}_t.$$

In direction i, let $g_{i,-1} = 0$, $g_{i,0} = 1$, and $g_{i,t+1} = (1 + \mu - \eta\lambda_i)g_{i,t} - \mu g_{i,t-1}$. The deterministic response to the initial condition is $\phi_{i,t} := g_{i,t} - \mu g_{i,t-1}$, so that $\mathbb{E}b_{i,t} = \phi_{i,t}e_{0,i}$. The innovation kernel of the full average is

$$K_{i,r}^{\mathrm{hb,full}} = \frac{1}{T}\sum_{j=0}^{T-2-r} g_{i,j} = \frac{1 - \phi_{i,T-1-r}}{T\eta\lambda_i},$$

where the second equality follows from $\phi_{i,L} - \phi_{i,L-1} = -\eta\lambda_i g_{i,L-1}$. For completeness, the deterministic initial response satisfies

$$\sum_{t\ge 0} \phi_{i,t}^2 = \frac{\tau^2(1+\mu) + 2\mu d_i}{\tau d_i[2(1+\mu) - d_i]}. \tag{J.6}$$

To derive this identity, use $\phi_t = g_t - \mu g_{t-1}$ together with Eq. (I.1):

$$\sum_{t\ge 0}\phi_t^2 = (1+\mu^2)L_i - 2\mu\sum_{t\ge 1} g_t g_{t-1} = L_i\left(1 + \mu^2 - \frac{2\mu(1+\mu-d_i)}{1+\mu}\right),$$

which simplifies to Eq. (J.6). Furthermore,

$$1 + \ell_i = \frac{2[\tau(1+\mu) + \mu d_i]}{\tau[2(1+\mu) - d_i]}. \tag{J.7}$$

Dividing Eq. (J.6) by Eq. (J.7) gives $\frac{1}{1+\ell_i}\sum_{t\ge 0}\phi_{i,t}^2 = \frac{\tau^2(1+\mu)+2\mu d_i}{2d_i[\tau(1+\mu)+\mu d_i]}$..

Let $\mathbf{D}_t^{\mathrm{hb}} := \mathbb{E}[\mathbf{b}_t \otimes \mathbf{b}_t]$. Repeating the martingale-innovation argument in the augmented two-dimensional state gives

$$B_T^{\mathrm{hb,full}} = \frac{1}{2}\sum_i \lambda_i e_{0,i}^2 \left(\frac{1}{T}\sum_{t=0}^{T-1}\phi_{i,t}\right)^2 + \frac{\eta^2}{2}\sum_{r=0}^{T-2}\sum_i \lambda_i [K_{i,r}^{\mathrm{hb,full}}]^2 [\mathcal{N}(\boldsymbol{D}_r^{\mathrm{hb}})]_{ii}. \tag{J.8}$$

This is the complete SHB bias. Its second line is not label-noise variance: it is generated by the random Hessians acting on the initialization trajectory. For the mean term, summing the deterministic response gives the exact identity

$$\sum_{t=0}^{T-1}\phi_{i,t} = \frac{\tau}{d_i}(1-\phi_{i,T-1}) + g_{i,T-1}, \qquad d_i := \eta\lambda_i, \quad \tau := 1-\mu.$$

In the nonnegative-root region, and under the same condition $T\tau \geq 64$ used in the full-SHB filter theorem, the geometric representation of the impulse response yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\phi_{i,t} \asymp \min\left\{1, \frac{\tau}{Td_i}\right\} = \min\left\{1, \frac{1}{T\nu\lambda_i}\right\}, \qquad \nu := \eta/\tau.$$

The two branches meet at $T\nu\lambda_i \asymp 1$; hence the statistical transition $\lambda \asymp (\nu T)^{-1}$ is derived from the mean response. In the oscillatory and negative-root regions, cancellation can make the signed mean response smaller, so a uniform two-sided cutoff formula is not valid there. The exact expression above should be retained; the root-uniform estimate $|T^{-1}\sum_{t<T}\phi_{i,t}| \lesssim (T\tau)^{-1}$ follows from the partial-sum bound in the SHB filter proof.

It remains to control the second line of equation J.8. Let $X_t^b = (\boldsymbol{X}_{i,t}^b)_i$ be the augmented covariance blocks and $\mathscr{R}(X_t^b) = \sum_i \lambda_i (\mathbf{X}_{i,t}^b)_{11}$. Retaining the directional negative term gives the positive comparison recursion

$$(\mathscr{K} X_t^b)_i = \mathbf{A}_i \mathbf{X}_{i,t}^b \mathbf{A}_i^\top + \eta^2 \left\{\psi\lambda_i \mathscr{R}(X_t^b) - \lambda_i^2 (\mathbf{X}_{i,t}^b)_{11}\right\} \boldsymbol{E}_{11}.$$

If $Y_t^b := \mathscr{K}^t X_0^b$, positivity gives $X_t^b \preceq Y_t^b$. The discounted-resolvent calculation, followed by monotone convergence, yields

$$\sum_{t\geq 0}\mathscr{R}(Y_t^b) = \frac{1}{\Delta_*}\sum_i \frac{\lambda_i e_{0,i}^2}{1+\ell_i}\sum_{t\geq 0}\phi_{i,t}^2. \tag{J.9}$$

To make the implication for the readout explicit, set

$$j_{\mathrm{hb},r} := \eta^2 \sum_i \lambda_i^2 [K_{i,r}^{\mathrm{hb,full}}]^2, \quad J_T^{\mathrm{hb}} := \max_{0\leq r\leq T-2} j_{\mathrm{hb},r}.$$

The fourth-moment comparison and $X_r^b \preceq Y_r^b$ give

$$B_{T,\mathrm{fluc}}^{\mathrm{hb,full}} \leq \frac{\psi}{2}\sum_{r=0}^{T-2} j_{\mathrm{hb},r}\mathscr{R}(Y_r^b) \leq \frac{\psi J_T^{\mathrm{hb}}}{2\Delta_*}\sum_i \frac{\lambda_i e_{0,i}^2}{1+\ell_i}\sum_{t\geq 0}\phi_{i,t}^2.$$

The readout factor can be evaluated explicitly. Since $\eta\lambda_i K_{i,r}^{\mathrm{hb,full}} = (1-\phi_{i,T-1-r})/T$, the root estimates used for the full-SHB filter give

$$J_T^{\mathrm{hb}} \lesssim \frac{k_{\mathrm{hb}}(T)}{T^2} + \nu^2 \sum_{i>k_{\mathrm{hb}}(T)} \lambda_i^2.$$

The index is introduced after balancing the two directional branches 1 and $(T\nu\lambda_i)^2$, which yields $\lambda_i \asymp (\nu T)^{-1}$. Substituting the exact weighted-history identity then gives

$$B_{T,\mathrm{fluc}}^{\mathrm{hb,full}} \lesssim \frac{\psi}{2\Delta_*}\left[\frac{k_{\mathrm{hb}}(T)}{T^2} + \nu^2 \sum_{i>k_{\mathrm{hb}}(T)}\lambda_i^2\right]\sum_i \lambda_i e_{0,i}^2 \frac{\tau^2(1+\mu)+2\mu d_i}{2d_i[\tau(1+\mu)+\mu d_i]}.$$

Likewise, the lower fourth-moment comparison gives $B_{T,\mathrm{fluc}}^{\mathrm{hb,full}} \geq (\beta/2)\sum_{r<T-1} j_{\mathrm{hb},r}\mathbb{E}\|\mathbf{b}_r\|_{\mathbf{H}}^2$. Thus the exact identity and the two comparison bounds retain both the signed deterministic response and the initialization-induced Hessian fluctuation. An explicit polynomial rate requires a source condition on $\boldsymbol{e}_0$; without one, these spectral sums are the appropriate instance-dependent description.

K Limitations and Future Directions

Our analysis is developed for constant-stepsize stochastic methods on population least squares with fresh independent samples. Although we do not require conditionally mean-zero noise, bounded relative noise, or commutativity between $\mathbf{H}$ and $\mathbf{\Sigma}$, the results rely on trace-class covariance and noise operators together with fourth-moment comparison conditions. Extending the framework to finite-data sampling with repeated observations, dependent data streams, time-varying stepsizes, or general nonquadratic objectives would require additional control beyond the present linear covariance recursions.

For SHB, the feedback closure applies to general deterministic linear readouts, and the full-averaging filter admits a matching spectral characterization. We do not, however, derive an equally explicit matching two-cutoff formula for tail-averaged SHB. Establishing such a result would clarify how the averaging window and momentum timescale interact and whether momentum changes only the spectral transition scales or also the attainable finite-time rates. The same transfer function and renewal viewpoint may also be useful for other momentum and multistep stochastic methods.

The experiments are designed to isolate the mechanisms covered by the theory. The synthetic models permit direct evaluation of the exact second-moment recursion and its predicted power-law rates, whereas the CIFAR-10 experiment trains only a linear predictor on frozen ResNet-18 features. Consequently, the latter tests the relevance of the spectral predictions on real neural representations but does not establish the same conclusions for end-to-end nonconvex training. Studying evolving representations, data-dependent choices of the burn-in and averaging window, and adaptive optimization methods provides a natural direction for future work.