How Dynamic Batch Schedules Shape Final Prediction Risk

Anonymous authors
Paper under double-blind review

Abstract

We develop a finite-horizon statistical theory of dynamic mini-batching under warmup–stable–decay learning rates (WSD) for sketched linear regression with power-law spectral and source conditions. WSD induces two effective time scales: deterministic learning is governed by the full horizon γJ, whereas stochastic noise retains only a shorter terminal memory $\gamma J/\log J$. Although a dynamic batch schedule has one degree of freedom per update, its leading stochastic contribution to final risk collapses to an influence-weighted scalar summary: B_{eff} for one-pass WSD-SGD and ρ_{eff} under data reuse. This compression also reveals a structural distinction. One-pass batching controls centered SGD variance, while multi-pass batching controls only fluctuation around the full-batch WSD-GD reference; approximation, GD bias, and GD variance remain schedule-independent at a fixed update horizon. We obtain matching one-pass variance bounds and nearly matching multi-pass fluctuation bounds. As a consequence, minimizing the influence-weighted stochastic terms yields the oracle allocation $B_t \propto \sqrt{\kappa_t}$.

1 Introduction

Scaling laws have become a standard language for describing progress in modern machine learning: across many domains, prediction error follows regular power laws in model size, data size, and compute. This pattern has been documented from early large-scale studies across translation, language, vision, and speech (Hestness et al., 2017) to modern language-model scaling analyses (Kaplan et al., 2020; Hoffmann et al., 2022) and multimodal autoregressive modeling (Henighan et al., 2020). As a result, scaling laws are now used not only to summarize experiments, but also to guide forecasting, resource allocation, and training design (Rosenfeld et al., 2020; Zhai et al., 2022; Alabdulmohsin et al., 2022; Besiroglu et al., 2024; Muennighoff et al., 2023; Paquette et al., 2024).

Empirical scaling laws do not by themselves explain where the exponents come from, which parts of the risk they describe, or how algorithmic choices change them. Rigorous results are therefore rarer. In statistically transparent models, however, approximation, optimization, and sampling effects can be separated. Sketched linear regression with spectral and source conditions provides a setting in which these components admit explicit scaling laws. It also lets us isolate two questions that are not visible in a single-horizon or static-batch description: how WSD separates persistent learning from terminal stochastic memory, and how a time-varying batch schedule is aggregated according to its influence on final risk. This perspective connects to theoretical accounts based on manifold arguments (Sharma & Kaplan, 2020; Bahri et al., 2024), constructive learning curves and solvable models (Hutter, 2021; Maloney et al., 2022), dynamical and feature-learning analyses (Bordelon et al., 2024; 2025), and renormalized high-dimensional asymptotics, quantization, and emergent scaling behavior (Michaud et al., 2023; Atanasov et al., 2024; Dohmatob et al., 2024; Paquette et al., 2024; Ren et al., 2025).

Batch size is one of the most important large-scale training knobs because it affects hardware utilization and gradient noise. Empirically, large-batch rules enabled ImageNet training with batches up to 8192 (Goyal et al., 2017), LARS pushed convolutional training to 8K and 32K (You et al., 2017), and later studies showed that the gains are highly workload-dependent and eventually saturate (Shallue et al., 2019; Golmant et al., 2019). An influential synthesis is the gradient-noise-scale viewpoint of McCandlish et al. (2018), and batch-dependent empirical scaling laws have recently been studied for language models (Shuai et al., 2024). Most recently, the theory-guided empirical

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.

study of Li et al. (2026) proposed joint loss laws for learning-rate and batch-size sequences, used them to derive budget-aware schedules and validated the predictions on deep-model settings.

There is also a growing theoretical understanding of batch size B through SGD noise. Smith & Le (2018) modeled SGD by an SDE with noise scale proportional to $\epsilon N/B$, suggesting that the optimal batch size should grow with both the learning rate and the dataset size N. Smith et al. (2018) studied the closely related strategy of increasing batch size instead of decaying the learning rate. In least-squares and related settings, the statistical effects of mini-batching, multiple passes, tail averaging, and implicit regularization have also been analyzed extensively (Lin & Rosasco, 2017; Jain et al., 2017; Mücke et al., 2019; Ge et al., 2019; Zou et al., 2021; Wu et al., 2022; Pillaud-Vivien et al., 2018). Our focus is a finite-horizon statistical account of how WSD separates learning from stochastic memory and how every update of an arbitrary dynamic schedule contributes to final risk under one pass and data reuse.

This paper develops such a theory for sketched linear regression trained with mini-batch methods. We study two stochastic procedures: one-pass WSD-SGD and multi-pass WSD-SGD, with each multi-pass batch sampled uniformly without replacement. For one-pass training we allow an arbitrary deterministic batch-size schedule $B_1, \dots, B_T$ with $\sum_t B_t = N$; for multi-pass training we allow any deterministic schedule $B_1, \dots, B_L$ with $1 \le B_t \le N$. We work within the sketched linear-regression framework of Lin et al. (2024; 2025), while focusing on the mechanisms specific to WSD and dynamic batching. Our main contributions are as follows.

- **Two effective horizons under WSD.** We give a precise finite-horizon characterization: deterministic learning uses the full cumulative horizon γJ, while stochastic noise is governed by the shorter terminal-memory scale $\gamma J/\log J$.
- **Final-risk influence and schedule compression.** The influence $\kappa_t^{(J)}$ quantifies how much noise injected at update t survives all subsequent updates. Consequently, an arbitrary one-pass schedule enters the leading variance scaling only through B_{eff}, and an arbitrary multi-pass schedule enters the leading fluctuation scaling only through ρ_{eff}; see Theorems 3.1 and 3.2.
- **Data reuse changes the role of batching.** One-pass batching controls centered SGD variance. Under data reuse it controls fluctuation around full-batch WSD-GD, while approximation, GD bias, and GD variance are schedule-independent at fixed horizon.
- **Influence-matched allocation as a consequence.** At fixed horizon and sample-processing budget, minimizing the resulting influence-weighted objectives gives $B_t^\star \propto \sqrt{\kappa_t}$; see Corollary 3.1.

Notation. For positive functions f and g, we write $f \lesssim g$ and $f \gtrsim g$ for upper and lower bounds up to logarithmic factors, and $f \asymp g$ when both hold. We use $\Theta(g)$ for two-sided scaling up to positive constant (*i.e.*, fixed and independent of compute footings M, N, T, B, L) factors only, whereas $\widetilde{\Theta}(g)$ may additionally suppress multiplicative logarithmic factors. Vector inner products are denoted by $\langle u, v\rangle$ or $u^\top v$, and matrix inner products by $\langle A, B\rangle := \mathrm{tr}(A^\top B)$. The symbol $\|\cdot\|$ denotes the operator norm for matrices and the ℓ_2-norm for vectors; $\|v\|_A^2 := v^\top A v$ for PSD A. We write $A \preceq B$ when $B - A$ is PSD, $\mu_j(A)$ and $r(A)$ for the eigenvalues and rank of a symmetric matrix, and $\log$ for the natural logarithm.

2 Preliminaries

We use WSD-SGD to denote mini-batch stochastic gradient descent with a warmup–stable–decay learning-rate schedule, and WSD-GD for its full-batch counterpart. We study the full-batch iterate θ_t, one-pass WSD-SGD, and multi-pass WSD-SGD. Each stochastic update t averages a mini-batch of size B_t. When $B_t = 1$ at every update, the one-pass procedure reduces to single-sample WSD-SGD. In the multi-pass method, each batch is sampled without replacement, independently across iterations; this procedure should therefore not be confused with random reshuffling across an epoch.

Problem setup. Let $\mathcal{H}$ be a finite- or countably infinite-dimensional Hilbert space, and let P be a Borel probability measure on $\mathcal{H} \times \mathbb{R}$. A generic input–response pair is denoted by $(x, y) \sim P$.

The input covariance is $H := \mathbb{E}[xx^\top]$, with eigenvalues $(\lambda_i)_{i\geq 1}$. For $w \in \mathcal{H}$, define the population square-loss risk and a population minimizer by

$$R(w) := \mathbb{E}\big[(\langle x, w\rangle - y)^2\big], \qquad w^* \in \arg\min_{w\in\mathcal{H}} R(w).$$

Rather than observing x directly, we observe its M-dimensional sketch Sx, where $S : \mathcal{H} \to \mathbb{R}^M$. Throughout the main results, S is a Gaussian sketch: in an orthonormal eigenbasis of H, its entries are independent $\mathcal{N}(0, 1/M)$ random variables. A sketched parameter $u \in \mathbb{R}^M$ represents the original-space predictor $S^\top u$, whose population risk is

$$R_M(u) := R(S^\top u) = \mathbb{E}\big[(\langle Sx, u\rangle - y)^2\big].$$

Given N i.i.d. samples $D = \{(x_i, y_i)\}_{i=1}^N$ drawn from P, we observe $O = \{(Sx_i, y_i)\}_{i=1}^N$. Write the empirical covariates and responses as $X := (x_1, \dots, x_N)^\top$ and $y := (y_1, \dots, y_N)^\top$, and define

$$\Sigma := SHS^\top, \qquad \widehat{\Sigma} := \frac{1}{N} SX^\top XS^\top, \qquad \widehat{b} := \frac{1}{N} SX^\top y.$$

Conditioned on S, the minimizer of R_M is

$$u^* = (SHS^\top)^{-1} SHw^* = \Sigma^{-1} SHw^*.$$

Optimization procedures. We compare three methods using the same warmup–stable–decay (WSD) schedule family. For a generic update horizon J, choose a warmup endpoint J_{w} and a stable-phase endpoint J_{s} with $1 \leq J_{\mathrm{w}} < J_{\mathrm{s}} < J$, and define

$$\gamma_t^{(J)} = \begin{cases} \gamma t / J_{\mathrm{w}}, & 1 \leq t \leq J_{\mathrm{w}}, \\ \gamma, & J_{\mathrm{w}} < t \leq J_{\mathrm{s}}, \\ \gamma \exp\Big(-\frac{(t-J_{\mathrm{s}})\log J}{J - J_{\mathrm{s}}}\Big), & J_{\mathrm{s}} < t \leq J. \end{cases} \tag{1}$$

For fixed constants $c_{\mathrm{s}}, c_{\mathrm{d}} > 0$, we assume $J_{\mathrm{s}} - J_{\mathrm{w}} \geq c_{\mathrm{s}} J$ and $J - J_{\mathrm{s}} \geq c_{\mathrm{d}} J$.

For one-pass WSD-SGD we instantiate $J = T$ and drop the superscript on $\gamma_t^{(T)}$. For full-batch WSD-GD and multi-pass WSD-SGD we instantiate $J = L$ and write the nominal rates as γ_t. The nominal learning horizons scale as γT and γL, while the final stochastic memory scales as $\gamma T/\log T$ and $\gamma L/\log L$, respectively. We retain distinct symbols because T is constrained by the one-pass data partition, whereas L is an independently chosen data-reuse horizon.

Below, the superscripts op and mp distinguish procedure-specific iterates and empirical batch statistics. Shared quantities, including B_t and I_t, remain unsuperscripted. For data reuse, η_t denotes the safeguarded rate defined below.

1. Full-Batch WSD-GD. Using the same safeguarded WSD rates η_t as the multi-pass method below, the full-batch reference iterate is

$$\theta_t = \theta_{t-1} - \eta_t \widehat{\Sigma}\theta_{t-1} + \eta_t \widehat{b}, \qquad \theta_0 = 0. \tag{2}$$

2. One-pass WSD-SGD. Fix $T \geq 3$ and arbitrary positive integers $B_1, \dots, B_T$ satisfying $\sum_{t=1}^T B_t = N$, and partition $[N]$ into disjoint batches $[N] = I_1 \sqcup \cdots \sqcup I_T$, $|I_t| = B_t$. For each batch, define

$$\widehat{\Sigma}_t^{\mathrm{op}} := \frac{1}{B_t} \sum_{i\in I_t} Sx_i x_i^\top S^\top, \qquad \widehat{b}_t^{\mathrm{op}} := \frac{1}{B_t} \sum_{i\in I_t} Sx_i y_i.$$

The one-pass WSD-SGD iterate is

$$u_t^{\mathrm{op}} = u_{t-1}^{\mathrm{op}} - \gamma_t \widehat{\Sigma}_t^{\mathrm{op}} u_{t-1}^{\mathrm{op}} + \gamma_t \widehat{b}_t^{\mathrm{op}}, \qquad t = 1, \dots, T, \qquad u_0^{\mathrm{op}} = 0. \tag{3}$$

Thus the method performs T updates and uses every observation exactly once; the one-pass constraint is $\sum_{t=1}^T B_t = N$, not $T = N/B$.

3. Multi-pass WSD-SGD. Fix a deterministic schedule $B_1, \dots, B_L$ with $1 \le B_t \le N$. At each step $t \in [L]$, independently sample a subset $I_t \subset [N]$ of size B_t, uniformly without replacement within the mini-batch.

Following the data-reuse safeguard of Lin et al. (2025), set

$$\gamma_{\text{safe}} := \min\left\{\gamma, \frac{1}{8\max_{i\in[N]}\|Sx_i\|_2^2}\right\}, \qquad r := \frac{\gamma_{\text{safe}}}{\gamma}, \qquad \eta_t := r\gamma_t. \tag{4}$$

Both multi-pass WSD-SGD and its full-batch reference use these same rates, fixed conditional on the training features. One-pass rates are unchanged; the scaling laws and influence weights use the nominal rates γ_t.

For each sampled batch I_t, define

$$\widehat{\Sigma}_t^{\text{mp}} := \frac{1}{B_t}\sum_{i\in I_t} Sx_i x_i^\top S^\top, \qquad \widehat{b}_t^{\text{mp}} := \frac{1}{B_t}\sum_{i\in I_t} Sx_i y_i.$$

The multi-pass WSD-SGD iterate is

$$u_t^{\text{mp}} = u_{t-1}^{\text{mp}} - \eta_t\widehat{\Sigma}_t^{\text{mp}}u_{t-1}^{\text{mp}} + \eta_t\widehat{b}_t^{\text{mp}}, \qquad t = 1, \dots, L, \qquad u_0^{\text{mp}} = 0. \tag{5}$$

The number of updates remains L, independently of the schedule. There is no one-pass constraint on $\sum_t B_t$; when a gradient-evaluation budget is needed, we write $D_{\text{grad}} := \sum_{t=1}^{L} B_t$ to avoid conflict with the dataset notation D. If $B_t = N$ at every step, this procedure is exactly full-batch WSD-GD.

In summary, all three procedures optimize the same sketched empirical problem under a common WSD schedule. The one-pass procedure allows a deterministic schedule $(B_t)_{t=1}^{T}$ with $\sum_t B_t = N$, whereas the multi-pass procedure allows any deterministic schedule $(B_t)_{t=1}^{L}$ with $1 \le B_t \le N$. When comparing a stochastic iterate with its deterministic reference, we always use the same stepsize schedule and the same number of updates: T for one-pass case and L for multi-pass case.

Unless a subscript indicates otherwise, expectations in the theorem statements are taken over w^*, the sampled dataset, and the mini-batch randomness when applicable. Probability statements over the Gaussian sketch S are stated separately. In Proposition 3.1, $\bar{u}_T$ is instead the conditional mean given (S, w^*), and the displayed decompositions condition on this pair.

3 Main Results

This section develops scaling laws for one-pass and multi-pass sketched WSD-SGD under the power-law/source-condition model. The central phenomenon is that WSD creates two distinct time scales: deterministic learning depends on the full optimization trajectory, whereas stochastic error is governed by a shorter terminal-memory horizon. Dynamic batch sizes do not alter the deterministic contributions at a fixed update horizon; instead, their effects are aggregated through final-risk influence weights in the one-pass variance and multi-pass fluctuation.

We first state the assumptions needed for the WSD and dynamic-batch analysis, then give a common risk decomposition and derive the two scaling laws.

Assumption 1 (Data assumptions)**.** *Assume the following conditions on the data distribution P.*

- ***A. Gaussian design.*** *The feature vector satisfies $x \sim \mathcal{N}(0, H)$.*
- ***B. Homoscedastic label model.*** *Conditional on w^*, the response satisfies*
$$y = \langle x, w^*\rangle + \varepsilon, \qquad \varepsilon \perp x, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{E}[\varepsilon^2] = \sigma^2.$$
- ***C. Power-law spectrum.*** *Eigenvalues of H satisfy $\lambda_i \asymp i^{-a}$ for some $a > 1$.*

Interpretation. Assumption 1 is a stylized random-design regression model. Gaussian features and homoscedastic noise are idealizations, but they are standard in least-squares theory because they isolate spectral effects and provide exact moment identities; related SGD analyses work under broader moment-controlled designs (Dieuleveut & Bach, 2016; Jain et al., 2017). The power law means that the data contain many progressively weaker directions, rather than having a sharp finite-dimensional cutoff. Such spectral decay is a common capacity model in learning theory and has also been observed in high-dimensional neural population responses (Stringer et al., 2019).

Assumption 2 (Source condition in diagonal coordinates)**.** *Assume without loss of generality that H is diagonal with non-increasing diagonal entries $H = \mathrm{diag}(\lambda_1, \lambda_2, \dots)$. Assume the true parameter w^* satisfies: for some $b > 1$,*

$$\mathbb{E}[w_i^* w_j^*] = 0 \quad \textit{for all } i \neq j; \qquad \mathbb{E}[\lambda_i (w_i^*)^2] \asymp i^{-b} \quad \textit{for all } i \geq 1.$$

Interpretation. Assumption 2 measures how the target aligns with the eigendirections of the data covariance. Larger b places less target energy in weak, high-index directions and therefore represents a smoother or easier prediction problem. Conditions of this type are usually called source conditions and are standard in kernel regression and statistical inverse problems because, together with eigenvalue decay, they determine approximation and learning rates (Caponnetto & De Vito, 2007; Dieuleveut & Bach, 2016; Pillaud-Vivien et al., 2018). Diagonalizing H is only a coordinate choice; the uncorrelated-coordinate prior is the substantive average-case simplification.

For both methods, let $\mu_1, \dots, \mu_M$ be the eigenvalues of $\Sigma = SHS^\top$, and define the final-risk influence of update t over a total horizon J by

$$\kappa_t^{(J)} := \gamma_t^2 \sum_{j=1}^{M} \mu_j^2 \prod_{s=t+1}^{J} (1 - \gamma_s \mu_j)^2, \qquad t = 1, \dots, J.$$

Similarly, $J = T$ denotes the one-pass influence and $J = L$ for multi-pass. For one-pass and multi-pass respectively, define the effective batch size and effective fluctuation prefactor

$$\frac{1}{B_{\text{eff}}} := \Big(\sum_{t=1}^{T} \frac{\kappa_t^{(T)}}{B_t}\Big) \Big/ \Big(\sum_{t=1}^{T} \kappa_t^{(T)}\Big), \quad \rho_{\text{eff}} := \Big(\sum_{t=1}^{L} \rho_t \kappa_t^{(L)}\Big) \Big/ \Big(\sum_{t=1}^{L} \kappa_t^{(L)}\Big), \quad \text{where } \rho_t := \frac{N - B_t}{B_t (N-1)}.$$

On the full-rank sketch event the denominator is positive. If $B_{\min} := \min_t B_t$ and $B_{\max} := \max_t B_t$, then $B_{\min} \leq B_{\text{eff}} \leq B_{\max}$; a constant schedule $B_t \equiv B$ gives $B_{\text{eff}} = B$.

Remark 3.1 (Final-risk influence and schedule compression)**.** The nominal weight $\kappa_t^{(J)}$ measures how much stochastic-gradient noise injected at iteration t remains in the prediction risk at the final iterate. Under the power-law spectrum and before the spectral cutoff reaches the model dimension, it has the scaling

$$\kappa_t^{(J)} \asymp \gamma_t^2 \big(1 + R_t^{(J)}\big)^{1/a - 2}, \qquad R_t^{(J)} := \sum_{s=t+1}^{J} \gamma_s;$$

see Appendix A.8 for the justification and more interpretation. Thus different batch trajectories can be equivalent at the level of leading final-risk scaling: with all other parameters fixed, comparable B_{eff} gives the same leading one-pass variance scaling, while comparable ρ_{eff} gives comparable leading multi-pass fluctuation. This concerns expected final prediction risk, not equality of realized optimization trajectories.

The next proposition packages the two stochastic procedures into one structural statement. Their risks share the same baseline contribution: the one-pass method further splits into bias plus variance, while the multi-pass method splits into a common GD reference contribution plus a fluctuation term.

Proposition 3.1 (Risk decompositions for the two stochastic procedures)**.** Assume Assumption 1.B. Define $\bar{u}_T(S, w^*) := \mathbb{E}[u_T^{\text{op}} \mid S, w^*]$. The following identities condition on (S, w^*); we suppress this conditioning in $\mathbb{E}$ and the arguments of $\bar{u}_T$:

$$\mathbb{E}[R_M(u_T^{\text{op}})] = \underbrace{R_M(u^*)}_{\substack{\text{common baseline risk}}} + \underbrace{R_M(\bar{u}_T) - R_M(u^*)}_{\substack{\text{one-pass bias}\\ \text{excess risk}}} + \underbrace{\mathbb{E}\big[R_M(u_T^{\text{op}}) - R_M(\bar{u}_T)\big]}_{\substack{\text{one-pass variance}\\ \text{excess risk}}}.$$

$$\mathbb{E}[R_M(u_L^{\text{mp}})] = \underbrace{R_M(u^*)}_{\substack{\text{common baseline risk}}} + \underbrace{\mathbb{E}\big[R_M(\theta_L) - R_M(u^*)\big]}_{\substack{\text{common GD-reference}\\ \text{excess risk}}} + \underbrace{\mathbb{E}\big[R_M(u_L^{\text{mp}}) - R_M(\theta_L)\big]}_{\substack{\text{multi-pass fluctuation}\\ \text{excess risk}}}.$$

Note that $R_M(u^*) = \underbrace{R(w^*)}_{\substack{\text{irreducible risk}}} + \underbrace{\big[R_M(u^*) - R(w^*)\big]}_{\substack{\text{approximation risk}}}$. Taking outer expectations over w^* gives the decompositions used in the scaling theorems.

The proof is given at the end of Appendix A. Proposition 3.1 is the organizing principle for the rest of the section: the next theorem specializes the one-pass decomposition, and the theorem after that specializes the multi-pass decomposition.

Theorem 3.1 (One-pass dynamic-batch WSD-SGD scaling law)**.** *Assume Assumptions 1.A, 1.B, 1.C, and 2, together with the one-pass WSD conditions in Section 2. Suppose*

$$b < a + 1, \qquad \sigma^2 \asymp 1, \qquad \frac{\gamma T}{\log T} \gtrsim 1, \qquad 0 < \gamma \le c.$$

Then every deterministic schedule $B_1, \dots, B_T \in \mathbb{N}$ satisfying $\sum_t B_t = N$ obeys, with probability at least $1 - \exp(-\Omega(M))$ over the randomness of S,

$$\mathbb{E}\big[R_M(u_T^{\mathrm{op}})\big] - \sigma^2 = \underbrace{\Theta\Big(M^{1-b} + (\gamma T)^{\frac{1-b}{a}}\Big)}_{\text{Approx+Bias}} + \underbrace{\Theta\left(\frac{\log T}{T B_{\text{eff}}} \min\left\{M, \left(\frac{\gamma T}{\log T}\right)^{\frac{1}{a}}\right\}\right)}_{\text{Var}}.$$

For the static schedule $B_t \equiv B$, where $B \mid N$ and $T = N/B \ge 3$, one has $B_{\text{eff}} = B$.

Interpretation. The irreducible term σ^2 is the label-noise floor. Within the approximation-plus-bias contribution, M^{1-b} measures the information lost by restricting prediction to the M-dimensional sketched model; since $b > 1$, it decreases as the sketch dimension M grows. The term $(\gamma T)^{(1-b)/a}$ is the optimization bias remaining after the full WSD trajectory, and it decreases with the learning horizon γT. Neither deterministic term depends on how a fixed number of updates is divided into batches. The variance term instead reflects stochastic gradient noise. Its spectral factor counts the directions that remain active over the shorter WSD memory horizon $\gamma T / \log T$, while B_{eff} averages the batch sizes according to their final-risk influence. Consequently, larger batches reduce variance most when assigned to influential updates, and two schedules with the same arithmetic mean N/T can have different risks. For the static schedule $B_t \equiv B$, one recovers $B_{\text{eff}} = B$ and the usual horizon–noise tradeoff with $T = N/B$. See Appendix B, and in particular Appendix B.1, for the proof of Theorem 3.1.

We now turn to the multi-pass setting. The key question is how mini-batching affects the decomposed risks, where the next theorem shows that only the fluctuation changes.

Theorem 3.2 (Multi-pass dynamic-batch WSD-SGD scaling law)**.** *Assume Assumptions 1.A, 1.B, 1.C, and 2, together with the multi-pass WSD conditions in Section 2. For a sufficiently small constant $c_1 > 0$,*

$$b < a + 1, \quad \sigma^2 \asymp 1, \quad \frac{\gamma L}{\log L} \gtrsim 1, \quad \gamma \le \frac{c}{\log N}, \quad (\gamma L)^{1/a} \le c_1 M, \quad \gamma L \log N \le c_1 N.$$

Then every schedule $1 \le B_t \le N$ satisfies, with probability at least $1 - \exp(-\Omega(M))$ over S,

$$\mathbb{E}[R_M(u_L^{\mathrm{mp}})] - \sigma^2 = \underbrace{\Theta(M^{1-b})}_{\text{Approx}} + \underbrace{\Theta((\gamma L)^{\frac{1-b}{a}})}_{\text{GD Bias}} + \underbrace{\Theta\left(\frac{(\gamma L)^{\frac{1}{a}}}{N}\right)}_{\text{GD Var}} + \underbrace{\widetilde{\Theta}\left(\rho_{\text{eff}} \frac{\log L}{L} \left(\frac{\gamma L}{\log L}\right)^{\frac{1}{a}}\right)}_{\text{Fluc}}.$$

For the static schedule $B_t \equiv B$, with $1 \le B \le N$, ρ_{eff} is replaced by $\rho_B := \frac{N-B}{B(N-1)}$. Here $\widetilde{\Theta}$ hides only the $\log N$ gap between the matching lower and upper bounds.

Interpretation. The noise floor, approximation error, GD bias, and GD variance have their standard meanings: they respectively represent irreducible label noise, information lost through the finite sketch dimension, incomplete optimization along the full GD trajectory, and estimation error from the finite dataset. The new schedule-dependent term is the fluctuation around that full-batch WSD-GD trajectory. A batch sampled without replacement contributes the finite-population factor $\rho_t = (N - B_t)/(B_t(N-1))$, and ρ_{eff} averages these factors according to the final-risk influence of each update. Thus enlarging an influential batch reduces the fluctuation, while a full batch has $\rho_t = 0$ and introduces none. The remaining factor is governed by the shorter WSD stochastic-memory horizon. Hence, when L and the learning-rate trajectory are fixed, the batch schedule changes only the fluctuation; the approximation, GD bias, and GD variance remain unchanged. See Appendix B.2 for the proof of Theorem 3.2.

Remark 3.2 (Multi-pass sampling with replacement)**.** If each multi-pass mini-batch is sampled with replacement, the same scaling law holds after defining $\rho_t = 1/B_t$ in ρ_{eff}; all other terms are unchanged. See Appendix I.4, in particular Lemma I.11.

3.1 Discussion and Related Works

Empirical evidence. A substantial empirical literature shows that batch size is a central training control. Large-batch ImageNet training motivated the linear learning-rate scaling rule (Goyal et al., 2017); increasing the batch size during training reproduced much of the effect of learning-rate decay while improving parallelism (Smith et al., 2018); and measurements across models and workloads identified a task-dependent critical batch size beyond which additional parallelism yields diminishing returns (McCandlish et al., 2018; Shallue et al., 2019). These studies are primarily empirical and focus on training speed, hardware utilization, or generalization in deep networks. The theory-inspired empirical work of Li et al. (2026) additionally fits joint loss laws to complete learning-rate and batch-size sequences and validates the resulting budget-aware schedules across deep-model settings. Our results address a complementary question: within a tractable statistical model under one pass and data reuse, which iterations' batch sizes determine the final prediction risk?

Theoretical comparisons. Jain et al. (2017) analyze mini-batching and averaging for finite-dimensional least squares, while Mücke et al. (2019) give rates for tail-averaged one-pass kernel SGD under spectral and source conditions. The fast-catch-up phenomenon of Wang et al. (2026) is related to stochastic-noise forgetting and motivates increasing or late-switch schedules for training efficiency. Our contribution is complementary: a discrete finite-horizon risk characterization yielding the explicit γJ versus $\gamma J/\log J$ separation, per-update influence weights, and the one-pass/data-reuse structural distinction.

What batch size changes. In our fixed-horizon decomposition, batch size acts primarily as a noise-control parameter rather than a deterministic regularizer. In the one-pass theorem, batching lowers the centered variance; in the multi-pass theorem, it lowers only the fluctuation around the GD reference path. This interpretation is consistent with the gradient-noise-scale viewpoint of Smith & Le (2018); McCandlish et al. (2018) and with empirical large-batch studies showing that large batches can be effective when properly tuned but that their gains eventually saturate (Goyal et al., 2017; Smith et al., 2018; Shallue et al., 2019). Our theorems make this principle explicit in the present linear-regression setting: once the stochastic terms fall below the approximation and optimization terms, further increasing the influential batch sizes no longer changes the leading statistical scaling. In the multi-pass setting, the finite-population correction decreases as the batch grows. The benefit is therefore greatest when the most influential iterations receive substantial batches.

3.2 Influence-matched batch allocation

The two theorems also yield an oracle allocation rule when the number of updates and the total sample-processing budget are fixed.

Corollary 3.1 (Influence-matched batch allocation)**.** *Assume the respective conditions of Theorems 3.1 and 3.2. In the continuous interior regime, the allocations minimizing the nominal objectives* $\sum_t(\kappa_t^{(T)}/B_t)$ *and* $\sum_t(\rho_t\kappa_t^{(L)})$*, respectively, are*

$$\boxed{B_t^\star = (N\sqrt{\kappa_t^{(T)}})/(\sum_{s=1}^{T}\sqrt{\kappa_s^{(T)}})} \qquad \textit{(one pass, fixed } T\textit{,} \sum_{t=1}^{T} B_t = N\textit{),}$$

and

$$\boxed{B_t^\star = (D_{\text{grad}}\sqrt{\kappa_t^{(L)}})/(\sum_{s=1}^{L}\sqrt{\kappa_s^{(L)}})} \qquad \textit{(multi-pass, fixed } L\textit{,} \sum_{t=1}^{L} B_t = D_{\text{grad}}\textit{).}$$

If a batch-size bound is active, the corresponding square-root allocation is clipped through the usual KKT conditions; budget-preserving integer rounding need not preserve optimality.

Interpretation and related work. In both cases, $B_t^\star \propto \sqrt{\kappa_t}$: more samples are assigned to updates whose noise has greater final-risk influence. This is a fixed-horizon oracle rule based on the population sketched covariance. Optimality is exact for the proxies, not necessarily for the true stochastic errors. Over the same admissible integer schedule class, uniform two-sided bounds make exact proxy minimizers constant-factor optimal for one-pass variance and $O(\log N)$-factor optimal for multi-pass fluctuation. The rule follows from final-risk attribution rather than a direct search for a monotone schedule, so it need not be monotone or single-peaked. The proof is in Appendix A.7.

3.3 Proof sketch

The common idea is to examine each spectral direction and ask two questions: how much signal is learned over the entire trajectory, and how much noise injected at iteration t remains at the final iterate. WSD gives different answers. Its cumulative stepsize controls learning and bias, while the decay phase rapidly forgets earlier noise, leaving a shorter stochastic-memory horizon.

For one-pass WSD-SGD, write the error as its mean plus a centered fluctuation, $e_t = m_t + \delta_t$. The mean follows population GD, so its final filter uses the full WSD trajectory. By contrast, a batch of size B_t injects noise of order $1/B_t$, which is then contracted by all subsequent updates. The quantity $\kappa_t^{(T)}$ measures exactly how much of this time-t noise survives in the final risk. Summing over independent batches therefore produces the schedule dependence B_{eff}. Separate covariance- and label-noise comparisons give matching upper and lower bounds; the WSD decay phase then determines the shorter variance horizon.

For multi-pass WSD-SGD, we instead subtract the full-batch WSD-GD trajectory and study $\Delta_t = u_t^{\mathrm{mp}} - \theta_t$. Conditioned on the dataset, each new batch produces a centered perturbation. Without-replacement sampling scales its covariance by ρ_t, and future updates again damp it according to the safeguarded spectral filter. Comparison with its nominal counterpart then gives the schedule dependence $\sum_{t=1}^{L} \rho_t \kappa_t^{(L)}$. This also explains why a full batch contributes no fluctuation. Covariance comparison and martingale arguments turn this intuition into the two-sided fluctuation estimates, up to the stated logarithmic factor.

Finally, the power-law spectrum and source condition convert these spectral filters into the rates in Theorems 3.1 and 3.2. Thus the proofs separate cleanly into a full-horizon deterministic filter and an influence-weighted stochastic filter with shorter memory.

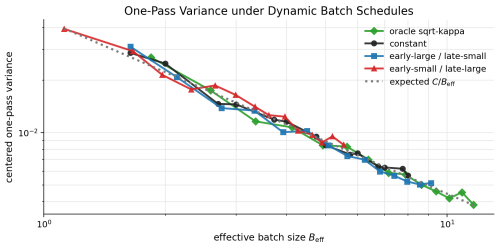
4 Experiments

We test schedule compression through final-risk influence and, secondarily, the allocation consequence in Corollary 3.1. We simulate the sketched Gaussian linear model with $a = 2$, $b = 1.5$, $d = 10^4$, $M = 64$, $N = 512$, $\sigma = 1$, and $\gamma = 0.05$. The sketch and target are fixed, while the dataset and optimization randomness are refreshed over 300 repetitions. The data-generation procedure, schedules, and statistical estimators are detailed in Appendix L.

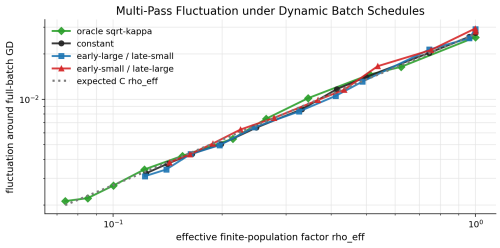
The first three experiments compare four schedules at each sample-processing budget: the oracle square-root allocation $B_t \propto \sqrt{\kappa_t}$, a constant allocation, an early-large/late-small allocation, and an early-small/late-large allocation. All schedules have the same number of updates and total budget. For one-pass WSD-SGD, we fix $T = 64$, vary N, and measure final-iterate variance. For multi-pass WSD-SGD, we fix $N = L = 512$, vary $D_{\mathrm{grad}} = \sum_t B_t$, and measure fluctuation errors.

Figure 1(a) shows that all four one-pass schedule families follow the predicted inverse trend when indexed by B_{eff}. Figure 1(b) gives the corresponding multi-pass collapse: fluctuation is approximately proportional to ρ_{eff}. Panels (a)–(b) therefore support schedule compression: after influence weighting, visibly different batch trajectories become equivalent at the level of leading final-risk scaling.

Experiment 3 compares the schedules at the largest displayed budgets, $N = 512$ for one pass and $D_{\mathrm{grad}} = 4096$ for multi-pass. Relative to the constant schedule, the oracle achieves one-pass variance 0.679 and multi-pass fluctuation 0.656, close to the predicted ratios 0.688 and 0.599. In contrast, the early-small/late-large schedule increases these errors to 1.502 and 1.181, compared with predictions 1.445 and 1.160. The full comparison in Figure 1(c) therefore supports both the influence-weighted laws and the square-root oracle.



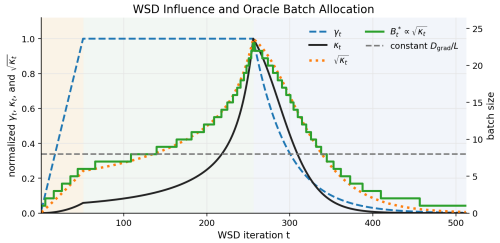
(a) One-pass B_{eff} collapse



(b) Multi-pass ρ_{eff} collapse

Schedule	OP var / const	OP pred / const	MP fluc / const	MP pred / const
oracle sqrt-kappa	0.679	0.688	0.656	0.599
early-large late-small	0.905	0.877	0.958	0.993
constant batch	1.000	1.000	1.000	1.000
early-small late-large	1.502	1.445	1.181	1.160

(c) Relative schedule comparison



(d) WSD influence and oracle allocation

Figure 1: Schedule compression and final-risk influence. Panels (a)–(b) show distinct schedules collapsing when indexed by B_{eff} or ρ_{eff}. Panel (c) compares measured and predicted relative errors. Panel (d) shows that influence and its induced allocation need not track the WSD stepsize.

Experiment 4 shows that batch importance is not determined by learning-rate magnitude alone. Although the stepsize is constant through most of the stable phase, κ_t increases because later perturbations undergo less future contraction. The influence peaks at iteration 256, where the oracle batch also reaches its maximum of 23, compared with the constant batch size 8. During decay, both $\sqrt{\kappa_t}$ and the oracle batch decrease. Thus importance combines the current stochastic perturbation with how much subsequent optimization forgets it; an influence-matched schedule need not track the learning rate or be globally monotone.

5 Conclusion

We developed a finite-horizon statistical theory of dynamic batching under WSD. Deterministic learning uses the full horizon, whereas stochastic noise retains a shorter terminal memory. Final-risk influence compresses arbitrary batch trajectories into scalar risk summaries and reveals a structural difference: batching controls one-pass variance but, under data reuse, only fluctuation around full-batch WSD-GD. The square-root allocation follows as a consequence. The experiments support both schedule compression and the distinction between learning-rate magnitude and final-risk influence.

Natural next steps include estimating the influence weights from data, optimizing the batch schedule jointly with the training horizon, and testing whether the two-horizon principle persists with momentum, adaptive methods, and nonlinear feature-learning models.

AI USE STATEMENT

The authors used ChatGPT and Codex to assist with literature review, manuscript editing, expanding mathematical arguments, and implementing the synthetic experiments. The authors take responsibility for the final manuscript, including all technical claims, proofs, and experimental results.

ETHICS AND REPRODUCIBILITY STATEMENT

Ethics. This work studies a mathematical model of optimization and uses synthetic Gaussian regression data. The reported experiments involve no human participants, personal data, or real-world deployment. The conclusions are restricted to the stated model and should not be interpreted as guarantees of fairness, privacy, or safety in downstream applications. Potential benefits and risks of applying the results beyond this setting are discussed in Appendix M.

Reproducibility. Section 2 specifies the optimization procedures, and Section 3 states the assumptions and main results. Appendix A provides supporting definitions and lemmas; Appendix B organizes the proofs and points to their detailed derivations. Appendix L describes the synthetic data distribution, parameter settings, randomization protocol, batch allocation and rounding, evaluation estimators, and reference-curve fitting used for the four experiments. These specifications support independent replication of the experimental protocol, without claiming bitwise-identical numerical results.

REFERENCES

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22300–22312, 2022.

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 4345–4382, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. In *International Conference on Learning Representations*, 2025.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

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

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11165–11197, 2024.

Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Noah Golmant, Nikita Vemuri, Zhewei Yao, Vladimir Feinberg, Amir Gholami, Kai Rothauge, Michael W. Mahoney, and Joseph Gonzalez. On the computational inefficiency of large batch sizes for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Ben Mann, Alec Radford, Aditya Ramesh, Daniel M. Ziegler, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

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, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jiaxiang Li, Zhiqi Bu, and Shiyun Xu. Towards joint scaling laws with optimal batch size schedules. *arXiv preprint arXiv:2607.27731*, 2026.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Licong Lin, Jingfeng Wu, and Peter L. Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems*, 2025.

Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems*, 2025.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020.

Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Xian Shuai, Yiding Wang, Yimeng Wu, Xin Jiang, and Xiaozhe Ren. Scaling law for language models training considering batch size. *arXiv preprint arXiv:2412.01505*, 2024.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019.

Jinbo Wang, Binghui Li, Zhanpeng Zhou, Mingze Wang, Yuxuan Sun, Jiaqi Zhang, Xunliang Cai, and Lei Wu. Fast catch-up, late switching: Optimal batch size scheduling via functional scaling laws. In *International Conference on Learning Representations*, 2026.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. The benefits of implicit regularization from stochastic gradient descent in least squares problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29773–29785, 2021.

How Dynamic Batch Schedules Shape Final Prediction Risk

Supplementary Material

Table of Contents

A Appendix Preliminaries **15**

A.1 One-pass and multi-pass WSD schedules 15

A.2 Spectral block notation 16

A.3 Full-batch WSD-GD setup 17

A.4 One-pass WSD-SGD setup 17

A.5 Multi-pass WSD-SGD setup 18

A.6 Proof of the main risk decomposition 18

A.7 Influence-matched batch allocation 19

A.8 Power-law interpretation of final-risk influence 20

A.9 Verification of derived multi-pass regularity 22

B Assembling the proofs of the main scaling theorems **23**

B.1 Assembling the proof of the one-pass theorem 24

B.2 Assembling the proof of the multi-pass theorem 24

C Approximation Error **25**

C.1 Upper bound 26

C.2 Lower bound 26

C.3 Bounds under our assumptions 27

D Bias Error for Full-Batch WSD-GD **27**

D.1 Upper and lower bounds 27

D.2 Example under the source condition 29

E Variance Error for Full-Batch WSD-GD **30**

E.1 Upper and lower bounds 30

E.2 Bounds in the main-theorem regime 30

F One-pass WSD-SGD: Excess Error Decomposition **31**

G Bias Error for One-pass WSD-SGD **33**

G.1 Upper and lower bounds for the bias term 34

G.2 Bounds under the source condition 36

H Variance Error for One-pass WSD-SGD **37**

H.1 Time-dependent batch moments and the raw kernel 37

H.2 The additive-noise component 38

H.3 The covariance-fluctuation component 39

H.4 Bounds under the source condition 40

H.5 The WSD influence kernel 41

I Auxiliary with-replacement fluctuation analysis **42**

I.1 Conditional covariance identities 43

I.2 A time-inhomogeneous stochastic approximation lemma 45

I.3 Influence kernels and covariance replacement 46

I.4 Dynamic fluctuation bounds 49

J Fluctuation Error under Multi-pass WSD-SGD without Replacement **51**

K Collected Auxiliary Lemmas **53**

K.1 General concentration lemmas 53

K.2 Power-law auxiliary lemmas 55

L Experimental setup **55**

M Limitations and Broader Effects **57**

A Appendix Preliminaries

We collect the common notation used throughout the appendix. Later appendix sections refer back to this section whenever possible, instead of restating the full stochastic-update setup each time.

Several proof ideas used throughout this appendix are borrowed from Lin et al. (2024; 2025). We adapt those ideas to the WSD and dynamic-batch setting developed in this paper.

A.1 One-pass and multi-pass WSD schedules

For the one-pass procedure only, choose integers

$$1 \le T_{\mathrm{w}} < T_{\mathrm{s}} < T, \qquad T_{\mathrm{d}} := T - T_{\mathrm{s}}, \qquad \tau_{\mathrm{d}}^{\mathrm{op}} := \frac{T_{\mathrm{d}}}{\log T},$$

and use the WSD schedule

$$\gamma_t = \begin{cases} \gamma t / T_{\mathrm{w}}, & 1 \le t \le T_{\mathrm{w}}, \\ \gamma, & T_{\mathrm{w}} < t \le T_{\mathrm{s}}, \\ \gamma \exp(-(t - T_{\mathrm{s}})/\tau_{\mathrm{d}}^{\mathrm{op}}), & T_{\mathrm{s}} < t \le T. \end{cases} \tag{6}$$

Assume

$$T_{\mathrm{s}} - T_{\mathrm{w}} \ge c_{\mathrm{s}} T, \qquad T - T_{\mathrm{s}} \ge c_{\mathrm{d}} T,$$

and define

$$A_T := \sum_{t=1}^{T} \gamma_t, \qquad A_{\mathrm{d}}^{\mathrm{op}} := \gamma \tau_{\mathrm{d}}^{\mathrm{op}}.$$

Lemma A.1 (Mass and product calculus for the one-pass WSD schedule)**.** *Under the preceding phase-length conditions,*

$$A_T \asymp \gamma T, \qquad A_{\mathrm{d}}^{\mathrm{op}} \asymp \frac{\gamma T}{\log T}.$$

If $0 \le \gamma\mu \le c_0 < 1$*, then*

$$\exp(-C\mu A_T) \le \prod_{t=1}^{T} (1 - \gamma_t \mu)^2 \le \exp(-c\mu A_T), \tag{7}$$

where $c, C > 0$ *depend only on* c_0*.*

Proof. Direct summation gives

$$A_T = \frac{\gamma(T_{\mathrm{w}}+1)}{2} + \gamma(T_{\mathrm{s}} - T_{\mathrm{w}}) + \gamma \sum_{r=1}^{T_{\mathrm{d}}} e^{-r/\tau_{\mathrm{d}}^{\mathrm{op}}}.$$

The stable-phase lower bound and $\gamma_t \le \gamma$ imply $A_T \asymp \gamma T$. Since $e^{-T_{\mathrm{d}}/\tau_{\mathrm{d}}^{\mathrm{op}}} = T^{-1}$ and $1 - e^{-1/\tau_{\mathrm{d}}^{\mathrm{op}}} \asymp (\tau_{\mathrm{d}}^{\mathrm{op}})^{-1}$, the final sum is $\asymp \tau_{\mathrm{d}}^{\mathrm{op}}$, proving the mass claims. Finally, summing $-u/(1-c_0) \le \log(1-u) \le -u$ with $u = \gamma_t \mu$ proves equation 7. □

For full-batch WSD-GD and multi-pass WSD-SGD, choose

$$1 \le L_{\mathrm{w}} < L_{\mathrm{s}} < L, \qquad L_{\mathrm{d}} := L - L_{\mathrm{s}}, \qquad \tau_{\mathrm{d}}^{\mathrm{mp}} := \frac{L_{\mathrm{d}}}{\log L},$$

and use the common WSD schedule

$$\gamma_t = \begin{cases} \gamma t / L_{\mathrm{w}}, & 1 \le t \le L_{\mathrm{w}}, \\ \gamma, & L_{\mathrm{w}} < t \le L_{\mathrm{s}}, \\ \gamma \exp(-(t - L_{\mathrm{s}})/\tau_{\mathrm{d}}^{\mathrm{mp}}), & L_{\mathrm{s}} < t \le L. \end{cases} \tag{8}$$

Assume

$$L_{\mathrm{s}} - L_{\mathrm{w}} \ge c_{\mathrm{s}} L, \qquad L - L_{\mathrm{s}} \ge c_{\mathrm{d}} L,$$

and define

$$A_L := \sum_{t=1}^{L} \gamma_t, \qquad A_{\mathrm{d}}^{\mathrm{mp}} := \gamma \tau_{\mathrm{d}}^{\mathrm{mp}}.$$

Lemma A.2 (Mass and product calculus for the multi-pass WSD schedule)**.** *Under the preceding phase-length conditions,*

$$A_L \asymp \gamma L, \qquad A_{\mathrm{d}}^{\mathrm{mp}} \asymp \frac{\gamma L}{\log L}.$$

If $0 \le \gamma\mu \le c_0 < 1$*, then*

$$\exp(-C\mu A_L) \le \prod_{t=1}^{L} (1 - \gamma_t \mu)^2 \le \exp(-c\mu A_L), \tag{9}$$

where $c, C > 0$ *depend only on* c_0.

Proof. The proof is the same direct calculation as for Lemma A.1, with T, T_{w}, and T_{s} replaced by L, L_{w}, and L_{s}. In particular, the stable phase supplies a constant fraction of γL, the decay sum is $\asymp \gamma \tau_{\mathrm{d}}^{\mathrm{mp}}$, and summing the scalar logarithmic bounds proves equation 9. □

Nominal and safeguarded rates for data reuse. Throughout the multi-pass and full-batch analysis, γ_t, A_L, and $A_{\mathrm{d}}^{\mathrm{mp}}$ are nominal quantities. The algorithms use $\eta_t = r\gamma_t$ from equation 4, with actual mass $\widetilde{A}_L = rA_L$. Conditional on the training features, the preceding product calculation applies to η_t, with mass $\widetilde{A}_L$; in particular, for empirical eigenvalues μ,

$$e^{-C\mu r A_L} \le \prod_{t=1}^{L} (1 - \eta_t \mu)^2 \le e^{-c\mu r A_L}.$$

No empirical contraction event is needed, since $\gamma_{\mathrm{safe}} \|\widehat{\Sigma}\|_2 \le 1/8$. Nominal masses remain outside expectations; random masses do not.

A.2 Spectral block notation

For integers $0 \le k_* \le k$ (allowing $k = \infty$), define

$$H_{k_*:k} := \mathrm{diag}(\lambda_{k_*+1}, \ldots, \lambda_k), \qquad w_{k_*:k} := (w_{k_*+1}, \ldots, w_k)^\top.$$

Similarly, let $S_{k_*:k}$ denote the submatrix of S consisting of columns $k_* + 1, \ldots, k$.

A.3 FULL-BATCH WSD-GD SETUP

This subsection records the notation for the full-batch WSD-GD procedure equation 2. Define

$$\Sigma := SHS^\top, \qquad \widehat{\Sigma} := \frac{SX^\top XS^\top}{N}, \qquad u^* := \Sigma^{-1}SHw^*, \qquad \widehat{b} := \frac{1}{N}SX^\top y.$$

Let

$$\widetilde{\varepsilon}_i := y_i - x_i^\top S^\top u^*, \qquad \widetilde{\varepsilon} := (\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_N)^\top, \qquad \widehat{c} := \frac{1}{N}SX^\top \widetilde{\varepsilon}.$$

The full-batch WSD-GD iterate satisfies

$$\theta_t = \theta_{t-1} - \eta_t \widehat{\Sigma}\theta_{t-1} + \eta_t \widehat{b}, \qquad \theta_0 = 0.$$

A.4 ONE-PASS WSD-SGD SETUP

This subsection records the notation for the one-pass WSD-SGD procedure equation 3. Fix positive integers $B_1, \dots, B_T$ satisfying $\sum_{t=1}^T B_t = N$, and define

$$T \geq 3, \qquad \Sigma := SHS^\top, \qquad u^* := \Sigma^{-1}SHw^*.$$

Partition $[N]$ into disjoint batches $I_1, \dots, I_T$ with $|I_t| = B_t$, and write each block as

$$I_t = \{i_{t,1}, \dots, i_{t,B_t}\}.$$

Define

$$\widehat{\Sigma}_t^{\mathrm{op}} := \frac{1}{B_t}\sum_{i \in I_t} Sx_i x_i^\top S^\top, \qquad \widehat{b}_t^{\mathrm{op}} := \frac{1}{B_t}\sum_{i \in I_t} Sx_i y_i,$$

and

$$\widehat{\xi}_t := \frac{1}{B_t}\sum_{i \in I_t} Sx_i\big(y_i - x_i^\top S^\top u^*\big).$$

The one-pass WSD-SGD iterate (u_t^{op}) satisfies

$$u_t^{\mathrm{op}} = u_{t-1}^{\mathrm{op}} - \gamma_t \widehat{\Sigma}_t^{\mathrm{op}} u_{t-1}^{\mathrm{op}} + \gamma_t \widehat{b}_t^{\mathrm{op}}, \qquad u_0^{\mathrm{op}} = 0.$$

With the centered error $e_t := u_t^{\mathrm{op}} - u^*$, this becomes

$$e_t = \big(I - \gamma_t \widehat{\Sigma}_t^{\mathrm{op}}\big)e_{t-1} + \gamma_t \widehat{\xi}_t.$$

For later use, we also write

$$z_{t,b} := Sx_{i_{t,b}}, \qquad \bar{Z}_t := \frac{1}{B_t}\sum_{b=1}^{B_t} z_{t,b}z_{t,b}^\top, \qquad \bar{\xi}_t := \frac{1}{B_t}\sum_{b=1}^{B_t} z_{t,b}\big(y_{i_{t,b}} - z_{t,b}^\top u^*\big),$$

so that equivalently

$$e_t = (I - \gamma_t \bar{Z}_t)e_{t-1} + \gamma_t \bar{\xi}_t.$$

Since the blocks are disjoint subsets of an i.i.d. sample, the pairs $(\bar{Z}_t, \bar{\xi}_t)$ are independent across t conditional on (S, w^*). The conditional mean error $m_t := \mathbb{E}[u_t^{\mathrm{op}} - u^* \mid S, w^*]$ therefore satisfies

$$m_t = (I - \gamma_t \Sigma)m_{t-1}, \qquad m_0 = -u^*.$$

Let $\mu_1, \dots, \mu_M$ denote the eigenvalues of Σ. Define

$$\kappa_t^{\mathrm{op}} := \gamma_t^2 \sum_{j=1}^M \mu_j^2 \prod_{s=t+1}^T (1 - \gamma_s \mu_j)^2 = \gamma_t^2 \left\langle \Sigma, \left[\prod_{s=t+1}^T (I - \gamma_s \Sigma)^2\right]\Sigma \right\rangle$$

and

$$\mathcal{K}_{B_{1:T}} := \sum_{t=1}^T \frac{\kappa_t^{\mathrm{op}}}{B_t}, \qquad \frac{1}{B_{\mathrm{eff}}} := \frac{\mathcal{K}_{B_{1:T}}}{\sum_{t=1}^T \kappa_t^{\mathrm{op}}}.$$

Thus B_{eff} is a weighted harmonic mean of the batch sizes. In particular,

$$B_{\min} := \min_t B_t, \qquad B_{\max} := \max_t B_t, \qquad B_{\min} \leq B_{\mathrm{eff}} \leq B_{\max}, \qquad B_t \equiv B \implies B_{\mathrm{eff}} = B.$$

Throughout the one-pass appendix sections, $(\gamma_t)_{t=1}^T$ denotes the WSD schedule in equation 6; A_T controls its bias filter and $A_{\mathrm{d}}^{\mathrm{op}} = \gamma \tau_{\mathrm{d}}^{\mathrm{op}}$ controls its variance kernel. The multi-pass notation below continues to use the unchanged smooth geometric schedule.

A.5 MULTI-PASS WSD-SGD SETUP

This subsection records the notation for the main-text multi-pass WSD-SGD procedure in equation 5, which samples without replacement within every mini-batch. The main text uses the superscript mp; only in the technical fluctuation proof do we write wor to distinguish this iterate from an auxiliary with-replacement process. Fix a deterministic schedule $B_1, \dots, B_L$ satisfying $1 \le B_t \le N$, and define

$$\Sigma := SHS^\top, \qquad \widehat{\Sigma} := \frac{SX^\top XS^\top}{N}, \qquad u^* := \Sigma^{-1} SHw^*, \qquad \widehat{b} := \frac{1}{N} SX^\top y.$$

Let

$$\widetilde{\varepsilon}_i := y_i - x_i^\top S^\top u^*, \qquad \widetilde{\varepsilon} := (\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_N)^\top, \qquad \widehat{c} := \frac{1}{N} SX^\top \widetilde{\varepsilon}, \qquad \mathcal{D}_{\mathrm{data}} := \sigma(S, w^*, D).$$

At each step $t \in [L]$, sample a subset

$$I_t \subset [N], \qquad |I_t| = B_t,$$

uniformly without replacement from $[N]$, independently across iterations. Define

$$\widehat{\Sigma}_t^{\mathrm{wor}} := \frac{1}{B_t} \sum_{i \in I_t} S x_i x_i^\top S^\top, \qquad \widehat{b}_t^{\mathrm{wor}} := \frac{1}{B_t} \sum_{i \in I_t} S x_i y_i,$$

and

$$\widehat{c}_t^{\mathrm{wor}} := \frac{1}{B_t} \sum_{i \in I_t} S x_i \widetilde{\varepsilon}_i.$$

The multi-pass WSD-SGD iterate without replacement is

$$u_t^{\mathrm{wor}} = u_{t-1}^{\mathrm{wor}} - \eta_t \widehat{\Sigma}_t^{\mathrm{wor}} u_{t-1}^{\mathrm{wor}} + \eta_t \widehat{b}_t^{\mathrm{wor}}, \qquad u_0^{\mathrm{wor}} = 0.$$

Define the multi-pass fluctuation process

$$\Delta_t^{\mathrm{wor}} := u_t^{\mathrm{wor}} - \theta_t, \qquad \Delta_0^{\mathrm{wor}} = 0.$$

Then

$$\Delta_t^{\mathrm{wor}} = \big(I - \eta_t \widehat{\Sigma}_t^{\mathrm{wor}}\big) \Delta_{t-1}^{\mathrm{wor}} + \eta_t \xi_t^{\mathrm{wor}},$$

where

$$\xi_t^{\mathrm{wor}} := -\big(\widehat{\Sigma}_t^{\mathrm{wor}} - \widehat{\Sigma}\big)(\theta_{t-1} - u^*) + \big(\widehat{c}_t^{\mathrm{wor}} - \widehat{c}\big).$$

Define

$$\mathcal{F}_t^{\mathrm{wor}} := \mathcal{D}_{\mathrm{data}} \vee \sigma(I_1, \dots, I_t).$$

Then

$$\mathbb{E}[\xi_t^{\mathrm{wor}} \mid \mathcal{F}_{t-1}^{\mathrm{wor}}] = 0.$$

A.6 PROOF OF THE MAIN RISK DECOMPOSITION

We record here the proof of the structural decomposition used later in Section 3, since it only uses the basic identities from the present section.

Proof of Proposition 3.1. First fix (S, w^*). Unless another conditioning is displayed, $\mathbb{E}$ in this proof denotes expectation conditional on this pair; in particular, Σ and u^* are fixed. Since u^* minimizes the sketched risk R_M, for every $u \in \mathbb{R}^M$ one has

$$R_M(u) = R_M(u^*) + \|\Sigma^{1/2}(u - u^*)\|_2^2.$$

For one-pass WSD-SGD, let $\bar{u}_T = \bar{u}_T(S, w^*) := \mathbb{E}[u_T^{\mathrm{op}} \mid S, w^*]$. Applying the previous display with $u = u_T^{\mathrm{op}}$ and taking expectation gives

$$\mathbb{E}[R_M(u_T^{\mathrm{op}})] = R_M(u^*) + \mathbb{E}\big[\|\Sigma^{1/2}(u_T^{\mathrm{op}} - u^*)\|_2^2\big].$$

Now write

$$u_T^{\mathrm{op}} - u^* = (\bar{u}_T - u^*) + (u_T^{\mathrm{op}} - \bar{u}_T).$$

Since $\mathbb{E}[u_T^{\mathrm{op}} - \bar{u}_T \mid S, w^*] = 0$, the cross term vanishes conditionally on (S, w^*), so

$$\mathbb{E}\big[\|\Sigma^{1/2}(u_T^{\mathrm{op}} - u^*)\|_2^2\big] = \big\|\Sigma^{1/2}(\bar{u}_T - u^*)\big\|_2^2 + \mathbb{E}\big[\|\Sigma^{1/2}(u_T^{\mathrm{op}} - \bar{u}_T)\|_2^2\big].$$

Since u^* minimizes R_M, the first term is exactly

$$R_M(\bar{u}_T) - R_M(u^*).$$

Moreover,

$$R_M(u_T^{\mathrm{op}}) - R_M(\bar{u}_T) = 2\big\langle \Sigma^{1/2}(\bar{u}_T - u^*), \Sigma^{1/2}(u_T^{\mathrm{op}} - \bar{u}_T)\big\rangle + \big\|\Sigma^{1/2}(u_T^{\mathrm{op}} - \bar{u}_T)\big\|_2^2,$$

so taking conditional expectation and using the same centering identity gives

$$\mathbb{E}\big[R_M(u_T^{\mathrm{op}}) - R_M(\bar{u}_T)\big] = \mathbb{E}\big[\|\Sigma^{1/2}(u_T^{\mathrm{op}} - \bar{u}_T)\|_2^2\big].$$

This proves the one-pass decomposition in excess-risk form.

For the main-text multi-pass iterate, identify $u_L^{\mathrm{mp}} = u_L^{\mathrm{wor}}$. Conditional on $\mathcal{D}_{\mathrm{data}}$, both algorithms use the same fixed η_t, so $\mathbb{E}_{\mathrm{batch}}[u_t^{\mathrm{mp}} \mid \mathcal{D}_{\mathrm{data}}] = \theta_t$ by induction. The safeguard depends only on features, not on responses or batch draws. The same identity gives

$$\mathbb{E}[R_M(u_L^{\mathrm{mp}})] = R_M(u^*) + \mathbb{E}\big[\|\Sigma^{1/2}(u_L^{\mathrm{mp}} - u^*)\|_2^2\big].$$

Writing

$$u_L^{\mathrm{mp}} - u^* = (\theta_L - u^*) + (u_L^{\mathrm{mp}} - \theta_L)$$

and expanding the squared norm produces the cross term

$$\mathbb{E}\big[\langle \Sigma^{1/2}(\theta_L - u^*), \Sigma^{1/2}(u_L^{\mathrm{mp}} - \theta_L)\rangle\big].$$

This term vanishes because the fluctuation process has zero conditional mean: the recursion defining $\Delta_t^{\mathrm{wor}} = u_t^{\mathrm{wor}} - \theta_t$ has martingale-difference driving noise with respect to the without-replacement sampling filtration, so

$$\mathbb{E}[u_L^{\mathrm{mp}} - \theta_L \mid \mathcal{D}_{\mathrm{data}}] = 0.$$

Hence

$$\mathbb{E}\big[\|\Sigma^{1/2}(u_L^{\mathrm{mp}} - u^*)\|_2^2\big] = \mathbb{E}\big[\|\Sigma^{1/2}(\theta_L - u^*)\|_2^2\big] + \mathbb{E}\big[\|\Sigma^{1/2}(u_L^{\mathrm{mp}} - \theta_L)\|_2^2\big],$$

where the first term is

$$\mathbb{E}\big[R_M(\theta_L) - R_M(u^*)\big]$$

because u^* minimizes R_M, and the second term is

$$\mathbb{E}\big[R_M(u_L^{\mathrm{mp}}) - R_M(\theta_L)\big]$$

because the corresponding cross term vanishes by the same conditional-centering argument. This proves the multi-pass decomposition in excess-risk form.

Finally, the tower property gives, for example,

$$\mathbb{E}[R_M(u_T^{\mathrm{op}}) \mid S] = \mathbb{E}_{w^*|S}\big[R_M(u^*) + \|\bar{u}_T(S, w^*) - u^*\|_\Sigma^2 + \mathbb{E}[\|u_T^{\mathrm{op}} - \bar{u}_T(S, w^*)\|_\Sigma^2 \mid S, w^*]\big].$$

The multi-pass identity is averaged in the same way. The main theorems retain the sketch conditioning and average over w^*; an additional expectation over S can be taken when an unconditional risk is desired. □

A.7 Proof of the influence-matched batch-allocation corollary

Proof of Corollary 3.1. In the appendix notation, $\kappa_t^{(T)} = \kappa_t^{\mathrm{op}}$ and $\kappa_t^{(L)} = \kappa_t^{\mathrm{mp}}$. We work in the continuous interior regime stated in the corollary. These are the nominal population weights, which are fixed conditional on S and independent of the training sample. We optimize the weighted proxies, not the exact risk including its unknown comparison constants. The safeguard is independent of the batch allocation.

One-pass allocation. For fixed T, all factors in the variance law except B_{eff} are independent of the allocation, and

$$\frac{1}{B_{\text{eff}}} = \frac{\sum_{t=1}^{T} \kappa_t^{(T)}/B_t}{\sum_{t=1}^{T} \kappa_t^{(T)}}.$$

Hence it suffices to minimize $\sum_t \kappa_t^{(T)}/B_t$ subject to $\sum_t B_t = N$. The Lagrange stationarity condition is

$$-\frac{\kappa_t^{(T)}}{B_t^2} + \lambda = 0,$$

so normalization by the budget gives

$$B_t^\star = \frac{N\sqrt{\kappa_t^{(T)}}}{\sum_{s=1}^{T} \sqrt{\kappa_s^{(T)}}}.$$

Multi-pass allocation. For without-replacement sampling,

$$\rho_t = \frac{N - B_t}{B_t(N-1)} = \frac{N}{N-1}\frac{1}{B_t} - \frac{1}{N-1}.$$

Consequently,

$$\sum_{t=1}^{L} \rho_t \kappa_t^{(L)} = \frac{N}{N-1}\sum_{t=1}^{L} \frac{\kappa_t^{(L)}}{B_t} - \frac{1}{N-1}\sum_{t=1}^{L} \kappa_t^{(L)}.$$

The second term is allocation-independent. Applying the same stationarity calculation under $\sum_t B_t = D_{\text{grad}}$ therefore yields

$$B_t^\star = \frac{D_{\text{grad}}\sqrt{\kappa_t^{(L)}}}{\sum_{s=1}^{L} \sqrt{\kappa_s^{(L)}}}.$$

If a lower or upper batch-size constraint is active, the KKT conditions clip the same square-root rule. A budget-preserving rounding then gives an integer schedule; naive coordinatewise rounding need not be exactly optimal. □

A.8 Power-law interpretation of final-risk influence

The exact influence weight retains the spectrum of the sketched covariance. Under the power-law assumption, however, it admits a simpler pointwise form in terms of the learning-rate mass remaining after an update.

Lemma A.3 (Pointwise power-law form of final-risk influence)**.** *Fix a horizon J and define*

$$R_t^{(J)} := \sum_{s=t+1}^{J} \gamma_s, \qquad \kappa_t^{(J)} := \gamma_t^2 \sum_{j=1}^{M} \mu_j^2 \prod_{s=t+1}^{J} (1-\gamma_s\mu_j)^2.$$

Suppose

$$\mu_j \asymp j^{-a}, \qquad a > 1, \qquad \gamma_s\mu_1 \le c_0 < 1 \quad \textit{for every } s \le J.$$

There is a sufficiently small constant $c_\star > 0$, depending only on the power-law and stability constants, such that whenever

$$\left(1 + R_t^{(J)}\right)^{1/a} \le c_\star M,$$

we have

$$\kappa_t^{(J)} \asymp \gamma_t^2 \left(1 + R_t^{(J)}\right)^{-(2-1/a)} = \gamma_t^2 \left(1 + R_t^{(J)}\right)^{1/a-2}. \tag{10}$$

Consequently, the comparison holds uniformly over t if

$$\left(1+\sum_{s=1}^{J}\gamma_s\right)^{1/a} \le c_\star M.$$

Proof. Write $R := R_t^{(J)}$. For $0 \le z \le c_0 < 1$, the scalar logarithmic bounds

$$-\frac{z}{1-c_0} \le \log(1-z) \le -z$$

give constants $c, C > 0$ such that

$$\exp(-C\mu_j R) \le \prod_{s=t+1}^{J}(1-\gamma_s\mu_j)^2 \le \exp(-c\mu_j R).$$

It therefore suffices to show that, for either fixed $c' > 0$,

$$\sum_{j=1}^{M}\mu_j^2 e^{-c'R\mu_j} \asymp (1+R)^{-(2-1/a)}. \tag{11}$$

If $R \le 1$, the upper bound follows from $\sum_j \mu_j^2 \lesssim \sum_j j^{-2a} \lesssim 1$, while the $j = 1$ term gives a constant lower bound. Now suppose $R > 1$ and put $k := \lceil R^{1/a} \rceil$. For the spectral tail,

$$\sum_{j\ge k}\mu_j^2 e^{-c'R\mu_j} \lesssim \sum_{j\ge k} j^{-2a} \lesssim k^{1-2a}.$$

For $j < k$, divide the indices into dyadic blocks

$$k2^{-(\ell+1)} < j \le k2^{-\ell}, \qquad \ell = 0, 1, \dots.$$

Each block contains at most $k2^{-\ell}$ indices, while on that block $R\mu_j \asymp 2^{a\ell}$. Hence its contribution is bounded by

$$CkR^{-2}2^{(2a-1)\ell}e^{-c''2^{a\ell}}.$$

Summing over ℓ gives an upper bound of order

$$kR^{-2} \asymp R^{1/a-2}.$$

For the lower bound, the cutoff condition ensures, after choosing $c_\star$ sufficiently small, that the block $k \le j \le 2k$ lies below M. On this block, $R\mu_j \asymp 1$, so every summand is of order R^{-2}. Since the block contains order k indices,

$$\sum_{j=k}^{2k}\mu_j^2 e^{-c'R\mu_j} \gtrsim kR^{-2} \asymp R^{1/a-2}.$$

This proves equation 11, and multiplication by γ_t^2 proves equation 10. □

Interpretation. The factor γ_t^2 measures the scale of the stochastic perturbation injected at update t. In direction j, the factor μ_j^2 combines the covariance scale of the injected gradient noise with its contribution to prediction risk, while the product over $s > t$ measures how much subsequent optimization contracts that perturbation. Thus $R_t^{(J)}$ acts as a remaining forgetting horizon: directions with $\mu_j R_t^{(J)} \gg 1$ are largely erased, whereas directions with $\mu_j R_t^{(J)} \lesssim 1$ remain visible at the final iterate.

Under $\mu_j \asymp j^{-a}$, the transition occurs near

$$j \asymp \left(1+R_t^{(J)}\right)^{1/a}.$$

There are order $(1 + R_t^{(J)})^{1/a}$ directions near this transition, each carrying squared spectral scale of order $(1 + R_t^{(J)})^{-2}$, which gives equation 10. It also yields the simpler oracle allocation rule

$$B_t^\star \propto \sqrt{\kappa_t^{(J)}} \asymp \gamma_t \big(1 + R_t^{(J)}\big)^{1/(2a)-1}.$$

During the stable phase of WSD, γ_t is constant while $R_t^{(J)}$ decreases, so the influence grows toward the end of that phase. During most of the exponential decay phase, $R_t^{(J)} \asymp \tau_{\mathrm{d}} \gamma_t$, where τ_{d} denotes the WSD decay time constant for the chosen horizon. When $R_t^{(J)} \gtrsim 1$, this gives $\kappa_t^{(J)} \asymp \tau_{\mathrm{d}}^{1/a-2} \gamma_t^{1/a}$, explaining why the influence typically decreases after the stable–decay transition.

A.9 Verification of derived multi-pass regularity

The following lemma discharges the trace, effective-dimension, and maximum-sample-norm conditions that are consequences of the model.

Lemma A.4 (Derived regularity under the Gaussian power-law model)**.** *Assume Assumptions 1.A and 1.C, and let S be the Gaussian sketch specified in Assumption 3. Suppose*

$$0 < \gamma \le \frac{c_\gamma}{\log N}, \qquad \gamma L \log N \le c_L N,$$

where $c_\gamma, c_L > 0$ are sufficiently small constants depending only on a. Then, for all sufficiently large N, with probability at least $1 - \exp(-\Omega(M))$ over S,

$$\operatorname{tr}(\Sigma) \asymp 1, \qquad \operatorname{tr}(\Sigma^2) \lesssim 1, \qquad \gamma \le \frac{c}{\operatorname{tr}(\Sigma)}, \tag{12}$$

and

$$\sum_{i=1}^{M} \frac{\mu_i(\Sigma)}{\mu_i(\Sigma) + 1/(\gamma L)} \le \frac{N}{4}. \tag{13}$$

Moreover, conditioned on every sketch in this event, there is an a-dependent constant $c_{\max} > 0$ such that, for every $u \ge 1$,

$$\mathbb{P}_X\left(8 \max_{i \in [N]} \|S x_i\|_2^2 > \frac{u}{\gamma} \;\middle|\; S \right) \le N^{-c_{\max} u}. \tag{14}$$

In particular, with conditional probability at least $1 - N^{-c_{\max}}$,

$$\gamma \mu_1(\widehat{\Sigma}) \le \gamma \max_{i \in [N]} \|S x_i\|_2^2 \le \frac{1}{8}.$$

The safeguarded rates, in contrast, satisfy $\eta_t C_A \le 1/8$ for every sample, with $C_A := \max_i \|S x_i\|_2^2$. With r from equation 4,

$$\begin{aligned} &\mathbb{P}_X(r < 1 \mid S) \le N^{-c_{\max}}, \qquad \mathbb{P}_X(r < u^{-1} \mid S) \le N^{-c_{\max} u} \quad (u \ge 1), \\ &\mathbb{E}_X[r^{-p} \mid S] \lesssim_p 1 \quad (p > 0). \end{aligned} \tag{15}$$

Proof. On the high-probability sketch event in Lemma K.8,

$$\mu_i(\Sigma) \asymp i^{-a}, \qquad i \in [M].$$

Since $a > 1$,

$$\operatorname{tr}(\Sigma) \asymp \sum_{i=1}^{M} i^{-a} \asymp 1, \qquad \operatorname{tr}(\Sigma^2) \asymp \sum_{i=1}^{M} i^{-2a} \lesssim 1.$$

After reducing c_γ if necessary, the imposed bound $\gamma \le c_\gamma / \log N$ therefore also implies $\gamma \le c / \operatorname{tr}(\Sigma)$, proving equation 12.

For the effective dimension, split the spectrum at

$$k := \left\lceil (\gamma L)^{1/a} \right\rceil.$$

The power law gives

$$\sum_{i=1}^{M} \frac{\mu_i(\Sigma)}{\mu_i(\Sigma) + 1/(\gamma L)} \le k + \gamma L \sum_{i>k} \mu_i(\Sigma) \lesssim 1 + (\gamma L)^{1/a}.$$

The horizon condition yields

$$(\gamma L)^{1/a} \le \left(\frac{c_L N}{\log N}\right)^{1/a}.$$

Because $a > 1$, this is $o(N)$. Choosing c_L sufficiently small and then N sufficiently large proves equation 13.

It remains to establish the maximum-norm tail. Conditioned on S, put $z_i := Sx_i$. Then z_i are independent $\mathcal{N}(0, \Sigma)$ vectors. The Gaussian quadratic-form inequality gives, for every $v \ge 1$,

$$\mathbb{P}_X\left(\|z_i\|_2^2 > \operatorname{tr}(\Sigma) + 2\sqrt{\operatorname{tr}(\Sigma^2)v} + 2\|\Sigma\|_2 v \;\middle|\; S\right) \le e^{-v}.$$

The trace bounds above and $\|\Sigma\|_2 \le \operatorname{tr}(\Sigma) \lesssim 1$ show that the threshold is at most $C(1+v)$. Taking $v = C_1 u \log N$, applying a union bound over $i \in [N]$, and choosing c_γ sufficiently small gives

$$\mathbb{P}_X\left(\max_{i\in[N]} \|z_i\|_2^2 > \frac{u}{8\gamma} \;\middle|\; S\right) \le N^{-c_{\max} u},$$

which proves equation 14. Finally,

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} z_i z_i^\top \preceq \left(\max_{i\in[N]} \|z_i\|_2^2\right) I,$$

so the stated nominal empirical contraction follows at threshold $u = 1$. For the safeguard ratio, $r^{-1} = \max\{1, 8\gamma C_A\}$. Thus the same tail bound proves the first two claims in equation 15, and

$$\mathbb{E}_X[r^{-p} \mid S] = 1 + p\int_1^\infty u^{p-1}\mathbb{P}_X(r^{-1} > u \mid S)\, du \lesssim_p 1.$$

This controls even the rare samples on which the safeguard is active. □

Lemma A.5 (Comparison with nominal influence weights)**.** *On the power-law sketch event, suppose $(1 + A_L)^{1/a} \le c_\star M$ and nominal population stability holds. Let κ_t use γ_t, and let $\widetilde{\kappa}_t$ use $\eta_t = r\gamma_t$ in the same population filter. Uniformly over t and $0 < r \le 1$,*

$$r^2\kappa_t \le \widetilde{\kappa}_t \lesssim r^{1/a}\kappa_t \le \kappa_t.$$

Under equation 15, for any nonnegative deterministic weights a_t, conditional on S,

$$\mathbb{E}_X \sum_t a_t \widetilde{\kappa}_t \asymp \sum_t a_t \kappa_t.$$

Proof. Decreasing every rate increases each nonnegative contraction factor, giving the lower bound. Put $R_t = \sum_{s>t} \gamma_s$ and $p_a = 2 - 1/a$. Lemma A.3 gives

$$\frac{\widetilde{\kappa}_t}{\kappa_t} \lesssim r^2\left(\frac{1+R_t}{1+rR_t}\right)^{p_a} \le r^{2-p_a} = r^{1/a}.$$

The unsaturated condition for rR_t follows from that for A_L. For the expectation lower bound restrict to the event $r = 1$, whose probability is bounded below by equation 15. The upper bound holds pathwise, so both comparisons survive arbitrary nonnegative weighting and summation. This lemma compares population filters only; empirical covariance replacement is treated separately below. □

B Assembling the proofs of the main scaling theorems

This section gathers the appendix ingredients used in the proofs of Theorems 3.1 and 3.2. The detailed derivations are carried out in the subsequent appendix sections; here we simply record how those results fit together. Throughout, we work on the intersection of the high-probability events from the cited lemmas. Since only finitely many such results are invoked, a union bound still yields probability

$$1 - \exp(-\Omega(M)).$$

B.1 Assembling the proof of Theorem 3.1

Start from Proposition 3.1, with the following identity and $\bar{u}_T = \bar{u}_T(S, w^*)$ conditional on (S, w^*):

$$\mathbb{E}[R_M(u_T^{\mathrm{op}})] = R_M(u^*) + \big(R_M(\bar{u}_T) - R_M(u^*)\big) + \mathbb{E}\big[R_M(u_T^{\mathrm{op}}) - R_M(\bar{u}_T)\big].$$

We then take the outer expectation over w^*, retaining S fixed on its high-probability event. The three terms are supplied by the appendix as follows.

(1) Common baseline risk. Appendix C defines

$$\mathrm{Approx} := R_M(u^*) - R(w^*).$$

Under Assumption 1.B, $R(w^*) = \sigma^2$, and therefore

$$R_M(u^*) = \sigma^2 + \mathrm{Approx}.$$

Lemma C.3 then gives

$$\mathbb{E}_{w^*}[\mathrm{Approx}] \asymp M^{1-b}.$$

(2) One-pass bias term. In Appendix F, the mean error satisfies $m_T = \bar{u}_T - u^*$, so Definition F.1 identifies

$$R_M(\bar{u}_T) - R_M(u^*) = \|\Sigma^{1/2}(\bar{u}_T - u^*)\|_2^2 = \mathrm{Bias}_T.$$

Lemma G.3 yields the general upper bound

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \lesssim \min\big\{M, A_T^{1/a}\big\}^{1-b}.$$

Moreover, when $A_T^{1/a} \le M/c_1$, the same lemma gives the matching lower bound

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim A_T^{(1-b)/a}.$$

(3) One-pass variance term. Proposition F.1 and Proposition F.2 identify the remaining term with the one-pass variance quantity $\mathrm{Var}_{B_{1:T}}$. Lemma H.3 then yields

$$\mathbb{E}_{w^*}[\mathrm{Var}_{B_{1:T}}] \asymp \frac{\min\{M, (\gamma\tau_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}} B_{\mathrm{eff}}}.$$

Here B_{eff} is the influence-weighted harmonic mean. It preserves the exact $1/B_t$ contribution of each update. For a constant schedule $B_t \equiv B$, one has $B_{\mathrm{eff}} = B$ and $T = N/B$.

From items **(1)** and **(2)**, the deterministic contribution obeys

$$\mathbb{E}_{w^*}[\mathrm{Approx} + \mathrm{Bias}_T] = \Theta\Big(M^{1-b} + A_T^{(1-b)/a}\Big).$$

Indeed, when $A_T^{1/a} \le M/c_1$, the lower bound on Bias_T gives the second scale directly; when $A_T^{1/a} > M/c_1$, one has $A_T^{(1-b)/a} \lesssim M^{1-b}$, so the approximation lower bound already controls that term. Substituting this deterministic bound and item **(3)** into the risk decomposition gives the dynamic-schedule display in Theorem 3.1. Substituting $A_T \asymp \gamma T$, $B_{\mathrm{eff}} = B$, and $T = N/B$ gives its static-batch display.

B.2 Assembling the proof of Theorem 3.2

The multi-pass iterate and the GD reference share η_t conditional on the features. All masses and influence weights in the bounds below are nominal; their conversion from the safeguarded filters is included in the GD and fluctuation lemmas. Start from Proposition 3.1, initially conditional on (S, w^*):

$$\mathbb{E}[R_M(u_L^{\mathrm{mp}})] = R_M(u^*) + \mathbb{E}\big[R_M(\theta_L) - R_M(u^*)\big] + \mathbb{E}\big[R_M(u_L^{\mathrm{mp}}) - R_M(\theta_L)\big].$$

Taking the outer target expectation, the three contributions are assembled as follows.

(1) Common baseline risk. As in the one-pass proof above,

$$R_M(u^*) = \sigma^2 + \text{Approx}, \qquad \mathbb{E}_{w^*}[\text{Approx}] \asymp M^{1-b}$$

by Assumption 1.B and Lemma C.3.

(2) Common GD-reference contribution. Sections D and E control the deterministic and stochastic pieces of the full-batch WSD-GD reference iterate θ_L. Their source-condition conclusions are Lemmas D.3 and E.2. The main-theorem condition $\gamma L \log N \le c_1 N$ implies the horizon restriction in the variance lemma. Since the multi-pass WSD mass satisfies $A_L \asymp \gamma L$, the theorem's explicit condition $(\gamma L)^{1/a} \le c_1 M$, with c_1 chosen sufficiently small, implies the appendix unsaturated condition $A_L^{1/a} \lesssim M$. The two lemmas therefore give

$$\mathbb{E}_{w^*}[\text{Bias}_{\text{GD}}(w^*)] \asymp A_L^{(1-b)/a},$$

and

$$\text{Var}_{\text{GD}} \asymp \frac{\min\{M, A_L^{1/a}\}}{N}.$$

(3) Multi-pass fluctuation term. Write

$$\mathcal{K}_{\text{d}} := \frac{\min\{M, (A_{\text{d}}^{\text{mp}})^{1/a}\}}{\tau_{\text{d}}^{\text{mp}}}.$$

Corollary J.1, whose proof uses the auxiliary with-replacement analysis in Lemma I.11, gives

$$\rho_{\text{eff,WSD}}^{\text{wor}} \mathcal{K}_{\text{d}} \lesssim \mathbb{E}[\text{Fluc}_{B_{1:L}}^{\text{wor}}] \lesssim \rho_{\text{eff,WSD}}^{\text{wor}} \log N \mathcal{K}_{\text{d}}.$$

Thus WSD has two distinct horizons: $A_L \asymp \gamma L$ controls GD learning in the nominal-scale expectation bounds, whereas $A_{\text{d}}^{\text{mp}} \asymp \gamma L / \log L$ controls the final stochastic-memory kernel. Combining the baseline term, the two GD-reference orders, and the fluctuation estimate gives the claimed four-part decomposition and the sharp fluctuation sandwich. For a static schedule, substituting $\rho_B = (N-B)/(B(N-1))$ gives the second display in the theorem. In the main-text notation, $u_L^{\text{mp}} = u_L^{\text{wor}}$ and $\rho_{\text{eff}} = \rho_{\text{eff,WSD}}^{\text{wor}}$.

C Approximation Error

The approximation analysis follows the sketched linear-regression arguments of Lin et al. (2024; 2025). We provide a short proof for completeness only; this term is independent of the learning-rate schedule. This section studies the approximation term, which is common to the three optimization procedures equation 2, equation 3, and equation 5.

Retaining the common setup from Section 2, define

$$\text{Approx} := \min_{u \in \mathbb{R}^M} R_M(u) - \min_{w \in \mathcal{H}} R(w) = R_M(u^*) - R(w^*)$$

which further yields, by substituting the value of u^*:

$$\text{Approx} = \left\| (I - H^{1/2} S^\top \Sigma^{-1} S H^{1/2}) H^{1/2} w^* \right\|_2^2. \tag{16}$$

Assumption 3 (Gaussian sketching)**.** *The sketching operator $S : \mathcal{H} \to \mathbb{R}^M$ is Gaussian, meaning that in the diagonal coordinates of H, its entries are i.i.d. distributed as $\mathcal{N}(0, 1/M)$.*

Assumption 4 (Source-condition regime)**.** *In addition to Assumptions 1.C and 2, we work in the regime*

$$1 < b < a + 1.$$

C.1 Upper bound

Lemma C.1 (Upper bound on the approximation error)**.** *Fix any integer $k \geq 0$ such that $r(H) \geq k + M$, and define*

$$A_k := S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top.$$

Then the approximation error satisfies the deterministic bound

$$\text{Approx} \lesssim \|w^*_{k:\infty}\|^2_{H_{k:\infty}} + w^{*\top}_{0:k}\big(H^{-1}_{0:k} + S^\top_{0:k} A_k^{-1} S_{0:k}\big)^{-1} w^*_{0:k}.$$

If in addition Assumption 3 holds and $k \leq M/2$, then with probability at least

$$1 - \exp(-\Omega(M))$$

over the randomness of S,

$$\text{Approx} \lesssim \|w^*_{k:\infty}\|^2_{H_{k:\infty}} + \beta_k \|w^*_{0:k}\|_2^2,$$

where

$$\beta_k := \frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}}.$$

Proof. Let $P := H^{1/2} S^\top \Sigma^{-1} S H^{1/2}$, an orthogonal projector. Splitting $H^{1/2} w^*$ into its head and tail and using $(I-P)^2 = I - P$ gives

$$\text{Approx} \leq 2 w^{*\top}_{0:k} H^{1/2}_{0:k} (I-P)_{0:k,0:k} H^{1/2}_{0:k} w^*_{0:k} + 2\|w^*_{k:\infty}\|^2_{H_{k:\infty}}.$$

The Woodbury identity yields

$$H^{1/2}_{0:k}(I-P)_{0:k,0:k} H^{1/2}_{0:k} = \big(H^{-1}_{0:k} + S^\top_{0:k} A_k^{-1} S_{0:k}\big)^{-1},$$

proving the deterministic bound. For Gaussian sketching, Lemma K.5 and Gaussian singular-value concentration give $\|A_k\|_2 \lesssim \beta_k$ and $S^\top_{0:k} S_{0:k} \succeq c I_k$ with the stated probability. Thus the inverse in the last display is bounded by $C\beta_k I_k$, which proves the probabilistic bound. For $k = 0$, the head term is absent. □

C.2 Lower bound

Lemma C.2 (Lower bound on the approximation error under the source condition)**.** *Assume Assumptions 1.C and 2, and let*

$$H_w := \mathbb{E}[w^* w^{*\top}].$$

Then, conditioned on the sketch matrix S,

$$\mathbb{E}_{w^*}[\text{Approx}] \gtrsim \sum_{i>M} \lambda_i i^{a-b}.$$

In particular, under Assumptions 1.C and 2,

$$\mathbb{E}_{w^*}[\text{Approx}] \gtrsim M^{1-b}.$$

Proof. With the projector P above and $Q := H^{1/2} H_w H^{1/2}$, the source condition gives

$$\mathbb{E}_{w^*}[\text{Approx}] = \text{tr}((I-P)Q) \geq \sum_{i>M} \mu_i(Q).$$

Indeed, P has rank at most M, so the trace captured by P is at most the sum of the largest M eigenvalues of Q. The diagonal entries of Q are comparable to i^{-b}; their decreasing rearrangement has the same order. Hence $\sum_{i>M} \mu_i(Q) \gtrsim \sum_{i>M} i^{-b} \asymp M^{1-b}$. Under the power law, $i^{-b} \asymp \lambda_i i^{a-b}$, giving the displayed form as well. □

C.3 Bounds under our assumptions

Lemma C.3 (Approximation error under our assumptions)**.** *Assume Assumptions 2, 3, and 4. Then with probability at least*

$$1 - \exp(-\Omega(M))$$

over the randomness of S,

$$\mathbb{E}_{w^*}[\mathrm{Approx}] \asymp M^{1-b}.$$

Proof. Choose $k = \lfloor M/2 \rfloor$. The power law and source condition imply

$$\beta_k \lesssim M^{-a}, \qquad \mathbb{E}_{w^*}\|w^*_{0:k}\|_2^2 \lesssim M^{a-b+1}, \qquad \mathbb{E}_{w^*}\|w^*_{k:\infty}\|^2_{H_{k:\infty}} \asymp M^{1-b},$$

where $a - b > -1$ is used in the head sum. Lemma C.1 therefore gives the upper bound CM^{1-b}, while Lemma C.2 gives the matching lower bound. Bounded sketch dimensions are absorbed into the constants. □

D Bias Error for Full-Batch WSD-GD

The baseline bias analysis follows the linear-regression framework of Lin et al. (2024; 2025), with the full-batch argument adapted from the latter work. We give short proofs for completeness only. The WSD adaptation uses the cumulative stepsize $A_L \asymp \gamma L$ and Lemma A.2. Empirical contraction is guaranteed by equation 4; A_L always denotes the nominal mass. We use the full-batch WSD-GD procedure equation 2 and the notation of Section 2:

$$\theta_t = \theta_{t-1} - \eta_t \widehat{\Sigma}\theta_{t-1} + \eta_t \widehat{b}, \qquad \theta_0 = 0,$$

with

$$\Sigma := SHS^\top, \qquad \widehat{\Sigma} := \frac{1}{N} SX^\top X S^\top, \qquad u^* := \Sigma^{-1} SHw^*, \qquad \widehat{b} := \frac{1}{N} SX^\top y.$$

Define

$$\widetilde{\varepsilon}_i := y_i - x_i^\top S^\top u^*, \qquad \widetilde{\varepsilon} := (\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_N)^\top, \qquad \widehat{c} := \frac{1}{N} SX^\top \widetilde{\varepsilon},$$

and introduce the shorthand

$$C_L := \prod_{t=1}^{L} (I - \eta_t \widehat{\Sigma}), \qquad V(\widehat{\Sigma}) := \frac{1}{N} \sum_{t=1}^{L} \eta_t \prod_{i=t+1}^{L} (I - \eta_i \widehat{\Sigma}).$$

Then

$$\theta_L - u^* = -C_L u^* + V(\widehat{\Sigma}) SX^\top \widetilde{\varepsilon}.$$

Accordingly, the GD bias term is

$$\mathrm{Bias}_{\mathrm{GD}}(w^*) := \mathbb{E}_X \left[\|\Sigma^{1/2} C_L u^*\|_2^2 \right].$$

D.1 Upper and lower bounds

Lemma D.1 (Upper bound on the WSD-GD bias term)**.** *Assume Assumptions 1.A, 1.C, and 3, use the safeguarded algorithm, and work on the derived regularity event in Lemma A.4. Suppose*

$$A_L \lesssim N^a.$$

Fix any integer $k \le M/3$ such that $\mathrm{rank}(H) \ge k + M$*, and define*

$$A_k := S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top, \qquad \widetilde{k} := \lceil N/2 \rceil.$$

Writing $\Sigma = \sum_{j=1}^{M} \mu_j q_j q_j^\top$, define its spectral tail

$$\Sigma_{>\widetilde{k}} := \sum_{j=\widetilde{k}+1}^{M} \mu_j q_j q_j^\top,$$

with the convention $\Sigma_{>\widetilde{k}} = 0$ *when* $\widetilde{k} \geq M$*. Then with probability at least*

$$1 - \exp(-\Omega(M))$$

over the randomness of S*,*

$$\mathrm{Bias}_{\mathrm{GD}}(w^*) \lesssim \frac{\|w^*_{0:k}\|_2^2}{A_L} \left(\frac{\mu_{M/2}(A_k)}{\mu_M(A_k)}\right)^2 + \overline{B}\,\|w^*_{k:\infty}\|^2_{H_{k:\infty}},$$

where

$$\overline{B} := 1 + (A_L^2 + 1)\left(\frac{\mathrm{tr}(\Sigma_{>\widetilde{k}})^2}{N^2} + \|\Sigma_{>\widetilde{k}}\|_2^2 + \frac{\mathrm{tr}(\Sigma^2_{>\widetilde{k}})}{N} + \sqrt{\frac{\mathrm{tr}(\Sigma^4_{>\widetilde{k}})}{N}}\right).$$

Proof. We use the head–tail argument of Lin et al. (2025), with the WSD mass A_L. Put $M_L = C_L \Sigma C_L$ and $\mathcal{B} = \mathbb{E}_X[\Sigma^{-1/2} M_L \Sigma^{-1/2}]$. Splitting SHw^* at k, and using $S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top \preceq \Sigma$, gives

$$\mathrm{Bias}_{\mathrm{GD}}(w^*) \lesssim \mathbb{E}_X \|M_L\|_2 \, \|\Sigma^{-1} S_{0:k} H_{0:k}\|_2^2 \|w^*_{0:k}\|_2^2 + \|\mathcal{B}\|_2 \|w^*_{k:\infty}\|^2_{H_{k:\infty}}.$$

For the head term, the safeguard and Lemma A.2 imply $\|C_L^2 \widehat{\Sigma}\|_2 \lesssim (rA_L)^{-1}$. Inserting $(\widehat{\Sigma} + \lambda I)^{1/2}$ with the deterministic $\lambda \asymp A_L^{-1}$ gives $\|M_L\|_2 \lesssim A_L^{-1} r^{-1} F_\lambda$, where $F_\lambda = \|(\widehat{\Sigma} + \lambda I)^{-1/2}(\Sigma + \lambda I)^{1/2}\|_2^2$. Cauchy–Schwarz, equation 15, and Lemma K.1 yield

$$\mathbb{E}_X \|M_L\|_2 \lesssim A_L^{-1}.$$

Lemma K.7 bounds the remaining head norm by $C\mu_{M/2}(A_k)/\mu_M(A_k)$.

For completeness, the tail estimate follows from the same Gaussian resolvent calculation. Write $\widehat{\Sigma} = \Sigma^{1/2} G^\top G \Sigma^{1/2}$, where $G_{ij} \sim \mathcal{N}(0, 1/N)$, and put $K = G\Sigma G^\top$. The residual polynomial

$$R(K) := \sum_{t=1}^{L} \eta_t \prod_{s<t} (I - \eta_s K)$$

satisfies $\|R(K)\|_2 \leq rA_L \leq A_L$ and $\|R(K)K\|_2 \leq 1$ by empirical contraction. Telescoping gives

$$\mathcal{B} \preceq 2I + 2\mathbb{E}_G Q_G, \qquad Q_G := G^\top R(K) G \Sigma^2 G^\top R(K) G.$$

Split the columns of G after $\min\{M, \lceil N/2 \rceil\}$, and let E be the corresponding tail covariance. On the Gaussian head singular-value event, with $K_{\mathrm{t}} = G_{\mathrm{t}} E G_{\mathrm{t}}^\top$,

$$Q_G \preceq C\big(1 + A_L^2 \|K_{\mathrm{t}}\|_2^2 + A_L^2 \|G_{\mathrm{t}} E^2 G_{\mathrm{t}}^\top\|_2\big) G^\top G.$$

Gaussian fourth-moment bounds and Cauchy–Schwarz, applied to $v^\top \mathbb{E}[ZG^\top G] v = \mathbb{E}[Z\|Gv\|_2^2]$, bound its expectation by

$$C\left[1 + A_L^2\left(\frac{\mathrm{tr}(E)^2}{N^2} + \|E\|_2^2 + \frac{\mathrm{tr}(E^2)}{N} + \sqrt{\frac{\mathrm{tr}(E^4)}{N}}\right)\right] \lesssim \overline{B}.$$

The complement of the head singular-value event is absorbed using its exponentially small probability, Gaussian moments, and $A_L \lesssim N^a$. When the tail is empty the bound is constant. Substituting these head and tail estimates proves the claim. □

Lemma D.2 (Lower bound on the WSD-GD bias term)**.** *Assume Assumptions 1.A and 3, and use the safeguarded algorithm. Let*

$$H_w := \mathbb{E}_{w^*}[w^* w^{*\top}], \qquad \Sigma_w := SHH_wHS^\top.$$

Then with probability at least

$$1 - \exp(-\Omega(M))$$

over the randomness of S*,*

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}(w^*)] \gtrsim \sum_{i=\lceil 2\tau \rceil + 1}^{\lfloor M/3 \rfloor} \frac{\mu_{3i}(\Sigma_w)}{\mu_i(\Sigma)},$$

where

$$\tau := \mathbb{E}_X\Big[\#\big\{i \in [M] : \mu_i(\widehat{\Sigma}) A_L > 1/4\big\}\Big].$$

The sum is understood to be zero when its lower endpoint exceeds $\lfloor M/3 \rfloor$*.*

Proof. Set $A = \mathbb{E}_X[C_L]$ and $B = \Sigma^{-1/2}\Sigma_w\Sigma^{-1/2}$. Empirical contraction makes C_L and A PSD. Gaussian sign invariance in an eigenbasis of Σ makes A diagonal in that basis: changing coordinate signs preserves every $\|Sx_i\|_2$, hence preserves r, and conjugates C_L by the same sign matrix. Thus $A\Sigma = \Sigma A$, even though the rates are sample-dependent. The PSD covariance identity then gives

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}(w^*)] \geq \mathrm{tr}(A^2B).$$

Let P_X project onto the empirical directions satisfying $A_L\mu_j(\widehat{\Sigma}) > 1/4$, using the nominal mass. Since $rA_L \leq A_L$, the WSD product bound implies $C_L \succeq c(I - P_X)$. Since $\mathrm{tr}(\mathbb{E}_X P_X) = \tau$, at most $\lceil 2\tau \rceil$ eigenvalues of $\mathbb{E}_X P_X$ exceed $1/2$. Thus A has at least $M - \lceil 2\tau \rceil$ eigenvalues bounded below by $c/2$. The lower rearrangement form of von Neumann's trace inequality therefore yields

$$\mathrm{tr}(A^2B) \geq \sum_{i=1}^{M} \mu_{M-i+1}(A)^2\mu_i(B) \gtrsim \sum_{i=\lceil 2\tau \rceil+1}^{\lfloor M/3 \rfloor} \mu_i(B).$$

Finally, apply the singular-value product inequality twice to $\Sigma_w = \Sigma^{1/2}B\Sigma^{1/2}$:

$$\mu_{3i-2}(\Sigma_w) \leq \mu_i(\Sigma)\mu_i(B).$$

Monotonicity gives $\mu_i(B) \geq \mu_{3i}(\Sigma_w)/\mu_i(\Sigma)$, which proves the result. An empty index set gives the stated zero bound. □

D.2 Example under the source condition

Lemma D.3 (Bias bounds under the source condition)**.** *Assume Assumptions 1.A, 1.C, 2, and 3, use the safeguarded algorithm and the derived regularity conclusions of Lemma A.4. Suppose*

$$a > b - 1, \qquad A_L \gtrsim 1, \qquad A_L \lesssim N^a.$$

Then there exists an (a, b)-dependent constant $c > 0$ such that, whenever

$$\gamma \leq \frac{c}{\log N},$$

we have with probability at least

$$1 - \exp(-\Omega(M))$$

over the randomness of S,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}(w^*)] \lesssim \min\{M, A_L^{1/a}\}^{1-b}.$$

Moreover, there exists an (a, b)-dependent constant $c_0 > 0$ such that, in the unsaturated regime

$$A_L^{1/a} \leq c_0 M,$$

the matching lower bound holds:

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}(w^*)] \asymp A_L^{(1-b)/a}.$$

Proof. On the spectral events in Lemmas K.8 and K.9, $\mu_i(\Sigma) \asymp i^{-a}$ and $\mu_{M/2}(A_k)/\mu_M(A_k) \lesssim 1$. The tail bracket defining $\overline{B}$ is $O(N^{-2a})$, so $A_L \lesssim N^a$ implies $\overline{B} \lesssim 1$. Choose an integer $k \asymp \min\{M, A_L^{1/a}\}$ with $k \leq M/3$. Lemma D.1 and the source condition give

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}] \lesssim A_L^{-1} \sum_{i \leq k} i^{a-b} + \sum_{i > k} i^{-b} \lesssim k^{1-b},$$

using $a - b > -1$ and $k^a \lesssim A_L$.

For the lower bound, Lemma K.11 gives $\tau \lesssim A_L^{1/a}$; its exponentially small exceptional event contributes at most $\min\{M, N\}\exp(-cN)$. Applying Lemma K.8 to HH_wH also gives $\mu_i(\Sigma_w) \asymp i^{-a-b}$. Therefore Lemma D.2 implies, when $A_L^{1/a} \leq c_0 M$,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_{\mathrm{GD}}] \gtrsim \sum_{CA_L^{1/a} \leq i \leq M/3} i^{-b} \gtrsim A_L^{(1-b)/a}.$$

Choosing c_0 sufficiently small leaves an interval of the required length. Outside this unsaturated regime only the upper bound is claimed. □

E VARIANCE ERROR FOR FULL-BATCH WSD-GD

The baseline variance analysis follows the linear-regression framework of Lin et al. (2024; 2025), with the full-batch argument adapted from the latter work. We provide short proofs for completeness only, recording the WSD filter comparison explicitly. As in the bias analysis, the shared safeguard supplies empirical contraction, and A_L is nominal.

Retain $V(\widehat{\Sigma})$ from Section D, and define

$$v_L(x) := 1 - \prod_{t=1}^{L}(1-\eta_t x), \qquad h_L(x) := \begin{cases} v_L(x)^2/x, & x > 0, \\ 0, & x = 0. \end{cases}$$

The normalized variance is

$$\mathrm{Var}_{\mathrm{GD}} := \mathbb{E}_X \operatorname{tr}\big(XS^\top V(\widehat{\Sigma})\Sigma V(\widehat{\Sigma})SX^\top\big) = \frac{1}{N}\mathbb{E}_X \operatorname{tr}\big(\Sigma h_L(\widehat{\Sigma})\big).$$

The zero-eigenvalue convention avoids assuming that $\widehat{\Sigma}$ is invertible when $M > N$.

E.1 UPPER AND LOWER BOUNDS

Lemma E.1 (Spectral bounds on the WSD-GD variance)**.** *Assume Assumptions 1.A, 1.C, and 3, use the safeguarded algorithm and the derived regularity conclusions of Lemma A.4. Suppose* $1 \lesssim A_L \lesssim N^a$. *Write* $\widehat{\Sigma} = \sum_j \widehat{\mu}_j \widehat{q}_j \widehat{q}_j^\top$, *and set*

$$g_{rA}(x) := \begin{cases} \min\{(rA_L)^2 x, x^{-1}\}, & x > 0, \\ 0, & x = 0. \end{cases}$$

Then

$$\mathrm{Var}_{\mathrm{GD}} \asymp \frac{1}{N}\mathbb{E}_X \sum_j (\widehat{q}_j^\top \Sigma \widehat{q}_j) g_{rA}(\widehat{\mu}_j) \lesssim \frac{D_U}{N}, \qquad D_U := \mathbb{E}_X \sum_j \min\{1, A_L\widehat{\mu}_j\}.$$

Proof. Empirical contraction and Lemma A.2 give $v_L(x) \asymp \min\{1, rA_L x\}$ throughout the empirical spectrum. Thus $h_L(x) \asymp g_{rA}(x)$, proving the two-sided identity in the empirical eigenbasis, with r kept inside the expectation. For the upper bound use $r \le 1$ and choose the deterministic $\lambda \asymp A_L^{-1}$. Inserting $(\widehat{\Sigma} + \lambda I)^{1/2}$ gives

$$\operatorname{tr}(\Sigma h_L(\widehat{\Sigma})) \le F_\lambda \sum_j (\widehat{\mu}_j + \lambda) h_L(\widehat{\mu}_j) \lesssim F_\lambda \sum_j \min\{1, A_L\widehat{\mu}_j\},$$

where $F_\lambda = \|\Sigma^{1/2}(\widehat{\Sigma} + \lambda I)^{-1/2}\|_2^2$. Lemma K.1 bounds F_λ by a constant outside an exponentially small event. On its complement, use $F_\lambda \le \|\Sigma\|_2/\lambda$, the rank bound $\sum_j \min\{1, A_L\widehat{\mu}_j\} \le N$, and $A_L \lesssim N^a$. This remainder is absorbed into D_U, which is bounded below by a positive constant under $A_L \gtrsim 1$ and the Gaussian power-law model. Taking expectation proves the claim. □

E.2 BOUNDS IN THE MAIN-THEOREM REGIME

Lemma E.2 (WSD-GD variance under power-law decay)**.** *Assume the conditions of Lemma E.1 and the main-theorem horizon restriction* $A_L \log N \le cN$, *for a sufficiently small constant* $c > 0$. *Then, with probability at least* $1 - \exp(-\Omega(M))$ *over* S,

$$\mathrm{Var}_{\mathrm{GD}} \asymp \frac{1}{N}\sum_{j=1}^{M} \min\{1, (A_L\mu_j)^2\} \asymp \frac{\min\{M, A_L^{1/a}\}}{N}.$$

Proof. Let $k_L = \min\{M, A_L^{1/a}\}$. On the empirical power-law event of Lemma K.11, splitting at $A_L^{1/a}$ gives

$$\sum_j \min\{1, A_L\widehat{\mu}_j\} \lesssim \min\left\{M,\ A_L^{1/a} + A_L \sum_{j > A_L^{1/a}} j^{-a}\right\} \lesssim k_L.$$

Its complement contributes at most $N\exp(-cN)$, proving the upper bound through Lemma E.1.

For the lower bound, restrict also to $r = 1$, whose complement has probability at most $N^{-c_{\max}}$ by equation 15, and use Lemma K.2 with $\lambda = A_L^{-1}$. The horizon restriction ensures $d_\lambda \lesssim A_L^{1/a} \ll N$. On this event, $\widehat{\Sigma} \preceq C(\Sigma + A_L^{-1} I)$, so

$$\widehat{q}_j^\top \Sigma \widehat{q}_j \geq C^{-1}\widehat{\mu}_j - A_L^{-1} \geq (2C)^{-1}\widehat{\mu}_j \quad \text{if } \widehat{\mu}_j \geq 2C/A_L.$$

The empirical lower power law supplies order k_L such eigenvalues by choosing a sufficiently small constant fraction of the first k_L indices; these indices lie below $c\min\{M, N\}$. Each contributes a positive constant to the spectral expression in Lemma E.1. When k_L is bounded, applying the covariance comparison at a sufficiently small fixed regularization level bounds the top empirical eigenvalue and its population quadratic form below by constants, giving the same lower order. The intersection event has probability bounded below, yielding $\mathrm{Var}_{\mathrm{GD}} \gtrsim k_L/N$. Finally, splitting $\sum_j \min\{1, (A_L\mu_j)^2\}$ at $A_L^{1/a}$ gives its equivalence to k_L. □

F One-pass WSD-SGD: Excess Error Decomposition

Throughout this section, unmarked expectations are conditional on (S, w^*), so Σ and u^* are fixed. Outer target expectations are taken explicitly when deriving source-condition rates.

This section focuses on the one-pass WSD-SGD procedure equation 3 and uses the notation from Section A.4. In particular, the centered error $e_t := u_t^{\mathrm{op}} - u^*$ satisfies

$$e_t = \big(I - \gamma_t \widehat{\Sigma}_t^{\mathrm{op}}\big) e_{t-1} + \gamma_t \widehat{\xi}_t. \tag{17}$$

F.1 Excess error decomposition

Since u^* is the minimizer of the sketched population risk, the excess error is

$$\mathcal{E}_{\mathrm{ex}} := R_M(u_T^{\mathrm{op}}) - R_M(u^*) = \|u_T^{\mathrm{op}} - u^*\|_\Sigma^2 = \|e_T\|_\Sigma^2.$$

Taking expectation, we will decompose

$$\mathbb{E}\mathcal{E}_{\mathrm{ex}} = \mathbb{E}\|e_T\|_\Sigma^2$$

into deterministic and stochastic contributions.

We use throughout the original mean-centered decomposition

$$e_T = m_T + (e_T - m_T), \qquad m_T := \mathbb{E}[e_T \mid S, w^*].$$

Definition F.1 (Bias and centered variance)**.** Define the mean iterate error and centered fluctuation by

$$m_t := \mathbb{E}[u_t^{\mathrm{op}} - u^* \mid S, w^*], \qquad \delta_t := e_t - m_t.$$

The one-pass bias and variance terms are

$$\mathrm{Bias}_T := \|m_T\|_\Sigma^2, \qquad \mathrm{Var}_{B_{1:T}} := \mathbb{E}\|\delta_T\|_\Sigma^2.$$

Since

$$\mathbb{E}[\widehat{\Sigma}_t^{\mathrm{op}}] = \Sigma, \qquad \mathbb{E}[\widehat{\xi}_t] = 0,$$

the mean recursion satisfies

$$m_t = (I - \gamma_t\Sigma) m_{t-1}, \qquad m_0 = -u^*,$$

and hence

$$m_T = -\prod_{t=1}^{T} (I - \gamma_t \Sigma) u^*.$$

Therefore

$$\boxed{\mathrm{Bias}_T = \left\| \prod_{t=1}^{T} (I - \gamma_t \Sigma) u^* \right\|_\Sigma^2.}$$

Proposition F.1 (Exact mean-centered decomposition of the excess error)**.** Assume Assumptions 1.A and 1.B. Under the one-pass WSD-SGD recursion equation 17,

$$\mathbb{E}\mathcal{E}_{\mathrm{ex}} = \mathbb{E}\|e_T\|_\Sigma^2 = \mathrm{Bias}_T + \mathrm{Var}_{B_{1:T}}.$$

Proof. Since $e_T = m_T + \delta_T$, we have

$$\|e_T\|_\Sigma^2 = \|m_T\|_\Sigma^2 + 2\langle m_T, \delta_T\rangle_\Sigma + \|\delta_T\|_\Sigma^2.$$

Taking conditional expectation and using $\mathbb{E}[\delta_T \mid S, w^*] = 0$ gives

$$\mathbb{E}\|e_T\|_\Sigma^2 = \|m_T\|_\Sigma^2 + \mathbb{E}\|\delta_T\|_\Sigma^2,$$

which is exactly the claimed identity. □

F.2 Fluctuation recursion and an exact split of the variance term

Subtracting the mean recursion from equation 17 gives the exact centered fluctuation recursion

$$\delta_t = \big(I - \gamma_t \widehat{\Sigma}_t^{\mathrm{op}}\big)\delta_{t-1} + \gamma_t\big(\Sigma - \widehat{\Sigma}_t^{\mathrm{op}}\big)m_{t-1} + \gamma_t\widehat{\xi}_t. \tag{18}$$

Equivalently, in the shorthand notation of Section A.4,

$$\delta_t = \big(I - \gamma_t \bar{Z}_t\big)\delta_{t-1} + \gamma_t(\Sigma - \bar{Z}_t)m_{t-1} + \gamma_t\bar{\xi}_t. \tag{19}$$

For the variance analysis it is convenient to split δ_t into the part caused by batch-covariance fluctuations and the part caused by the additive label noise.

Definition F.2 (Centered covariance and noise components)**.** Define two auxiliary processes (q_t) and (v_t) by

$$q_0 = 0, \qquad v_0 = 0,$$

and, for $t \geq 1$,

$$q_t = (I - \gamma_t\bar{Z}_t)q_{t-1} + \gamma_t(\Sigma - \bar{Z}_t)m_{t-1},$$
$$v_t = (I - \gamma_t\bar{Z}_t)v_{t-1} + \gamma_t\bar{\xi}_t.$$

Define the corresponding quadratic terms by

$$\mathrm{Var}_{B_{1:T}}^{\mathrm{cov}} := \mathbb{E}\|q_T\|_\Sigma^2, \qquad \mathrm{Var}_{B_{1:T}}^{\mathrm{noise}} := \mathbb{E}\|v_T\|_\Sigma^2.$$

Proposition F.2 (Exact split of the centered variance)**.** Assume Assumptions 1.A and 1.B. Then, for every t,

$$\delta_t = q_t + v_t, \qquad \mathbb{E}[q_t] = 0.$$

Moreover, if

$$\mathcal{G}_t := \sigma\big(z_{s,b} : 1 \leq s \leq t,\ 1 \leq b \leq B_s\big),$$

then

$$\mathbb{E}[v_t \mid \mathcal{G}_t, w^*] = 0,$$

and therefore

$$\mathrm{Var}_{B_{1:T}} = \mathrm{Var}_{B_{1:T}}^{\mathrm{cov}} + \mathrm{Var}_{B_{1:T}}^{\mathrm{noise}}.$$

Proof. The identity $\delta_t = q_t + v_t$ follows by induction from equation 19. Indeed, it is true at time $t = 0$, and if $\delta_{t-1} = q_{t-1} + v_{t-1}$, then

$$\delta_t = \big(I - \gamma_t\bar{Z}_t\big)(q_{t-1} + v_{t-1}) + \gamma_t(\Sigma - \bar{Z}_t)m_{t-1} + \gamma_t\bar{\xi}_t = q_t + v_t.$$

Next, $\mathbb{E}[q_0] = 0$. If $\mathbb{E}[q_{t-1}] = 0$, then using that q_{t-1} depends only on batches $1, \ldots, t-1$ while $\bar{Z}_t$ comes from the disjoint t-th batch, we may use independence of q_{t-1} and $\bar{Z}_t$, together with $\mathbb{E}[\bar{Z}_t] = \Sigma$ and the fact that m_{t-1} is deterministic, to obtain

$$\mathbb{E}[q_t] = (I - \gamma_t\Sigma)\mathbb{E}[q_{t-1}] + \gamma_t\mathbb{E}[(\Sigma - \bar{Z}_t)m_{t-1}] = 0.$$

Thus $\mathbb{E}[q_t] = 0$ for every t.

To prove the conditional mean-zero property for v_t, fix one sample in the current batch and write $z = Sx$. Conditioning on the sketch matrix S, the pair (x, z) is jointly Gaussian with

$$\mathbb{E}[x \mid z] = HS^\top (SHS^\top)^{-1} z = HS^\top \Sigma^{-1} z.$$

Since $u^* = \Sigma^{-1} SHw^*$, the Gaussian conditional-regression identity yields

$$\mathbb{E}[\langle x, w^* \rangle \mid z, w^*] = \langle \mathbb{E}[x \mid z], w^* \rangle = \langle HS^\top \Sigma^{-1} z, w^* \rangle = z^\top u^*.$$

On the other hand, Assumption 1.B implies

$$\mathbb{E}[y - \langle x, w^* \rangle \mid x, w^*] = 0.$$

Therefore, by the tower property,

$$\mathbb{E}[y - z^\top u^* \mid z, w^*] = \mathbb{E}[y - \langle x, w^* \rangle \mid z, w^*] + \mathbb{E}[\langle x, w^* \rangle - z^\top u^* \mid z, w^*] = 0.$$

Multiplying by z, which is measurable with respect to $\sigma(z, w^*)$, gives

$$\mathbb{E}\big[z(y - z^\top u^*) \mid z, w^*\big] = 0.$$

Applying this to each summand in $\bar{\xi}_t = \frac{1}{B_t} \sum_{b=1}^{B_t} z_{t,b}(y_{i_{t,b}} - z_{t,b}^\top u^*)$, and conditioning on $\mathcal{G}_t$, which fixes the current sketched covariates $(z_{t,b})_{b=1}^{B_t}$, we obtain

$$\mathbb{E}[\bar{\xi}_t \mid \mathcal{G}_t, w^*] = 0.$$

Because v_{t-1} depends only on the first $t-1$ batches, it is independent of the current block $(z_{t,b})_{b=1}^{B}$. Hence

$$\mathbb{E}[v_t \mid \mathcal{G}_t, w^*] = (I - \gamma_t \bar{Z}_t)\mathbb{E}[v_{t-1} \mid \mathcal{G}_t, w^*] + \gamma_t \mathbb{E}[\bar{\xi}_t \mid \mathcal{G}_t, w^*].$$

Since conditioning on the additional current block does not change $\mathbb{E}[v_{t-1} \mid \mathcal{G}_t, w^*]$, the induction hypothesis gives $\mathbb{E}[v_{t-1} \mid \mathcal{G}_t, w^*] = \mathbb{E}[v_{t-1} \mid \mathcal{G}_{t-1}, w^*] = 0$. Starting from $v_0 = 0$, we therefore obtain by induction that

$$\mathbb{E}[v_t \mid \mathcal{G}_t, w^*] = 0 \qquad \text{for all } t.$$

Finally, q_T is $\mathcal{G}_T$-measurable, so

$$\mathbb{E}\langle q_T, v_T \rangle_\Sigma = \mathbb{E}\Big[\mathbb{E}\big[\langle q_T, v_T \rangle_\Sigma \mid \mathcal{G}_T, w^*\big]\Big] = 0.$$

Using $\delta_T = q_T + v_T$, we conclude that

$$\mathbb{E}\|\delta_T\|_\Sigma^2 = \mathbb{E}\|q_T\|_\Sigma^2 + \mathbb{E}\|v_T\|_\Sigma^2 = \mathrm{Var}^{\mathrm{cov}}_{B_{1:T}} + \mathrm{Var}^{\mathrm{noise}}_{B_{1:T}},$$

as claimed. □

G BIAS ERROR FOR ONE-PASS WSD-SGD

This section focuses on the one-pass WSD-SGD procedure equation 3 and uses the notation from Section A.4. In particular,

$$A_T := \sum_{t=1}^{T} \gamma_t \asymp \gamma T, \qquad \Sigma := SHS^\top, \qquad u^* := \Sigma^{-1} SHw^*.$$

The mean m_t is conditional on (S, w^*); target expectations are written explicitly. By Definition F.1, this section studies the exact one-pass bias term

$$\mathrm{Bias}_T = \|m_T\|_\Sigma^2.$$

Define

$$C_T := \prod_{t=1}^{T} (I - \gamma_t \Sigma).$$

Since the mean recursion satisfies $m_t = (I - \gamma_t \Sigma) m_{t-1}$ and $m_0 = -u^*$, we have

$$m_T = -C_T u^*, \qquad \mathrm{Bias}_T = \|C_T u^*\|_\Sigma^2.$$

G.1 Upper and lower bounds for the bias term

Lemma G.1 (Upper bound on the one-pass batch bias)**.** *Assume Assumptions 2 and 3, and the WSD conditions of Theorem 3.1. Fix any integer $k \le M/3$ such that $\mathrm{rank}(H) \ge k + M$, and define*

$$A_k := S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top.$$

Then, with probability at least

$$1 - \exp(-\Omega(M))$$

over the randomness of S,

$$\mathrm{Bias}_T \lesssim \frac{\|w^*_{0:k}\|_2^2}{A_T} \left(\frac{\mu_{M/2}(A_k)}{\mu_M(A_k)} \right)^2 + \|w^*_{k:\infty}\|^2_{H_{k:\infty}}.$$

In particular,

$$\mathrm{Bias}_T \lesssim \frac{\|u^*\|_2^2}{A_T}.$$

Proof. By Assumption 3, we may work in the diagonal coordinates of H, as encoded in Assumption 2. Define

$$M_T := C_T \Sigma C_T.$$

By Lemma A.1, for a constant $c > 0$,

$$M_T \preceq e^{-cA_T\Sigma} \Sigma e^{-cA_T\Sigma} =: M.$$

Therefore,

$$\begin{aligned} \mathrm{Bias}_T &= \langle M_T, u^* u^{*\top} \rangle \\ &\le \langle M, u^* u^{*\top} \rangle \\ &= w^{*\top} H S^\top \Sigma^{-1} M \Sigma^{-1} S H w^*. \end{aligned} \tag{20}$$

Now decompose

$$SH = (S_{0:k} H_{0:k},\ S_{k:\infty} H_{k:\infty}).$$

Then using the idea of spectral truncation, we have

$$2T_1 + 2T_2,$$

where

$$T_1 := w^{*\top}_{0:k} H_{0:k} S^\top_{0:k} \Sigma^{-1} M \Sigma^{-1} S_{0:k} H_{0:k} w^*_{0:k},$$

and

$$T_2 := w^{*\top}_{k:\infty} H_{k:\infty} S^\top_{k:\infty} \Sigma^{-1} M \Sigma^{-1} S_{k:\infty} H_{k:\infty} w^*_{k:\infty}.$$

For the head term, we have

$$T_1 \le \|M\|_2 \cdot \|\Sigma^{-1} S_{0:k} H_{0:k}\|_2^2 \cdot \|w^*_{0:k}\|_2^2.$$

By spectral calculus,

$$\|M\|_2 = \max_{x \in [0, \|\Sigma\|_2]} x e^{-2cA_T x} \lesssim \frac{1}{A_T}.$$

Moreover, Lemma K.7 yields

$$\|\Sigma^{-1} S_{0:k} H_{0:k}\|_2 \lesssim \frac{\mu_{M/2}(A_k)}{\mu_M(A_k)}$$

with probability at least $1 - \exp(-\Omega(M))$. Hence

$$T_1 \lesssim \frac{\|w^*_{0:k}\|_2^2}{A_T} \left(\frac{\mu_{M/2}(A_k)}{\mu_M(A_k)} \right)^2.$$

For the tail term, we also have

$$0 \preceq M = e^{-cA_T\Sigma} \Sigma e^{-cA_T\Sigma} \preceq \Sigma.$$

Therefore,

$$\begin{aligned} T_2 &\le w_{k:\infty}^{*\top} H_{k:\infty} S_{k:\infty}^\top \Sigma^{-1} \Sigma \Sigma^{-1} S_{k:\infty} H_{k:\infty} w_{k:\infty}^* \\ &= w_{k:\infty}^{*\top} H_{k:\infty}^{1/2} \Big(H_{k:\infty}^{1/2} S_{k:\infty}^\top \Sigma^{-1} S_{k:\infty} H_{k:\infty}^{1/2} \Big) H_{k:\infty}^{1/2} w_{k:\infty}^* \\ &\le \|w_{k:\infty}^*\|_{H_{k:\infty}}^2, \end{aligned}$$

where the last step uses

$$0 \preceq H_{k:\infty}^{1/2} S_{k:\infty}^\top \Sigma^{-1} S_{k:\infty} H_{k:\infty}^{1/2} \preceq I.$$

Combining the bounds on T_1 and T_2 proves the first claim.

For the simpler bound, we may ignore the head–tail split and use

$$\mathrm{Bias}_T \le \|M\|_2 \, \|u^*\|_2^2 \lesssim \frac{\|u^*\|_2^2}{A_T}.$$

□

Lemma G.2 (Lower bound on the one-pass batch bias)**.** *Assume the WSD conditions of Theorem 3.1. Let*

$$H_w := \mathbb{E}_{w^*}[w^* w^{*\top}], \qquad \Sigma_w := S H H_w H S^\top.$$

Then, conditioned on the sketch matrix S,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \sum_{i:\, \mu_i(\Sigma) < 1/A_T} \frac{\mu_i(\Sigma_w)}{\mu_i(\Sigma)}.$$

Proof. Let us define

$$M_T' := \Sigma e^{-C A_T \Sigma},$$

where $C > 0$ is a sufficiently large universal constant. The lower product comparison in Lemma A.1 gives

$$\prod_{t=1}^{T} (1 - \gamma_t x)^2 \ge e^{-C A_T x} \qquad \text{for every } x \in [0, \|\Sigma\|_2].$$

By functional calculus, this gives

$$M_T := C_T \Sigma C_T \succeq M_T'.$$

Therefore,

$$\mathrm{Bias}_T = \langle M_T, u^* u^{*\top} \rangle \ge \langle M_T', u^* u^{*\top} \rangle.$$

Since

$$\mathbb{E}_{w^*}[u^* u^{*\top}] = \Sigma^{-1} \Sigma_w \Sigma^{-1},$$

we have

$$\begin{aligned} \mathbb{E}_{w^*}[\mathrm{Bias}_T] &\ge \mathrm{tr}\big(M_T' \Sigma^{-1} \Sigma_w \Sigma^{-1}\big) \\ &= \mathrm{tr}\big(\Sigma^{-1} M_T' \Sigma^{-1} \Sigma_w\big). \end{aligned} \tag{21}$$

Applying von Neumann's trace inequality to equation 21 gives

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \sum_{i=1}^{M} \mu_{M-i+1}(\Sigma^{-1} M_T' \Sigma^{-1}) \, \mu_i(\Sigma_w).$$

Since M_T' is positive definite, the identity $\mu_{M-i+1}(A) = \mu_i(A^{-1})^{-1}$ yields

$$\mu_{M-i+1}(\Sigma^{-1} M_T' \Sigma^{-1}) = \frac{1}{\mu_i(\Sigma^2 M_T'^{-1})} = \frac{1}{\mu_i\big(\Sigma e^{C A_T \Sigma}\big)}.$$

Therefore,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \sum_{i=1}^{M} \frac{\mu_i(\Sigma_w)}{\mu_i\big(\Sigma e^{C A_T \Sigma}\big)}.$$

If $\mu_i(\Sigma) < 1/A_T$, then

$$e^{CA_T\mu_i(\Sigma)} \lesssim 1,$$

so the corresponding denominator is comparable to $\mu_i(\Sigma)$. Restricting the sum to this index set gives

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \sum_{i:\,\mu_i(\Sigma)<1/A_T} \frac{\mu_i(\Sigma_w)}{\mu_i(\Sigma)},$$

which is exactly the claimed bound. □

G.2 Bounds under the source condition

Lemma G.3 (Bounds on the one-pass batch bias under the source condition)**.** *Assume Assumptions 1.C, 2, and 3, together with the WSD conditions in Theorem 3.1. Suppose moreover that*

$$a > b - 1.$$

Then there exists an (a, b)-dependent constant $c_1 > 0$ such that, with probability at least

$$1 - \exp(-\Omega(M))$$

over the randomness of S,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \lesssim \min\{M, A_T^{1/a}\}^{1-b}.$$

Moreover, if in addition

$$A_T^{1/a} \le \frac{M}{c_1},$$

then, on the same event,

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim A_T^{(1-b)/a}.$$

Proof. In this proof, we verify the upper and lower bounds separately.

For the upper bound, choose

$$k := \min\{M/3, A_T^{1/a}\}.$$

By Lemma K.9,

$$\frac{\mu_{M/2}(A_k)}{\mu_M(A_k)} \lesssim 1$$

with probability at least $1 - \exp(-\Omega(M))$. Under Assumption 2,

$$\mathbb{E}_{w^*}\|w^*_{0:k}\|_2^2 \asymp \sum_{i=1}^{k} i^{a-b}, \qquad \mathbb{E}_{w^*}\|w^*_{k:\infty}\|_{H_{k:\infty}}^2 \asymp \sum_{i>k} i^{-b}.$$

Plugging these relations into Lemma G.1 gives

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \lesssim \frac{1}{A_T}\sum_{i=1}^{k} i^{a-b} + \sum_{i>k} i^{-b} \lesssim k^{1-b}.$$

Since $k \asymp \min\{M, A_T^{1/a}\}$, this yields

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \lesssim \min\{M, A_T^{1/a}\}^{1-b}.$$

For the lower bound, set

$$\theta := A_T^{1/a}.$$

Lemma K.8 gives

$$\mu_i(\Sigma) \asymp i^{-a} \qquad \text{for } i \in [M]$$

with probability at least $1 - \exp(-\Omega(M))$. Therefore, the condition

$$\mu_i(\Sigma) < \frac{1}{A_T}$$

forces

$$i \gtrsim \theta.$$

Moreover, under Assumption 2, the operator HH_wH has eigenvalues of order i^{-a-b}, so Lemma K.8 applied to HH_wH yields

$$\mu_i(\Sigma_w) \asymp i^{-a-b}.$$

Combining these estimates with Lemma G.2, we obtain

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \sum_{i \gtrsim \theta} \frac{\mu_i(\Sigma_w)}{\mu_i(\Sigma)} \asymp \sum_{i \gtrsim \theta} i^{-b}.$$

If $\theta \le M/c_1$ for a sufficiently large (a,b)-dependent constant c_1, then the last sum contains the range $[C\theta, 2C\theta]$ for some constant $C > 0$, and hence

$$\mathbb{E}_{w^*}[\mathrm{Bias}_T] \gtrsim \theta^{1-b} \gtrsim A_T^{(1-b)/a}.$$

Combining the upper and lower bounds proves the claim. □

H VARIANCE ERROR FOR ONE-PASS WSD-SGD

This section proves the one-pass variance bound for an arbitrary deterministic batch-size schedule $B_1, \dots, B_T$ satisfying $\sum_{t=1}^T B_t = N$. We use the notation from Section A.4. Unless an outer target expectation is indicated, expectations are conditional on (S, w^*), as in Definition F.1. In particular,

$$\tau_{\mathrm{d}}^{\mathrm{op}} := \frac{T - T_{\mathrm{s}}}{\log T}, \qquad A_{\mathrm{d}}^{\mathrm{op}} := \gamma \tau_{\mathrm{d}}^{\mathrm{op}},$$

$$\bar{Z}_t := \frac{1}{B_t} \sum_{r=1}^{B_t} z_{t,r} z_{t,r}^\top, \qquad \bar{\xi}_t := \frac{1}{B_t} \sum_{r=1}^{B_t} z_{t,r}(y_{i_{t,r}} - z_{t,r}^\top u^*).$$

Throughout this section, $(\gamma_t)_{t=1}^T$ is the one-pass WSD schedule from equation 6.

Let α be the fourth-moment constant from Lemma K.3; for Gaussian design one may take $\alpha = 3$. Set

$$R_\star^2 := \alpha \operatorname{tr}(\Sigma), \qquad \widetilde{\sigma}^2(w^*) := 2\big(\sigma^2 + \alpha \|w^*\|_H^2\big),$$

and

$$\overline{\sigma}^2(w^*) := \widetilde{\sigma}^2(w^*) + \alpha \|w^*\|_H^2.$$

The constant in the one-pass stepsize assumption is chosen so that

$$\gamma R_\star^2 \le \frac{1}{4}. \tag{22}$$

This is the same primitive stepsize condition already used in the fixed-batch analysis. It is uniform over every schedule because $B_t \ge 1$; no additional assumption involving $B_{\min}$ is required.

H.1 TIME-DEPENDENT BATCH MOMENTS AND THE RAW KERNEL

For any PSD matrix A, independence within the t-th batch gives

$$\mathbb{E}[\bar{Z}_t A \bar{Z}_t] = \frac{1}{B_t} \mathbb{E}[z z^\top A z z^\top] + \left(1 - \frac{1}{B_t}\right) \Sigma A \Sigma, \tag{23}$$

$$\mathbb{E}[(\bar{Z}_t - \Sigma) A (\bar{Z}_t - \Sigma)] = \frac{1}{B_t} \left(\mathbb{E}[z z^\top A z z^\top] - \Sigma A \Sigma\right). \tag{24}$$

Similarly, the conditional centering proved in Proposition F.2 yields

$$\mathbb{E}[\bar{\xi}_t \bar{\xi}_t^\top] = \frac{1}{B_t} \mathbb{E}\big[z (y - z^\top u^*)^2 z^\top\big]. \tag{25}$$

Definition H.1 (Dynamic raw kernel)**.** Define

$$\mathcal{K}_{B_{1:T}} := \sum_{t=1}^{T} \frac{\gamma_t^2}{B_t} \left\langle \Sigma, \left[\prod_{s=t+1}^{T} (I - \gamma_s \Sigma)^2 \right] \Sigma \right\rangle = \sum_{t=1}^{T} \frac{\kappa_t^{\mathrm{op}}}{B_t}. \tag{26}$$

By the definition of B_{eff},

$$\mathcal{K}_{B_{1:T}} = \frac{1}{B_{\mathrm{eff}}} \sum_{t=1}^{T} \kappa_t^{\mathrm{op}}. \tag{27}$$

The main issue is to retain the factor $1/B_t$ after the random covariance operators are iterated. A direct replacement by $1/B_{\min}$ would be stable but would lose the claimed effective-batch rate. The next two subsections use a uniform covariance bound only to control the multiplicative perturbation; the perturbation at time t itself still carries the exact factor $1/B_t$.

H.2 The additive-noise component

Recall

$$v_t = (I - \gamma_t \bar{Z}_t) v_{t-1} + \gamma_t \bar{\xi}_t, \qquad v_0 = 0,$$

and let $C_t := \mathbb{E}[v_t v_t^\top]$. For a deterministic PSD matrix A, write

$$\mathcal{A}_t(A) := \mathbb{E}[(I - \gamma_t \bar{Z}_t) A (I - \gamma_t \bar{Z}_t)]$$

and

$$\mathcal{D}_t(A) := (I - \gamma_t \Sigma) A (I - \gamma_t \Sigma).$$

Lemma H.1 (Time-dependent covariance recursion for additive noise)**.** *Under Assumptions 1.A and 1.B, define $\Sigma_{\xi,t} := \mathbb{E}[\bar{\xi}_t \bar{\xi}_t^\top]$. Then*

$$\frac{\sigma^2}{B_t} \Sigma \preceq \Sigma_{\xi,t} \preceq \frac{\widetilde{\sigma}^2(w^*)}{B_t} \Sigma$$

and

$$C_t = \mathcal{A}_t(C_{t-1}) + \gamma_t^2 \Sigma_{\xi,t}. \tag{28}$$

Moreover, under equation 22,

$$C_t \preceq \frac{\gamma \widetilde{\sigma}^2(w^*)}{1 - \gamma R_\star^2} I \qquad \textit{for every } t. \tag{29}$$

Proof. For the lower covariance bound, write

$$\varepsilon := y - x^\top w^*, \qquad r := w^* - S^\top u^*.$$

Then $y - z^\top u^* = \varepsilon + x^\top r$. By the well-specified, homoscedastic assumption,

$$\mathbb{E}[\varepsilon^2 z z^\top] = \sigma^2 \Sigma, \qquad \mathbb{E}[\varepsilon (x^\top r) z z^\top] = 0.$$

The remaining quadratic term is PSD, so

$$\mathbb{E}[z(y - z^\top u^*)^2 z^\top] \succeq \sigma^2 \Sigma.$$

Together with equation 25, this gives the lower bound on $\Sigma_{\xi,t}$. The upper bound follows from Lemma K.3. The mixed terms in the second moment of v_t vanish after conditioning on the current sketched covariates and w^*, which proves equation 28.

For the uniform bound, the fourth-moment inequality and equation 23 imply

$$\mathbb{E}[\bar{Z}_t^2] \preceq R_\star^2 \Sigma \qquad \text{for all } t.$$

Set

$$K_v := \frac{\gamma \widetilde{\sigma}^2(w^*)}{1 - \gamma R_\star^2}.$$

If $C_{t-1} \preceq K_v I$, monotonicity of the fourth-moment operator, $B_t \geq 1$, and $\gamma_t \leq \gamma$ give

$$C_t \preceq K_v\big(I - 2\gamma_t\Sigma + \gamma_t^2 R_\star^2\Sigma\big) + \gamma_t^2\widetilde{\sigma}^2(w^*)\Sigma \preceq K_v I.$$

Indeed, using $\gamma\widetilde{\sigma}^2(w^*) = K_v(1-\gamma R_\star^2)$, the coefficient of Σ in $C_t - K_v I$ is at most

$$-K_v\gamma_t\big(1 + (\gamma - \gamma_t)R_\star^2\big) \leq 0.$$

Since $C_0 = 0$, induction proves equation 29. □

Theorem H.1 (Two-sided bound for the additive-noise component)**.** *Under the assumptions of Lemma H.1 and equation 22,*

$$\sigma^2\mathcal{K}_{B_{1:T}} \leq \mathrm{Var}^{\mathrm{noise}}_{B_{1:T}} \leq \frac{\widetilde{\sigma}^2(w^*)}{1-\gamma R_\star^2}\,\mathcal{K}_{B_{1:T}}.$$

Proof. For every PSD A,

$$\mathcal{A}_t(A) - \mathcal{D}_t(A) = \gamma_t^2\mathbb{E}[(\bar{Z}_t - \Sigma)A(\bar{Z}_t - \Sigma)] \succeq 0.$$

Combining this with the lower bound on $\Sigma_{\xi,t}$, then unrolling equation 28, gives

$$C_T \succeq \sigma^2\sum_{t=1}^{T}\frac{\gamma_t^2}{B_t}\left[\prod_{s=t+1}^{T}(I-\gamma_s\Sigma)^2\right]\Sigma.$$

Taking the inner product with Σ proves the lower bound.

For the upper bound, equation 23, Lemma K.3, and equation 29 imply

$$\begin{aligned}\mathcal{A}_t(C_{t-1}) &\preceq \mathcal{D}_t(C_{t-1}) + \frac{\alpha\gamma_t^2}{B_t}\,\mathrm{tr}(\Sigma C_{t-1})\Sigma \\ &\preceq \mathcal{D}_t(C_{t-1}) + \frac{\alpha K_v\,\mathrm{tr}(\Sigma)\gamma_t^2}{B_t}\Sigma.\end{aligned}$$

After adding the driving covariance,

$$C_t \preceq \mathcal{D}_t(C_{t-1}) + \frac{\gamma_t^2}{B_t}\left(\widetilde{\sigma}^2(w^*) + \alpha K_v\,\mathrm{tr}(\Sigma)\right)\Sigma.$$

Since

$$\widetilde{\sigma}^2(w^*) + \alpha K_v\,\mathrm{tr}(\Sigma) = \frac{\widetilde{\sigma}^2(w^*)}{1-\gamma R_\star^2},$$

unrolling the deterministic maps $\mathcal{D}_t$ and taking the inner product with Σ proves the upper bound. Notice that the intermediate uniform estimate does not replace B_t: every driving term retains its own factor $1/B_t$. □

H.3 The covariance-fluctuation component

Recall

$$q_t = (I - \gamma_t\bar{Z}_t)q_{t-1} + \gamma_t(\Sigma - \bar{Z}_t)m_{t-1}, \qquad q_0 = 0,$$

and put

$$Q_t := \mathbb{E}[q_t q_t^\top], \qquad \Lambda_t := \mathbb{E}[(\Sigma - \bar{Z}_t)m_{t-1}m_{t-1}^\top(\Sigma - \bar{Z}_t)].$$

Lemma H.2 (Time-dependent covariance recursion for covariance fluctuation)**.** *Under Assumptions 1.A and 1.B,*

$$\Lambda_t \preceq \frac{\alpha\|m_{t-1}\|_\Sigma^2}{B_t}\Sigma \preceq \frac{\alpha\|u^*\|_\Sigma^2}{B_t}\Sigma,$$

and

$$Q_t = \mathcal{A}_t(Q_{t-1}) + \gamma_t^2\Lambda_t. \tag{30}$$

Under equation 22,

$$Q_t \preceq \frac{\gamma\alpha\|u^*\|_\Sigma^2}{1-\gamma R_\star^2}\,I \qquad \textit{for every } t.$$

Proof. Equation equation 24 and Lemma K.3 give

$$\Lambda_t \preceq \frac{\alpha}{B_t} \operatorname{tr}(\Sigma m_{t-1} m_{t-1}^\top)\Sigma.$$

The deterministic mean recursion is contractive in the Σ-norm, so $\|m_{t-1}\|_\Sigma \le \|u^*\|_\Sigma$.

The cross terms in the second moment of q_t vanish: q_{t-1} is independent of the current batch and $\mathbb{E}[q_{t-1}] = 0$. This proves equation 30. The uniform estimate follows from the same induction as in Lemma H.1, with $\widetilde{\sigma}^2(w^*)$ replaced by $\alpha\|u^*\|_\Sigma^2$. □

Theorem H.2 (Bound on the covariance-fluctuation component)**.** *Under the assumptions of Lemma H.2 and equation 22,*

$$0 \le \mathrm{Var}^{\mathrm{cov}}_{B_{1:T}} \le \frac{\alpha\|u^*\|_\Sigma^2}{1-\gamma R_\star^2}\,\mathcal{K}_{B_{1:T}}.$$

Proof. Let

$$K_q := \frac{\gamma\alpha\|u^*\|_\Sigma^2}{1-\gamma R_\star^2}.$$

As in the additive-noise proof,

$$\mathcal{A}_t(Q_{t-1}) \preceq \mathcal{D}_t(Q_{t-1}) + \frac{\alpha K_q \operatorname{tr}(\Sigma)\gamma_t^2}{B_t}\Sigma.$$

Combining this with the bound on Λ_t gives

$$Q_t \preceq \mathcal{D}_t(Q_{t-1}) + \frac{\gamma_t^2}{B_t}\frac{\alpha\|u^*\|_\Sigma^2}{1-\gamma R_\star^2}\Sigma.$$

Unrolling and taking the inner product with Σ proves the result. □

Proposition H.1 (Raw dynamic-batch variance bound)**.** Under Assumptions 1.A and 1.B, and under equation 22,

$$\sigma^2\mathcal{K}_{B_{1:T}} \le \mathrm{Var}^{\mathrm{noise}}_{B_{1:T}} \le \frac{\widetilde{\sigma}^2(w^*)}{1-\gamma R_\star^2}\mathcal{K}_{B_{1:T}},$$

$$0 \le \mathrm{Var}^{\mathrm{cov}}_{B_{1:T}} \le \frac{\alpha\|u^*\|_\Sigma^2}{1-\gamma R_\star^2}\mathcal{K}_{B_{1:T}},$$

and consequently

$$\sigma^2\mathcal{K}_{B_{1:T}} \le \mathrm{Var}_{B_{1:T}} \le \frac{\overline{\sigma}^2(w^*)}{1-\gamma R_\star^2}\mathcal{K}_{B_{1:T}}.$$

Proof. Combine Proposition F.2, Theorem H.1, and Theorem H.2. To compare the two norms, set

$$P := H^{1/2}S^\top\Sigma^{-1}SH^{1/2}.$$

Since $\Sigma = SHS^\top$, one has $P^2 = P$, so P is an orthogonal projector. Therefore

$$\|u^*\|_\Sigma^2 = w^{*\top}H^{1/2}PH^{1/2}w^* \le \|w^*\|_H^2,$$

which yields the stated last constant. □

H.4 Bounds under the source condition

Lemma H.3 (Source-condition bounds for the exact one-pass variance terms)**.** *Assume Assumptions 1.A, 1.B, 1.C, 2, and 3, together with the WSD conditions in Theorem 3.1. Suppose*

$$\sigma^2 \asymp 1, \qquad A_{\mathrm{d}}^{\mathrm{op}} = \gamma\tau_{\mathrm{d}}^{\mathrm{op}} \gtrsim 1.$$

Then, with probability at least $1 - \exp(-\Omega(M))$ *over* S*,*

$$\mathbb{E}_{w^*}[\mathrm{Var}^{\mathrm{noise}}_{B_{1:T}}] \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}} B_{\mathrm{eff}}},$$

$$\mathbb{E}_{w^*}[\mathrm{Var}^{\mathrm{cov}}_{B_{1:T}}] \lesssim \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}} B_{\mathrm{eff}}},$$

and therefore

$$\mathbb{E}_{w^*}[\mathrm{Var}_{B_{1:T}}] \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}} B_{\mathrm{eff}}}.$$

Proof. On the power-law sketch event, $\mathrm{tr}(\Sigma) \lesssim \sum_{j\geq 1} j^{-a} \lesssim 1$. Thus the constant in $\gamma \leq c$ can be chosen so that $\gamma R_\star^2 \leq 1/4$, and Proposition H.1 applies. Under the source condition and $b > 1$,

$$\mathbb{E}_{w^*}\|w^*\|_H^2 \asymp \sum_{j\geq 1} j^{-b} \lesssim 1.$$

Hence $\mathbb{E}_{w^*}\widetilde{\sigma}^2(w^*) \lesssim 1$ and $\mathbb{E}_{w^*}\overline{\sigma}^2(w^*) \lesssim 1$. Apply Proposition H.1 and Lemma H.5 below. The additive-noise lower bound supplies the matching lower order for the total centered variance. □

H.5 The WSD influence kernel

Lemma H.4 (Two-sided WSD influence-kernel bound under power-law decay)**.** *Assume the WSD phase-length and stability conditions in Theorem 3.1. On the power-law sketch event $\mu_j(\Sigma) \asymp j^{-a}$, suppose $A_{\mathrm{d}}^{\mathrm{op}} = \gamma\tau_{\mathrm{d}}^{\mathrm{op}} \gtrsim 1$. Then*

$$\sum_{t=1}^{T} \kappa_t^{\mathrm{op}} \asymp \frac{1}{\tau_{\mathrm{d}}^{\mathrm{op}}} \sum_{j=1}^{M} \min\left\{1, (A_{\mathrm{d}}^{\mathrm{op}}\mu_j(\Sigma))^2\right\}. \tag{31}$$

Consequently,

$$\sum_{t=1}^{T} \kappa_t^{\mathrm{op}} \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}}}. \tag{32}$$

Proof. Write

$$R_t := \sum_{s=t+1}^{T} \gamma_s.$$

The stability condition and the logarithmic product bound imply

$$\prod_{s=t+1}^{T} (1 - \gamma_s\mu)^2 \leq e^{-c\mu R_t}.$$

For a power-law spectrum we repeatedly use

$$\sum_{j=1}^{M} \mu_j^2 e^{-r\mu_j} \lesssim (1+r)^{-(2-1/a)}. \tag{33}$$

First consider the decay phase. With $t = T_{\mathrm{s}} + r$, its rates are $\gamma e^{-r/\tau_{\mathrm{d}}^{\mathrm{op}}}$, $1 \leq r \leq T_{\mathrm{d}}$, and $e^{-T_{\mathrm{d}}/\tau_{\mathrm{d}}^{\mathrm{op}}} = T^{-1}$. The geometric filter calculation, using the logarithmic product bounds valid for $\gamma\mu_j \leq c_0 < 1$, gives

$$\sum_{t=T_{\mathrm{s}}+1}^{T} \kappa_t^{\mathrm{op}} \asymp \frac{1}{\tau_{\mathrm{d}}^{\mathrm{op}}} \sum_{j=1}^{M} \min\{1, (A_{\mathrm{d}}^{\mathrm{op}}\mu_j)^2\}. \tag{34}$$

This already proves the required lower bound.

It remains to show that the earlier phases do not have a larger aggregate order. Every warmup update has at least a constant fraction of the stable phase after it, so $R_t \gtrsim \gamma T$. Since $\sum_{t\leq T_{\mathrm{w}}} \gamma_t^2 \lesssim \gamma^2 T$, equation 33 gives

$$\sum_{t=1}^{T_{\mathrm{w}}} \kappa_t^{\mathrm{op}} \lesssim \gamma^{1/a} T^{1/a-1} \lesssim \frac{(A_{\mathrm{d}}^{\mathrm{op}})^{1/a}}{\tau_{\mathrm{d}}^{\mathrm{op}}}. \tag{35}$$

If $M < (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}$, all retained eigenvalues lie above the decay cutoff. In this case, writing $y_j := \gamma T \mu_j$ gives the direct bound

$$\sum_{t=1}^{T_{\mathrm{w}}} \kappa_t^{\mathrm{op}} \lesssim \frac{1}{T} \sum_{j=1}^{M} y_j^2 e^{-cy_j} \lesssim \frac{M}{T} \lesssim \frac{M}{\tau_{\mathrm{d}}^{\mathrm{op}}}.$$

For a stable-phase update write $k = T_{\mathrm{s}} - t$. The remaining learning-rate mass satisfies $R_t \gtrsim \gamma(\tau_{\mathrm{d}}^{\mathrm{op}} + k)$. Summing first over k yields, for each eigenvalue,

$$\sum_{t=T_{\mathrm{w}}+1}^{T_{\mathrm{s}}} \gamma^2 \mu_j^2 e^{-c\mu_j R_t} \lesssim \gamma \mu_j e^{-c' A_{\mathrm{d}}^{\mathrm{op}} \mu_j}.$$

Consequently, power-law summation in the unsaturated case and the elementary bound $y e^{-c'y} \lesssim 1$ in the saturated case give

$$\sum_{t=T_{\mathrm{w}}+1}^{T_{\mathrm{s}}} \kappa_t^{\mathrm{op}} \lesssim \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}}}. \tag{36}$$

Combining equation 34–equation 36 proves equation 31. Finally, splitting the spectral sum at $j \asymp (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}$ proves equation 32. □

Remark H.1 (Why the power-law condition is used)**.** For a completely arbitrary spectrum, very small eigenvalues may accumulate a larger stable-phase contribution than the decay-only expression in equation 34. The power-law summation in the preceding lemma shows that this tail is lower order in the present model. Thus the theorem's WSD kernel rate is valid without any extra batch-schedule assumption, but the intermediate equivalence is not asserted spectrum-free.

Lemma H.5 (Power-law bound for the dynamic raw kernel)**.** *Assume Assumptions 1.C and 3, and suppose* $A_{\mathrm{d}}^{\mathrm{op}} = \gamma \tau_{\mathrm{d}}^{\mathrm{op}} \gtrsim 1$*. Then, with probability at least* $1 - \exp(-\Omega(M))$ *over S,*

$$\sum_{t=1}^{T} \kappa_t^{\mathrm{op}} \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}}},$$

and

$$\mathcal{K}_{B_{1:T}} = \sum_{t=1}^{T} \frac{\kappa_t^{\mathrm{op}}}{B_t} \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{op}} B_{\mathrm{eff}}}.$$

Proof. Let $s := A_{\mathrm{d}}^{\mathrm{op}}$. On the high-probability sketch event from Lemma K.8, $\mu_j(\Sigma) \asymp j^{-a}$ for $j \le M$. Splitting the sum in Lemma H.4 at $k = \min\{M, \lfloor s^{1/a} \rfloor\}$ gives

$$\sum_{j=1}^{M} \min\{1, (s\mu_j)^2\} \asymp \min\{M, s^{1/a}\}.$$

This proves the first display. The second follows exactly from equation 27; no comparison between the individual B_t's is needed. □

Remark H.2 (Constant-batch consistency check)**.** If $B_t \equiv B$, then $B_{\mathrm{eff}} = B$ and $T = N/B$. Therefore

$$\mathcal{K}_{B_{1:T}} \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{op}})^{1/a}\}}{B\tau_{\mathrm{d}}^{\mathrm{op}}},$$

which is the fixed-batch WSD variance law.

I Auxiliary With-Replacement Fluctuation Analysis

This proof-only section develops a with-replacement comparison process for a deterministic batch-size schedule $B_1, \ldots, B_L$, with $1 \le B_t \le N$, and the safeguarded multi-pass WSD learning-rate schedule. Appendix J applies the resulting bounds, together with the finite-population correction, to the without-replacement multi-pass procedure studied in the main text.

Write

$$A_L := \sum_{t=1}^{L} \gamma_t, \quad A_{\mathrm{d}}^{\mathrm{mp}} := \gamma \tau_{\mathrm{d}}^{\mathrm{mp}}, \qquad R_t := \sum_{s=t+1}^{L} \gamma_s,$$
$$\rho_t^{\mathrm{wr}} := \frac{1}{B_t}, \qquad \rho_t^{\mathrm{wor}} := \frac{N - B_t}{B_t(N-1)}.$$

For later use, abbreviate the WSD kernel scale by

$$\mathcal{K}_{\mathrm{d}} := \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{mp}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{mp}}}.$$

Lemma A.2 gives $A_L \asymp \gamma L$ and $A_{\mathrm{d}}^{\mathrm{mp}} \asymp \gamma L / \log L$. These masses and the population weights κ_t^{mp} below are nominal. Actual recursions and empirical kernels use $\eta_t = r\gamma_t$. Conditional on the training features, r is fixed and $\gamma_{\mathrm{safe}} C_A \le 1/8$ holds for every realization. Condition on the sketched dataset sigma-field $\mathcal{D}_{\mathrm{data}} = \sigma(S, w^*, D)$, and use the filtration

$$\mathcal{F}_t^{\mathrm{wr}} := \mathcal{D}_{\mathrm{data}} \vee \sigma(i_{s,r} : 1 \le s \le t,\ 1 \le r \le B_s).$$

Define

$$\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wr}} := \mathbb{E}_{\mathrm{batch}}\big[\|\Sigma^{1/2}(u_L^{\mathrm{wr}} - \theta_L)\|_2^2\big].$$

The fluctuation process satisfies

$$\Delta_t = (I - \eta_t \widehat{\Sigma}_t^{\mathrm{wr}})\Delta_{t-1} + \eta_t \xi_t^{\mathrm{wr}}, \qquad \Delta_0 = 0, \tag{37}$$

where

$$\xi_t^{\mathrm{wr}} := -(\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma})(\theta_{t-1} - u^*) + (\widehat{c}_t^{\mathrm{wr}} - \widehat{c}). \tag{38}$$

Since θ_{t-1} is determined by $\mathcal{D}_{\mathrm{data}}$, the current batch is independent of $\mathcal{F}_{t-1}^{\mathrm{wr}}$ and

$$\mathbb{E}[\xi_t^{\mathrm{wr}} \mid \mathcal{F}_{t-1}^{\mathrm{wr}}] = 0. \tag{39}$$

I.1 Conditional covariance identities

For $i \in [N]$, define the centered single-sample perturbation

$$\zeta_t(i) := -(S x_i x_i^\top S^\top - \widehat{\Sigma})(\theta_{t-1} - u^*) + (S x_i \widetilde{\varepsilon}_i - \widehat{c}).$$

Then

$$\frac{1}{N}\sum_{i=1}^{N} \zeta_t(i) = 0, \qquad \xi_t^{\mathrm{wr}} = \frac{1}{B_t}\sum_{r=1}^{B_t} \zeta_t(i_{t,r}).$$

Consequently, if

$$G_t := \frac{1}{N}\sum_{i=1}^{N} \zeta_t(i)\zeta_t(i)^\top,$$

then

$$\mathbb{E}[\xi_t^{\mathrm{wr}}(\xi_t^{\mathrm{wr}})^\top \mid \mathcal{F}_{t-1}^{\mathrm{wr}}] = \rho_t^{\mathrm{wr}} G_t. \tag{40}$$

Let $z_i := S x_i$, $Z_i := z_i z_i^\top$, and

$$C_A := \max_{i \in [N]} \|z_i\|_2^2.$$

For any deterministic PSD matrix C, the transition covariance obeys

$$\mathbb{E}[(\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma}) C (\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma}) \mid \mathcal{F}_{t-1}^{\mathrm{wr}}] = \rho_t^{\mathrm{wr}} \frac{1}{N}\sum_{i=1}^{N} (Z_i - \widehat{\Sigma}) C (Z_i - \widehat{\Sigma}). \tag{41}$$

In particular,

$$\mathbb{E}[(\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma})^2 \mid \mathcal{F}_{t-1}^{\mathrm{wr}}] \preceq \rho_t^{\mathrm{wr}} C_A \widehat{\Sigma}.$$

These identities show that both the additive stochastic source and the random-transition correction created at iteration t carry the same time-dependent factor ρ_t^{wr}.

Lemma I.1 (Dynamic-batch conditional covariance bound)**.** *Define*

$$\sigma_\xi^2 := 2 \max_{\substack{i\in[N]\\ t\in[L]}} \left[\left(x_i^\top S^\top (\theta_{t-1} - u^*) \right)^2 + \widetilde{\varepsilon}_i^2 \right].$$

Then, almost surely with respect to $\mathcal{D}_{\mathrm{data}}$*,*

$$G_t \preceq \sigma_\xi^2 \widehat{\Sigma} \qquad \textit{for every } t,$$

and hence

$$\mathbb{E}[\xi_t^{\mathrm{wr}} (\xi_t^{\mathrm{wr}})^\top \mid \mathcal{F}_{t-1}^{\mathrm{wr}}] \preceq \rho_t^{\mathrm{wr}} \sigma_\xi^2 \widehat{\Sigma}. \tag{42}$$

Proof. Put $d_t := \theta_{t-1} - u^*$. The covariance of $-(Z_i - \widehat{\Sigma})d_t$ is bounded by its second moment, which is at most

$$\max_j (z_j^\top d_t)^2 \, \widehat{\Sigma}.$$

Likewise, the covariance of $z_i \widetilde{\varepsilon}_i - \widehat{c}$ is at most

$$\max_j \widetilde{\varepsilon}_j^2 \, \widehat{\Sigma}.$$

Using $(a+b)(a+b)^\top \preceq 2aa^\top + 2bb^\top$ proves the result. □

For the lower bound, the response noise itself supplies a non-degenerate stochastic source. Write $\varepsilon_i := y_i - \langle x_i, w^* \rangle$, and let $\mathbb{E}_\varepsilon$ denote expectation over $(\varepsilon_i)_{i=1}^N$ conditional on S, w^*, and the covariates.

Lemma I.2 (Lower covariance of the batch perturbation)**.** *Let* $B_\nu := \max_{i\in[N]} \|z_i\|_2$*, so* $C_A = B_\nu^2$*, and define*

$$\beta_\xi^{\mathrm{WSD}} := \left[\frac{N-1}{N} - \frac{C A_L B_\nu^2}{N} \right]_+,$$

where $C > 0$ *is a universal constant. Under Assumption 1.B, for every* $t \in [L]$*,*

$$\mathbb{E}_\varepsilon[G_t \mid S, w^*, X] \succeq \sigma^2 \beta_\xi^{\mathrm{WSD}} \widehat{\Sigma}. \tag{43}$$

Consequently,

$$\mathbb{E}_\varepsilon \left[\mathbb{E}_{\mathrm{batch}} \left[\xi_t^{\mathrm{wr}} (\xi_t^{\mathrm{wr}})^\top \mid \mathcal{F}_{t-1}^{\mathrm{wr}} \right] \middle| S, w^*, X \right] \succeq \rho_t^{\mathrm{wr}} \sigma^2 \beta_\xi^{\mathrm{WSD}} \widehat{\Sigma}.$$

Proof. Conditional on S, w^*, X, the full-batch WSD-GD iterate is affine in the response-noise vector. Split the single-sample perturbation as $\zeta_t(i) = \zeta_t^{(0)}(i) + \zeta_t^{(\varepsilon)}(i)$, where the first term is independent of the response noises and the second is linear in them. The conditional mean-zero and homoscedastic assumptions imply that the cross terms vanish and hence

$$\mathbb{E}_\varepsilon[G_t \mid S, w^*, X] \succeq \frac{1}{N} \sum_{i=1}^N \mathbb{E}_\varepsilon \left[\zeta_t^{(\varepsilon)}(i) \zeta_t^{(\varepsilon)}(i)^\top \mid S, w^*, X \right].$$

Let

$$C_{t-1} := \prod_{s=1}^{t-1} (I - \eta_s \widehat{\Sigma}).$$

Put

$$V_{t-1} := (I - C_{t-1}) \widehat{\Sigma}^\dagger, \qquad q := \frac{1}{N} \sum_{k=1}^N z_k \varepsilon_k, \qquad A_i := Z_i - \widehat{\Sigma}.$$

The response-noise part of the GD error is $V_{t-1} q$, and therefore

$$\zeta_t^{(\varepsilon)}(i) = \underbrace{z_i \varepsilon_i - q}_{a_i} - \underbrace{A_i V_{t-1} q}_{b_i}.$$

Expanding $(a_i - b_i)(a_i - b_i)^\top$ exactly, we use

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\varepsilon[a_i a_i^\top \mid S, w^*, X] = \sigma^2\frac{N-1}{N}\widehat{\Sigma}, \qquad \mathbb{E}_\varepsilon[qq^\top \mid S, w^*, X] = \frac{\sigma^2}{N}\widehat{\Sigma}.$$

Moreover,

$$\mathbb{E}_\varepsilon[a_i q^\top \mid S, w^*, X] = \frac{\sigma^2}{N}A_i,$$

so the two averaged cross terms have operator magnitude at most

$$\frac{2\sigma^2}{N^2}\sum_{i=1}^{N}A_i V_{t-1}A_i \preceq \frac{2\sigma^2 A_L C_A}{N}\widehat{\Sigma}.$$

The schedule-dependent filter satisfies

$$0 \preceq V_{t-1}\widehat{\Sigma}V_{t-1} = (I - C_{t-1})^2\widehat{\Sigma}^\dagger \preceq \big(I - C_{t-1}^2\big)\widehat{\Sigma}^\dagger \preceq 2\sum_{s=1}^{t-1}\eta_s I \preceq 2A_L I,$$

which follows from the scalar inequality $1 - \prod_s(1-u_s)^2 \le 2\sum_s u_s$. Finally,

$$\frac{1}{N}\sum_{i=1}^{N}(Z_i - \widehat{\Sigma})^2 \preceq \frac{1}{N}\sum_{i=1}^{N}Z_i^2 \preceq C_A\widehat{\Sigma}.$$

The averaged $b_i b_i^\top$ term is PSD and may be dropped in this lower bound. Combining these displays, and enlarging the universal constant C, proves equation 43. The batch statement follows by first applying equation 40 conditional on the full dataset and then averaging over the response noises. □

I.2 A TIME-INHOMOGENEOUS STOCHASTIC APPROXIMATION LEMMA

The next abstract lemma uses local deterministic rate symbols γ_t, γ; its application substitutes η_t and γ_{safe}, conditional on the training features.

Lemma I.3 (Stochastic approximation with dynamic covariance factors)**.** *Let*

$$\mu_t = (I - \gamma_t A_t)\mu_{t-1} + \gamma_t\xi_t, \qquad \mu_0 = 0,$$

and let $\mathcal{F}_t = \sigma((A_s, \xi_s) : 1 \le s \le t)$. Suppose the deterministic numbers ρ_t lie in $[0,1]$, the current pair (A_t, ξ_t) is independent of $\mathcal{F}_{t-1}$, and, uniformly in t,

$$\mathbb{E}[A_t] = \Sigma_\nu, \qquad \mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_{t-1}] \preceq \rho_t\sigma_\xi^2\Sigma_\nu,$$

$$\mathbb{E}[(A_t - \Sigma_\nu)^2] \preceq \rho_t C_A\Sigma_\nu, \qquad \mathbb{E}[A_t^2] \preceq C_A\Sigma_\nu, \qquad \|\Sigma_\nu\|_2 \le C_A.$$

Assume also $0 < \gamma_t \le \gamma$ and $\gamma C_A \le 1/8$. Let

$$P_{t+1:L} := \prod_{s=t+1}^{L}(I - \gamma_s\Sigma_\nu).$$

Then

$$\mathbb{E}[\mu_L\mu_L^\top] \preceq \frac{9}{8}\sigma_\xi^2\sum_{t=1}^{L}\rho_t\gamma_t^2 P_{t+1:L}\Sigma_\nu P_{t+1:L}. \tag{44}$$

Consequently, for $u \in [0,1]$,

$$\mathbb{E}\|\Sigma_\nu^{u/2}\mu_L\|_2^2 \lesssim \sigma_\xi^2\sum_{t=1}^{L}\rho_t\gamma_t^2\,\mathrm{tr}\big(\Sigma_\nu^{u+1}P_{t+1:L}^2\big).$$

Proof. The independence of the current pair from the past and the martingale difference property imply $\mathbb{E}[\mu_t] = 0$. They also make the mixed terms in the unconditional second-moment recursion vanish, even though A_t and ξ_t may be dependent within the current pair. Thus, with $C_t := \mathbb{E}[\mu_t\mu_t^\top]$,

$$C_t = \mathbb{E}[(I - \gamma_t A_t)C_{t-1}(I - \gamma_t A_t)] + \gamma_t^2\mathbb{E}[\xi_t\xi_t^\top].$$

First, induction gives the schedule-independent crude estimate

$$C_t \preceq \sigma_\xi^2 \gamma I.$$

Indeed, use $\rho_t \le 1$, $\mathbb{E}[A_t^2] \preceq C_A \Sigma_\nu$, and $\gamma C_A \le 1/8$.

Next, decompose the transition map around its mean:

$$\begin{aligned}\mathbb{E}[(I-\gamma_t A_t)C_{t-1}(I-\gamma_t A_t)] = (I-\gamma_t\Sigma_\nu)C_{t-1}(I-\gamma_t\Sigma_\nu) \\ + \gamma_t^2 \mathbb{E}[(A_t-\Sigma_\nu)C_{t-1}(A_t-\Sigma_\nu)].\end{aligned}$$

The crude estimate and the centered transition bound imply that the last term is at most

$$\rho_t \sigma_\xi^2 \gamma C_A \gamma_t^2 \Sigma_\nu.$$

Adding the driving covariance and using $\gamma C_A \le 1/8$ gives

$$C_t \preceq (I-\gamma_t\Sigma_\nu)C_{t-1}(I-\gamma_t\Sigma_\nu) + \frac{9}{8}\rho_t\sigma_\xi^2\gamma_t^2\Sigma_\nu.$$

Unrolling proves equation 44; taking the appropriate trace proves the second claim. □

Lemma I.4 (Verification for dynamic with-replacement batches)**.** *Conditioned on $\mathcal{D}_{\mathrm{data}}$, set*

$$A_t := \widehat{\Sigma}_t^{\mathrm{wr}}, \qquad \Sigma_\nu := \widehat{\Sigma}, \qquad \rho_t := \rho_t^{\mathrm{wr}}.$$

Then all transition-matrix assumptions of Lemma I.3 hold. Together with Lemma I.1, its noise assumptions also hold with the σ_ξ^2 defined there.

Proof. The conditional mean is $\widehat{\Sigma}$, and equation 41 gives the centered transition bound. Moreover, pathwise

$$0 \preceq A_t \preceq C_A I, \qquad A_t^2 \preceq C_A A_t.$$

Taking expectations gives $\mathbb{E}[A_t^2] \preceq C_A\widehat{\Sigma}$, while $\widehat{\Sigma} \preceq C_A I$. Apply the preceding abstract lemma with its rates γ_t replaced by η_t and its peak γ by γ_{safe}; then $\gamma_{\mathrm{safe}} C_A \le 1/8$ holds pathwise. □

I.3 Influence kernels and covariance replacement

For $u \in \{0,1\}$, define the empirical kernels

$$\widehat{\kappa}_t^{(u)} := \eta_t^2 \operatorname{tr}\left(\widehat{\Sigma}^{u+1}\prod_{s=t+1}^{L}(I-\eta_s\widehat{\Sigma})^2\right).$$

Define also the empirical cross-kernel

$$\widehat{\kappa}_t^{(\times)} := \eta_t^2 \operatorname{tr}\left(\Sigma\widehat{\Sigma}\prod_{s=t+1}^{L}(I-\eta_s\widehat{\Sigma})^2\right). \tag{45}$$

The population final-risk kernel is

$$\kappa_t^{\mathrm{mp}} := \gamma_t^2 \operatorname{tr}\left(\Sigma^2\prod_{s=t+1}^{L}(I-\gamma_s\Sigma)^2\right) = \gamma_t^2\sum_{j=1}^{M}\mu_j^2\prod_{s=t+1}^{L}(1-\gamma_s\mu_j)^2. \tag{46}$$

Define

$$\rho_{\mathrm{eff,WSD}}^{\mathrm{wr}} := \frac{\sum_{t=1}^{L}\rho_t^{\mathrm{wr}}\kappa_t^{\mathrm{mp}}}{\sum_{t=1}^{L}\kappa_t^{\mathrm{mp}}}, \qquad \rho_{\mathrm{eff,WSD}}^{\mathrm{wor}} := \frac{\sum_{t=1}^{L}\rho_t^{\mathrm{wor}}\kappa_t^{\mathrm{mp}}}{\sum_{t=1}^{L}\kappa_t^{\mathrm{mp}}}. \tag{47}$$

Lemma I.5 (Uniform empirical-to-population kernel comparison)**.** *Assume the power-law and sketch conditions, and let*

$$\lambda_L := A_L^{-1}.$$

Under the horizon condition

$$A_L \gtrsim 1, \qquad A_L \lesssim N^{(1-\varepsilon)a}, \qquad (1+A_L)^{1/a} \le c_\star M,$$

the empirical spectral event of Lemma K.11 gives, uniformly in t *and the safeguard ratio* $r \in (0, 1]$*,*

$$\widehat{\kappa}_t^{(1)} + \lambda_L \widehat{\kappa}_t^{(0)} \lesssim \kappa_t^{\mathrm{mp}}.$$

Consequently, for every deterministic sequence $\rho_t \in [0, 1]$*,*

$$\sum_{t=1}^{L} \rho_t \left(\widehat{\kappa}_t^{(1)} + \lambda_L \widehat{\kappa}_t^{(0)} \right) \lesssim \sum_{t=1}^{L} \rho_t \kappa_t^{\mathrm{mp}}.$$

Proof. The safeguard ensures $\eta_s \nu \le 1/8$ on the empirical spectrum. The scalar logarithmic bounds reduce its kernels to sums of the form

$$r^2 \gamma_t^2 \sum_j \nu_j^q e^{-c r R_t \nu_j}, \qquad q \in \{1, 2\}.$$

For a power-law spectrum,

$$\sum_j \nu_j^q e^{-cR\nu_j} \lesssim (1 + R)^{-(q - 1/a)}.$$

The empirical upper and lower power-law comparisons control these sums uniformly for $rR_t \le A_L$. The unsaturated condition allows the matching population lower estimate at the same remaining mass, by Lemma A.3.

Finally,

$$\lambda_L \frac{(1 + rR_t)^{-(1-1/a)}}{(1 + rR_t)^{-(2-1/a)}} \lesssim \frac{1 + rR_t}{A_L} \lesssim 1.$$

Thus the left side is bounded by the actual population weight $\widetilde{\kappa}_t$, which is at most $C\kappa_t^{\mathrm{mp}}$ by Lemma A.5. Multiplication by arbitrary nonnegative ρ_t's and summation preserve it. □

Lemma I.6 (Uniform lower comparison for the cross-kernel)**.** *Assume the power-law and Gaussian-design conditions. Suppose*

$$A_L \gtrsim 1, \qquad A_L^{1/a} \le c_0 M, \qquad A_L \log N \le c_0 N$$

for a sufficiently small a*-dependent constant* $c_0 > 0$*. Then, conditioned on the usual high-probability sketch event, with probability at least* $1 - N^{-c} - \exp(-\Omega(N))$ *over the sample, on an event contained in* $\{r = 1\}$*,*

$$\widehat{\kappa}_t^{(\times)} \gtrsim \kappa_t^{\mathrm{mp}} \qquad \textit{uniformly for } t \in [L]. \tag{48}$$

Proof. Set $\lambda_0 := A_L^{-1}$. Under the power law,

$$d_{\lambda_0} = \mathrm{tr}\big(\Sigma(\Sigma + \lambda_0 I)^{-1}\big) \lesssim A_L^{1/a}.$$

The assumptions therefore imply the two-sided comparison in Lemma K.2 at λ_0, and hence at every $\lambda \ge \lambda_0$. Intersect this event with the empirical power-law event in Lemma K.11 and with $r = 1$. The latter costs at most N^{-c} by equation 15; on it $\eta_t = \gamma_t$. The condition $A_L \log N \le c_0 N$ implies $A_L^{1/a} \le cN$, so all spectral indices used below lie in the range of that lemma.

Fix t, write $R := R_t$, and first suppose $R \ge R_0$, where R_0 is a sufficiently large absolute constant. Put $\lambda := R^{-1}$. Choose fixed constants $K_2 > K_1 > 1$, sufficiently large relative to the constants in equation 59, and let Q_t be the spectral projector of $\widehat{\Sigma}$ corresponding to eigenvalues in $[K_1\lambda, K_2\lambda]$. Lemma K.11 gives

$$\mathrm{rank}(Q_t) \asymp R^{1/a}.$$

Moreover, the upper half of the regularized comparison gives

$$\widehat{\Sigma} + \lambda I \preceq C(\Sigma + \lambda I).$$

Consequently, by choosing $K_1 > C - 1$,

$$Q_t \Sigma Q_t \succeq c\lambda Q_t, \qquad \mathrm{tr}(\Sigma Q_t) \gtrsim \lambda R^{1/a}.$$

Let

$$h_t(x) := x \prod_{s=t+1}^{L} (1-\gamma_s x)^2.$$

The stepsize stability condition and the logarithmic product bound imply

$$h_t(x) \gtrsim \lambda \qquad \text{for } x \in [K_1\lambda, K_2\lambda],$$

because xR is bounded above and below by fixed constants on that interval. Hence

$$\widehat{\kappa}_t^{(\times)} = \gamma_t^2 \operatorname{tr}\big(\Sigma h_t(\widehat{\Sigma})\big) \gtrsim \gamma_t^2 \lambda \operatorname{tr}(\Sigma Q_t) \gtrsim \gamma_t^2 R^{1/a-2}.$$

The population power law and the same product comparison give $\kappa_t^{\mathrm{mp}} \lesssim \gamma_t^2 R^{1/a-2}$.

It remains to consider $R < R_0$. On the empirical power-law event, $\|\widehat{\Sigma}\|_2 \lesssim 1$, so all future products are bounded below by a fixed positive constant. Therefore

$$\widehat{\kappa}_t^{(\times)} \gtrsim \gamma_t^2 \operatorname{tr}(\Sigma\widehat{\Sigma}).$$

The scalar Gaussian concentration bound applied to $N^{-1}\sum_i z_i^\top \Sigma z_i$, together with $\operatorname{tr}(\Sigma^2) \asymp 1$, yields $\operatorname{tr}(\Sigma\widehat{\Sigma}) \gtrsim 1$. On the other hand, $\kappa_t^{\mathrm{mp}} \lesssim \gamma_t^2 \operatorname{tr}(\Sigma^2) \lesssim \gamma_t^2$. This proves the uniform comparison. □

Lemma I.7 (Shape and total mass of the multi-pass WSD influence kernel)**.** *Under the power-law spectrum,*

$$\kappa_t^{\mathrm{mp}} \lesssim \gamma_t^2 (1+R_t)^{-(2-1/a)}.$$

Moreover, if $A_{\mathrm{d}}^{\mathrm{mp}} \gtrsim 1$, then

$$\sum_{t=1}^{L} \kappa_t^{\mathrm{mp}} \asymp \frac{1}{\tau_{\mathrm{d}}^{\mathrm{mp}}} \sum_{j=1}^{M} \min\big\{1, (A_{\mathrm{d}}^{\mathrm{mp}}\mu_j)^2\big\} \asymp \frac{\min\{M, (A_{\mathrm{d}}^{\mathrm{mp}})^{1/a}\}}{\tau_{\mathrm{d}}^{\mathrm{mp}}}. \tag{49}$$

The constants may depend on the fixed WSD phase fractions and on the stepsize-stability constant.

Proof. The first claim follows from the logarithmic product comparison and the power-law spectral-sum estimate with exponent two. We split the time sum into warmup, stable, and decay phases.

For $t \le L_{\mathrm{w}}$, the remaining stable phase has mass $R_t \gtrsim \gamma L$. Since $\sum_{t\le L_{\mathrm{w}}} \gamma_t^2 \lesssim \gamma^2 L$,

$$\sum_{t=1}^{L_{\mathrm{w}}} \kappa_t^{\mathrm{mp}} \lesssim \gamma^{1/a} L^{1/a-1}.$$

For the stable phase put $k = L_{\mathrm{s}} - t$. Direct summation of the decay schedule gives

$$R_t \asymp \gamma(k + \tau_{\mathrm{d}}^{\mathrm{mp}}).$$

Since $A_{\mathrm{d}}^{\mathrm{mp}} \gtrsim 1$, summing the pointwise kernel bound over $k \ge 0$ yields

$$\sum_{t=L_{\mathrm{w}}+1}^{L_{\mathrm{s}}} \kappa_t^{\mathrm{mp}} \lesssim \gamma^{1/a} (\tau_{\mathrm{d}}^{\mathrm{mp}})^{1/a-1} = \frac{(A_{\mathrm{d}}^{\mathrm{mp}})^{1/a}}{\tau_{\mathrm{d}}^{\mathrm{mp}}}.$$

For the decay phase put $q = e^{-1/\tau_{\mathrm{d}}^{\mathrm{mp}}}$ and $r = t - L_{\mathrm{s}}$. Then $\gamma_t = \gamma q^r$, $q^{L_{\mathrm{d}}} = L^{-1}$, and $1-q \asymp 1/\tau_{\mathrm{d}}^{\mathrm{mp}}$. For each eigenvalue μ, a geometric-series/integral comparison gives

$$\sum_{t=L_{\mathrm{s}}+1}^{L} \gamma_t^2 \mu^2 \prod_{s=t+1}^{L} (1-\gamma_s\mu)^2 \asymp \frac{1}{\tau_{\mathrm{d}}^{\mathrm{mp}}} \min\big\{1, (A_{\mathrm{d}}^{\mathrm{mp}}\mu)^2\big\}. \tag{50}$$

Indeed, after setting $z_r = A_{\mathrm{d}}^{\mathrm{mp}}\mu q^r$, the sum is comparable to $(\tau_{\mathrm{d}}^{\mathrm{mp}})^{-2}\sum_r z_r^2 e^{-cz_r}$, which is of order $(\tau_{\mathrm{d}}^{\mathrm{mp}})^{-1}\min\{1, z_1^2\}$. Thus the decay phase alone supplies the lower bound.

For the upper bound, the warmup estimate is smaller than the decay order by a fixed power of $\log L$. The stable estimate has the same order when the cutoff lies below M; when $M < (A_{\mathrm{d}}^{\mathrm{mp}})^{1/a}$, the

per-eigenvalue bound $ze^{-cz} \lesssim 1$, with $z = A_{\mathrm{d}}^{\mathrm{mp}}\mu_j$, improves it to $O(M/\tau_{\mathrm{d}}^{\mathrm{mp}})$. Hence all three phases are bounded by the decay expression. Finally,

$$\sum_{j=1}^{M} \min\{1, (A_{\mathrm{d}}^{\mathrm{mp}}\mu_j)^2\} \asymp \min\{M, (A_{\mathrm{d}}^{\mathrm{mp}})^{1/a}\}$$

under $\mu_j \asymp j^{-a}$, completing the proof. □

Remark I.1 (Exact weights are essential)**.** The phase estimates determine the total mass, but do not give a uniform phasewise proxy for arbitrary dynamic batch schedules. We therefore retain the exact κ_t^{mp} in $\rho_{\mathrm{eff,WSD}}$, especially near phase boundaries and the final endpoint.

I.4 Dynamic fluctuation bounds

We average the response residuals before bounding their covariance. This avoids both a maximum of finite-variance label noises and a leave-one-out independence claim for the sample-dependent safeguard. Put $Z = XS^\top$, $K = ZZ^\top/N$, and $\mathcal{D}_z = \sigma(S, w^*, Z)$. Conditional on S, w^*, Gaussian design and the label model give

$$y = Zu^* + \widetilde{\varepsilon}, \qquad \mathbb{E}[\widetilde{\varepsilon} \mid \mathcal{D}_z] = 0, \qquad \mathbb{E}[\widetilde{\varepsilon}\widetilde{\varepsilon}^\top \mid \mathcal{D}_z] = \widetilde{\sigma}^2 I_N, \quad \widetilde{\sigma}^2 = \sigma^2 + \mathrm{Approx}(w^*).$$

Here $\mathbb{E}_{\mathrm{res}}$ denotes integration over this residual vector conditional on $\mathcal{D}_z$, including unobserved feature directions. It requires only second moments of the label noise.

Lemma I.8 (Residual covariance bound uniform over safeguarded filters)**.** *Order the positive eigenvalues of K decreasingly, let $d = \mathrm{rank}(Z)$, and let P_j project onto eigenvectors $j, \dots, d$, for $1 \le j \le d$. Define*

$$\Psi := \widetilde{\sigma}^2 + \max_{i\in[N],\, 1\le j\le d} |(P_j Z u^*)_i|^2.$$

For all t, $\mathbb{E}_{\mathrm{res}}[G_t \mid \mathcal{D}_z] \preceq \Psi\widehat{\Sigma}$. Moreover, for each fixed $p \ge 1$,

$$\left(\mathbb{E}_Z[\Psi^p \mid S, w^*]\right)^{1/p} \lesssim_p \log(2N)(\sigma^2 + \|w^*\|_H^2). \tag{51}$$

Proof. The training residual at time $t-1$ is $f_{t-1}(K)y$, where $f_{t-1}(x) = \prod_{s<t}(1-\eta_s x)$. The safeguard makes f_{t-1} decreasing and valued in $[0,1]$ on the spectrum of K. Summation by parts therefore expresses $f_{t-1}(K)Zu^*$ as a convex combination of $0, P_1Zu^*, \dots, P_dZu^*$. Also $f_{t-1}(K)^2 \preceq I$. Consequently every squared residual has conditional expectation at most Ψ. Dropping the centering term in the covariance of $z_i(y_i - z_i^\top\theta_{t-1})$ gives $\mathbb{E}_{\mathrm{res}}G_t \preceq N^{-1}\sum_i \Psi z_i z_i^\top = \Psi\widehat{\Sigma}$.

For the moment bound, condition on $Z^\top Z$. Gaussian design is invariant under every orthogonal transformation of the sample rows. Thus each P_jZu^* has a uniformly random orientation in $\mathbb{R}^N$, with length at most $\|Zu^*\|_2$. The spherical-coordinate tail bound and a union bound over at most N^2 coordinates give, for $u \ge 1$,

$$\mathbb{P}\left(\max_{i,j} |(P_jZu^*)_i|^2 > C\frac{\|Zu^*\|_2^2}{N}(\log(2N) + u) \;\middle|\; Z^\top Z, S, w^*\right) \le 2e^{-u}.$$

Integrate this tail and use $\|Zu^*\|_2^2/\|u^*\|_\Sigma^2 \sim \chi_N^2$ when $u^* \neq 0$ (the zero case is immediate), $\|u^*\|_\Sigma^2 \le \|w^*\|_H^2$, and $\widetilde{\sigma}^2 \le \sigma^2 + \|w^*\|_H^2$. This proves equation 51. The envelope is uniform over all the filters, so no rotational invariance of the safeguard itself is being assumed. □

Set $\lambda_L = A_L^{-1}$ and

$$R_p := \|(\Sigma + \lambda_L I)^{1/2}(\widehat{\Sigma} + \lambda_L I)^{-1/2}\|_2^2, \qquad F_{\mathrm{mp}} := R_p\Psi.$$

Lemma I.9 (Upper bound for dynamic with-replacement batches)**.** *Assume Assumptions 1.A, 1.B, 1.C, and 3, and use the safeguarded algorithm. For a fixed $\varepsilon \in (0,1)$, suppose*

$$A_L \gtrsim 1, \qquad A_L \lesssim N^{(1-\varepsilon)a}, \qquad (1 + A_L)^{1/a} \le c_\star M.$$

On the empirical spectral event of Lemma I.5,

$$\mathbb{E}_{\mathrm{res}}[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wr}} \mid \mathcal{D}_z] \lesssim F_{\mathrm{mp}} \sum_{t=1}^{L} \rho_t^{\mathrm{wr}}\kappa_t^{\mathrm{mp}}.$$

Proof. For each fixed response vector, the batch-mean fluctuation is zero. Hence its covariance recursion has no mixed terms, and this remains true after averaging over the residual responses conditional on $\mathcal{D}_z$. The transition maps depend only on Z. By Lemma I.8, the averaged driving covariance is at most $\rho_t \Psi \widehat{\Sigma}$. The second-moment induction in Lemma I.3 therefore applies with $\sigma_\xi^2 = \Psi$, rates η_t, and peak γ_{safe}. The covariance replacement

$$\|\Sigma^{1/2}\Delta_L\|_2^2 \le R_p \left(\|\widehat{\Sigma}^{1/2}\Delta_L\|_2^2 + \lambda_L \|\Delta_L\|_2^2 \right)$$

then gives, for every feature realization,

$$\mathbb{E}_{\mathrm{res}}[\mathrm{Fluc}^{\mathrm{wr}}_{B_{1:L}} \mid \mathcal{D}_z] \lesssim F_{\mathrm{mp}} \sum_{t=1}^{L} \rho_t^{\mathrm{wr}} \left(\widehat{\kappa}_t^{(1)} + \lambda_L \widehat{\kappa}_t^{(0)} \right).$$

This all-sample bound will also control the exceptional spectral event. On its complement, Lemma I.5 proves the claim. □

Lemma I.10 (Empirical lower bound for dynamic with-replacement batches)**.** *Under the notation above,*

$$\mathbb{E}_\varepsilon \left[\mathrm{Fluc}^{\mathrm{wr}}_{B_{1:L}} \mid S, w^*, X \right] \ge \sigma^2 \beta_\xi^{\mathrm{WSD}} \sum_{t=1}^{L} \rho_t^{\mathrm{wr}} \widehat{\kappa}_t^{(\times)}. \tag{52}$$

Proof. First condition on the entire sketched dataset. Since $\mathbb{E}_{\mathrm{batch}}[\Delta_t] = 0$, the mixed terms vanish in the conditional second-moment recursion. For every deterministic PSD matrix C,

$$\begin{aligned}
&\mathbb{E}_{\mathrm{batch}} \Big[(I - \eta_t \widehat{\Sigma}_t^{\mathrm{wr}}) C (I - \eta_t \widehat{\Sigma}_t^{\mathrm{wr}}) \mid \mathcal{F}_{t-1}^{\mathrm{wr}} \Big] \\
&\quad = (I - \eta_t \widehat{\Sigma}) C (I - \eta_t \widehat{\Sigma}) \\
&\qquad + \eta_t^2 \mathbb{E}_{\mathrm{batch}} \Big[(\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma}) C (\widehat{\Sigma}_t^{\mathrm{wr}} - \widehat{\Sigma}) \mid \mathcal{F}_{t-1}^{\mathrm{wr}} \Big] \\
&\quad \succeq (I - \eta_t \widehat{\Sigma}) C (I - \eta_t \widehat{\Sigma}).
\end{aligned}$$

Unrolling this lower recursion and then taking expectation over the response noises gives, by Lemma I.2,

$$\mathbb{E}_{\varepsilon,\mathrm{batch}}[\Delta_L \Delta_L^\top \mid S, w^*, X] \succeq \sigma^2 \beta_\xi^{\mathrm{WSD}} \sum_{t=1}^{L} \rho_t^{\mathrm{wr}} \eta_t^2 \prod_{s=t+1}^{L} (I - \eta_s \widehat{\Sigma}) \widehat{\Sigma} \prod_{s=t+1}^{L} (I - \eta_s \widehat{\Sigma}).$$

Multiplying by Σ, taking the trace, and using equation 45 proves the claim. □

Lemma I.11 (Source-condition bound for dynamic with-replacement batches)**.** *Assume the conditions of Theorem 3.2. Then, with probability at least* $1 - \exp(-\Omega(M))$ *over* S*,*

$$\sum_{t=1}^{L} \rho_t^{\mathrm{wr}} \kappa_t^{\mathrm{mp}} \lesssim \mathbb{E}[\mathrm{Fluc}^{\mathrm{wr}}_{B_{1:L}}] \lesssim \log N \sum_{t=1}^{L} \rho_t^{\mathrm{wr}} \kappa_t^{\mathrm{mp}}, \tag{53}$$

$$\rho^{\mathrm{wr}}_{\mathrm{eff,WSD}} \mathcal{K}_{\mathrm{d}} \lesssim \mathbb{E}[\mathrm{Fluc}^{\mathrm{wr}}_{B_{1:L}}] \lesssim \rho^{\mathrm{wr}}_{\mathrm{eff,WSD}} \log N \mathcal{K}_{\mathrm{d}}. \tag{54}$$

In particular, the fluctuation rate is sharp up to the logarithmic factor.

Proof. Fix S on its power-law event and first fix w^*. The horizon restriction implies $A_L \lesssim N^{(1-\varepsilon)a}$ for some fixed $\varepsilon \in (0, 1 - 1/a)$. Let $\mathcal{E}$ be the intersection of the empirical spectral event in Lemma I.5 and the regularized covariance event in Lemma K.1. Then $\mathbb{P}_Z(\mathcal{E}^c \mid S) \le Ce^{-cN}$, and $R_p \lesssim 1$ on $\mathcal{E}$. By the conditional upper bound and equation 51,

$$\mathbb{E}_Z[\mathbf{1}_{\mathcal{E}} \mathbb{E}_{\mathrm{res}}[\mathrm{Fluc} \mid \mathcal{D}_z] \mid S, w^*] \lesssim (\sigma^2 + \|w^*\|_H^2) \log N \sum_t \rho_t^{\mathrm{wr}} \kappa_t^{\mathrm{mp}}.$$

The safeguard also controls $\mathcal{E}^c$: use the all-sample bound in the preceding proof, $R_p \le 1 + A_L \|\Sigma\|_2$, and

$$\widehat{\kappa}_t^{(1)} + A_L^{-1} \widehat{\kappa}_t^{(0)} \le \gamma_t^2 \left[\mathrm{tr}(\widehat{\Sigma}^2) + A_L^{-1} \mathrm{tr}(\widehat{\Sigma}) \right].$$

Gaussian covariance moments, equation 51, and Hölder's inequality bound the exceptional contribution by

$$C(\sigma^2 + \|w^*\|_H^2)(1 + A_L)\log(2N)e^{-cN}\sum_t \rho_t^{\mathrm{wr}}\gamma_t^2.$$

Here fixed moments of the two empirical traces are bounded because $\mathrm{tr}(\Sigma) \lesssim 1$. By Lemma A.3, $\kappa_t^{\mathrm{mp}} \gtrsim \gamma_t^2(1 + A_L)^{1/a-2}$. Since $A_L \lesssim N^a$, the exceptional contribution is absorbed into the preceding upper bound, uniformly in the batch schedule. Finally average over w^*, using $\mathbb{E}\|w^*\|_H^2 \asymp \sum_i i^{-b} \lesssim 1$. Only the second-moment source condition is used; the higher moments above were taken over features with w^* fixed.

For the lower bound, condition on a sketch in the power-law event. Gaussian concentration and a union bound give

$$C_A = \max_{i\in[N]} \|Sx_i\|_2^2 \lesssim \log N$$

with probability at least $1 - N^{-c}$ over the covariates. If the constant in $A_L \log N \le c_0 N$ is sufficiently small, this event implies $\beta_\xi^{\mathrm{WSD}} \ge c_\xi$ for an absolute $c_\xi > 0$. Intersect it with the event in Lemma I.6. On this intersection, the safeguard is inactive, so the empirical lower bound uses the nominal rates without any random weight being taken outside an expectation. Lemma I.10 gives

$$\mathbb{E}_\varepsilon\left[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wr}} \mid S, w^*, X\right] \gtrsim \sum_{t=1}^{L} \rho_t^{\mathrm{wr}}\kappa_t^{\mathrm{mp}}.$$

The intersection has probability bounded below by an absolute constant, so taking expectation over the covariates and w^* proves the lower half of equation 53.

Finally,

$$\sum_t \rho_t^{\mathrm{wr}}\kappa_t^{\mathrm{mp}} = \rho_{\mathrm{eff,WSD}}^{\mathrm{wr}} \sum_t \kappa_t^{\mathrm{mp}},$$

and the sharp total-mass estimate equation 49 proves equation 54. □

Remark I.2 (Constant-batch sharp rate)**.** If $B_t \equiv B$, then $\rho_t^{\mathrm{wr}} \equiv 1/B$ and $\rho_{\mathrm{eff,WSD}}^{\mathrm{wr}} = 1/B$ algebraically. Consequently,

$$\frac{\mathcal{K}_{\mathrm{d}}}{B} \lesssim \mathbb{E}[\mathrm{Fluc}_B^{\mathrm{wr}}] \lesssim \frac{\log N}{B}\mathcal{K}_{\mathrm{d}}.$$

J Fluctuation Error under Multi-pass WSD-SGD without Replacement

This section treats the deterministic schedule $B_1, \dots, B_L$ when the batch at each iteration is sampled uniformly without replacement and the learning rates follow the multi-pass WSD schedule. Define

$$\rho_t^{\mathrm{wor}} := \frac{N - B_t}{B_t(N-1)}, \qquad \mathcal{F}_t^{\mathrm{wor}} := \mathcal{D}_{\mathrm{data}} \vee \sigma(I_1, \dots, I_t).$$

The fluctuation process is

$$\Delta_t^{\mathrm{wor}} = (I - \eta_t\widehat{\Sigma}_t^{\mathrm{wor}})\Delta_{t-1}^{\mathrm{wor}} + \eta_t\xi_t^{\mathrm{wor}}, \qquad \Delta_0^{\mathrm{wor}} = 0,$$

and

$$\mathbb{E}[\xi_t^{\mathrm{wor}} \mid \mathcal{F}_{t-1}^{\mathrm{wor}}] = 0.$$

Set

$$\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wor}} := \mathbb{E}_{\mathrm{batch}}\|\Sigma^{1/2}\Delta_L^{\mathrm{wor}}\|_2^2.$$

Lemma J.1 (Finite-population covariance identity)**.** *Let $v_1, \dots, v_N$ be vectors in a finite-dimensional inner-product space and suppose $N^{-1}\sum_i v_i = 0$. If $I \subset [N]$ is uniform among subsets of size B, then*

$$\mathbb{E}_I\left[\left(\frac{1}{B}\sum_{i\in I} v_i\right)\left(\frac{1}{B}\sum_{i\in I} v_i\right)^\top\right] = \frac{N-B}{B(N-1)}\frac{1}{N}\sum_{i=1}^{N} v_i v_i^\top.$$

The identity also holds for centered matrices V_i in the form

$$\mathbb{E}_I\left[\left(\frac{1}{B}\sum_{i\in I}V_i\right)C\left(\frac{1}{B}\sum_{i\in I}V_i\right)\right]=\frac{N-B}{B(N-1)}\frac{1}{N}\sum_{i=1}^{N}V_iCV_i$$

for every deterministic PSD matrix C.

Proof. Expand the two sums and use

$$\mathbb{P}(i\in I)=\frac{B}{N},\qquad \mathbb{P}(i,j\in I)=\frac{B(B-1)}{N(N-1)}\quad (i\neq j),$$

together with $\sum_{i\neq j}v_iv_j^\top=-\sum_i v_iv_i^\top$. The matrix statement is the same calculation applied to $V_iC^{1/2}$. □

For the single-sample perturbations $\zeta_t(i)$ from Appendix I,

$$\xi_t^{\mathrm{wor}}=\frac{1}{B_t}\sum_{i\in I_t}\zeta_t(i).$$

Conditionally on $\mathcal{F}_{t-1}^{\mathrm{wor}}$, the family $(\zeta_t(i))_{i=1}^N$ is deterministic and centered. Therefore

$$\mathbb{E}[\xi_t^{\mathrm{wor}}(\xi_t^{\mathrm{wor}})^\top\mid\mathcal{F}_{t-1}^{\mathrm{wor}}]=\rho_t^{\mathrm{wor}}G_t. \tag{55}$$

Likewise, with $Z_i=Sx_ix_i^\top S^\top$,

$$\begin{aligned}&\mathbb{E}[(\widehat{\Sigma}_t^{\mathrm{wor}}-\widehat{\Sigma})C(\widehat{\Sigma}_t^{\mathrm{wor}}-\widehat{\Sigma})\mid\mathcal{F}_{t-1}^{\mathrm{wor}}]\\&\quad=\rho_t^{\mathrm{wor}}\frac{1}{N}\sum_{i=1}^{N}(Z_i-\widehat{\Sigma})C(Z_i-\widehat{\Sigma}).\end{aligned} \tag{56}$$

In particular,

$$\mathbb{E}[(\widehat{\Sigma}_t^{\mathrm{wor}}-\widehat{\Sigma})^2\mid\mathcal{F}_{t-1}^{\mathrm{wor}}]\preceq\rho_t^{\mathrm{wor}}C_A\widehat{\Sigma}.$$

Corollary J.1 (Dynamic fluctuation bounds without replacement)**.** *Under the assumptions of Lemma I.9, on its empirical spectral event and with the residual expectation and F_{mp} defined there,*

$$\mathbb{E}_{\mathrm{res}}[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wor}}\mid\mathcal{D}_z]\lesssim F_{\mathrm{mp}}\sum_{t=1}^{L}\rho_t^{\mathrm{wor}}\kappa_t^{\mathrm{mp}}.$$

Under the source-condition assumptions of Lemma I.11,

$$\sum_{t=1}^{L}\rho_t^{\mathrm{wor}}\kappa_t^{\mathrm{mp}}\lesssim\mathbb{E}[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wor}}]\lesssim\log N\sum_{t=1}^{L}\rho_t^{\mathrm{wor}}\kappa_t^{\mathrm{mp}}, \tag{57}$$

$$\rho_{\mathrm{eff},\mathrm{WSD}}^{\mathrm{wor}}\mathcal{K}_{\mathrm{d}}\lesssim\mathbb{E}[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wor}}]\lesssim\rho_{\mathrm{eff},\mathrm{WSD}}^{\mathrm{wor}}\log N\mathcal{K}_{\mathrm{d}}. \tag{58}$$

Proof. Equations equation 55 and equation 56 verify the two ρ_t-weighted covariance assumptions in Lemma I.3. The remaining conditions follow pathwise from

$$0\preceq\widehat{\Sigma}_t^{\mathrm{wor}}\preceq C_AI,\qquad(\widehat{\Sigma}_t^{\mathrm{wor}})^2\preceq C_A\widehat{\Sigma}_t^{\mathrm{wor}},$$

and from independence of the batches across iterations, using rates $\eta_t\le\gamma_{\mathrm{safe}}$ and $\gamma_{\mathrm{safe}}C_A\le1/8$. Averaging the residual covariance as in Lemma I.8 is compatible with the same finite-population identity. The proof of Lemmas I.9 and I.11 then applies with ρ_t^{wr} replaced by ρ_t^{wor}. For the lower bound, the finite-population identity gives

$$\mathbb{E}_\varepsilon\big[\mathbb{E}_{\mathrm{batch}}\big[\xi_t^{\mathrm{wor}}(\xi_t^{\mathrm{wor}})^\top\mid\mathcal{F}_{t-1}^{\mathrm{wor}}\big]\big|S,w^*,X\big]\succeq\rho_t^{\mathrm{wor}}\sigma^2\beta_\xi^{\mathrm{WSD}}\widehat{\Sigma}$$

by Lemma I.2. The transition comparison in the proof of Lemma I.10 is unchanged. Thus

$$\mathbb{E}_\varepsilon\left[\mathrm{Fluc}_{B_{1:L}}^{\mathrm{wor}}\mid S,w^*,X\right]\ge\sigma^2\beta_\xi^{\mathrm{WSD}}\sum_{t=1}^{L}\rho_t^{\mathrm{wor}}\widehat{\kappa}_t^{(\times)}.$$

Applying Lemma I.6 on its $r=1$ event, the $C_A\lesssim\log N$ event, and equation 49 proves both lower bounds. □

Remark J.1 (Full-batch consistency)**.** If $B_t = N$ for every t, then $\rho_t^{\mathrm{wor}} = 0$ and $\widehat{\Sigma}_t^{\mathrm{wor}} = \widehat{\Sigma}$, $\widehat{b}_t^{\mathrm{wor}} = \widehat{b}$. Thus

$$u_t^{\mathrm{wor}} = \theta_t \qquad \text{for all } t,$$

and the fluctuation error is exactly zero.

Remark J.2 (Constant-batch sharp rate)**.** If $B_t \equiv B$, then

$$\rho_t^{\mathrm{wor}} \equiv \frac{N-B}{B(N-1)} = \rho_{\mathrm{eff,WSD}}^{\mathrm{wor}}.$$

Consequently, in the sharp regime of the theorem,

$$\frac{N-B}{B(N-1)}\mathcal{K}_{\mathrm{d}} \lesssim \mathbb{E}[\mathrm{Fluc}_B^{\mathrm{wor}}] \lesssim \frac{N-B}{B(N-1)} \log N \mathcal{K}_{\mathrm{d}}.$$

K Collected Auxiliary Lemmas

This section collects the spectral, moment, and concentration ingredients used repeatedly across the appendix proofs for equation 3 and for the auxiliary multi-pass analyses in Appendices I and J.

We use the common notation from Section A, especially Section A.2. Let $(\lambda_i)_{i\ge 1}$ denote the eigenvalues of H in non-increasing order. For integers $0 \le k_* \le k \le \infty$, define

$$\Sigma_{k_*:k} := S_{k_*:k} H_{k_*:k} S_{k_*:k}^\top, \qquad \Sigma_{k:\infty} := S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top.$$

For any symmetric PSD matrix A, we write $\mu_1(A) \ge \mu_2(A) \ge \cdots$ for its eigenvalues.

K.1 General concentration lemmas

Lemma K.1 (Covariance replacement)**.** *Let $\lambda > 0$. If*

$$\sum_{i=1}^{M} \frac{\mu_i(\Sigma)}{\mu_i(\Sigma)+\lambda} \le \frac{N}{4},$$

then with probability at least $1 - \exp(-\Omega(N))$ over the sample,

$$\left\|(\widehat{\Sigma}+\lambda I)^{-1/2}(\Sigma+\lambda I)^{1/2}\right\|_2 \le 3.$$

Moreover,

$$\mathbb{E}_X \left\|(\widehat{\Sigma}+\lambda I)^{-1/2}(\Sigma+\lambda I)^{1/2}\right\|_2^4 \le 100 + \exp(-cN)\frac{\|\Sigma\|_2^2}{\lambda^2}$$

for some absolute constant $c > 0$.

Lemma K.2 (Two-sided regularized covariance comparison)**.** *Condition on S, and suppose $z = Sx \sim \mathcal{N}(0,\Sigma)$. For $\lambda > 0$, let*

$$d_\lambda := \mathrm{tr}\big(\Sigma(\Sigma+\lambda I)^{-1}\big).$$

There exist absolute constants $c_0, c_1, c_2, c_3 > 0$ such that, whenever $d_\lambda \le c_0 N$, with probability at least $1 - 2\exp(-c_1 N)$ over the sample,

$$c_2(\Sigma+\lambda I) \preceq \widehat{\Sigma}+\lambda I \preceq c_3(\Sigma+\lambda I). \tag{59}$$

Moreover, if equation 59 holds at λ_0, then, after changing only the absolute constants, it holds simultaneously for every $\lambda \ge \lambda_0$.

Proof. Put

$$g_i := (\Sigma+\lambda I)^{-1/2} z_i, \qquad K := \mathbb{E}[g_i g_i^\top] = (\Sigma+\lambda I)^{-1/2}\Sigma(\Sigma+\lambda I)^{-1/2}.$$

Then $0 \preceq K \preceq I$ and $\mathrm{tr}(K) = d_\lambda$. The standard Gaussian covariance concentration bound, applied to $N^{-1}\sum_i g_i g_i^\top$, gives

$$\left\|\frac{1}{N}\sum_{i=1}^{N} g_i g_i^\top - K\right\|_2 \le \frac{1}{2}$$

with probability at least $1-2\exp(-c_1 N)$, provided the absolute constant in $d_\lambda \le c_0 N$ is sufficiently small. Conjugating the inequalities $-I/2 \preceq N^{-1}\sum_i g_i g_i^\top - K \preceq I/2$ by $(\Sigma+\lambda I)^{1/2}$, and then adding λI, proves equation 59.

For $\lambda \ge \lambda_0$, add $(\lambda-\lambda_0)I$ to all three matrices in the comparison at λ_0. Since the comparison constants can be enlarged toward one, the same two-sided estimate follows uniformly for every larger regularization level. □

Lemma K.3 (Sketched fourth-moment and residual covariance bounds)**.** *Assume Assumptions 1.A and 1.B. Condition on the sketch matrix S and the target w^*, and let*

$$z := Sx, \qquad \Sigma := SHS^\top, \qquad u^* := \Sigma^{-1}SHw^*.$$

Then there exists an absolute constant $\alpha > 0$ such that the following hold.

(i) *For every PSD matrix $A \in \mathbb{R}^{M\times M}$,*

$$\mathbb{E}[zz^\top A zz^\top] \preceq \alpha\ \mathrm{tr}(\Sigma A)\Sigma.$$

(ii) *Writing*

$$\widetilde{\sigma}^2(w^*) := 2\big(\sigma^2 + \alpha\|w^*\|_H^2\big),$$

one has

$$\mathbb{E}\big[(y - z^\top u^*)^2 zz^\top\big] \preceq \widetilde{\sigma}^2(w^*)\,\Sigma.$$

Under Gaussian design, one may take $\alpha = 3$.

Proof. Since $z \sim \mathcal{N}(0,\Sigma)$, Isserlis' formula gives

$$\mathbb{E}[zz^\top A zz^\top] = \mathrm{tr}(\Sigma A)\Sigma + 2\Sigma A\Sigma.$$

For $A \succeq 0$, $\Sigma A\Sigma \preceq \mathrm{tr}(\Sigma A)\Sigma$, so part (i) follows with $\alpha = 3$.

For part (ii), write

$$\varepsilon := y - x^\top w^*, \qquad q := w^* - S^\top u^*.$$

Then $y - z^\top u^* = \varepsilon + x^\top q$. By Assumption 1.B,

$$\mathbb{E}[\varepsilon^2 zz^\top] = \mathbb{E}\big[\mathbb{E}[\varepsilon^2 \mid x, w^*] zz^\top\big] = \sigma^2\Sigma.$$

Moreover, another application of the Gaussian fourth-moment identity yields

$$\mathbb{E}[(x^\top q)^2 zz^\top] \preceq \alpha\|q\|_H^2\Sigma.$$

The definition $u^* = \Sigma^{-1}SHw^*$ makes $S^\top u^*$ the H-orthogonal projection of w^* onto $\mathrm{range}(S^\top)$, and hence $\|q\|_H \le \|w^*\|_H$. Therefore, using $(r+s)^2 \le 2r^2 + 2s^2$,

$$\mathbb{E}[(y - z^\top u^*)^2 zz^\top] \preceq 2\big(\sigma^2 + \alpha\|w^*\|_H^2\big)\Sigma,$$

as claimed. □

Lemma K.4 (Head–tail eigenvalue comparison)**.** *There exists an absolute constant $c > 1$ such that for every $0 \le k \le M$, with probability at least*

$$1 - \exp(-\Omega(M)) - \exp(-\Omega(k)),$$

we have for every $j \in [M]$,

$$\left|\mu_j(\Sigma) - \lambda_j - \frac{1}{M}\sum_{i>k}\lambda_i\right| \le c\left(\frac{k}{M}\lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right).$$

In particular, if $k \le M/c^2$, then the term $\frac{k}{M}\lambda_j$ can be absorbed into the left-hand side.

Lemma K.5 (Tail concentration). *For any $k \geq 0$, with probability at least $1-\delta$,*

$$\left\| \Sigma_{k:\infty} - \frac{1}{M} \sum_{i>k} \lambda_i \, I_M \right\|_2 \lesssim \lambda_{k+1} \left(1 + \frac{\log(1/\delta)}{M}\right) + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M} \left(1 + \frac{\log(1/\delta)}{M}\right)}.$$

In particular, with probability at least $1 - \exp(-\Omega(M))$,

$$\left\| \Sigma_{k:\infty} - \frac{1}{M} \sum_{i>k} \lambda_i \, I_M \right\|_2 \lesssim \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}}.$$

Lemma K.6 (Head concentration). *For any $k \geq 1$, with probability at least $1-\delta$,*

$$\left|\mu_j(\Sigma_{0:k}) - \lambda_j\right| \lesssim \frac{k + \log(1/\delta)}{M} \lambda_j, \qquad j \leq k.$$

In particular, with probability at least $1 - \exp(-\Omega(k))$,

$$\left|\mu_j(\Sigma_{0:k}) - \lambda_j\right| \lesssim \frac{k}{M} \lambda_j, \qquad j \leq k.$$

Lemma K.7 (Head–tail resolvent estimate). *Fix an integer $k \leq M/3$ such that $\operatorname{rank}(H) \geq k + M$, and define*

$$A_k := S_{k:\infty} H_{k:\infty} S_{k:\infty}^\top, \qquad \Sigma = S_{0:k} H_{0:k} S_{0:k}^\top + A_k.$$

Then, with probability at least

$$1 - \exp(-\Omega(M)),$$

one has

$$\|\Sigma^{-1} S_{0:k} H_{0:k}\|_2 \lesssim \frac{\mu_{M/2}(A_k)}{\mu_M(A_k)}.$$

K.2 Power-law auxiliary lemmas

Lemma K.8 (Power-law spectrum of the sketched covariance). *Suppose the population spectrum obeys $\lambda_j \asymp j^{-a}$ for some $a > 1$. Then, with probability at least $1 - \exp(-\Omega(M))$,*

$$\mu_j(\Sigma) \asymp j^{-a}, \qquad j \in [M].$$

Lemma K.9 (Tail spectral ratio under power law). *Suppose $\lambda_j \asymp j^{-a}$ with $a > 1$. Then there exists an a-dependent constant $c > 0$ such that for any $k \geq 0$,*

$$\frac{\mu_{M/2}(\Sigma_{k:\infty})}{\mu_M(\Sigma_{k:\infty})} \leq c$$

with probability at least $1 - \exp(-\Omega(M))$.

Lemma K.10 (Source-condition approximation bound). *Suppose the source condition with exponent $b > 1$ holds. Then, with probability at least $1 - \exp(-\Omega(M))$ over the sketch matrix,*

$$M^{1-b} \lesssim \mathbb{E}_{w^*}[\text{Approx}] \lesssim M^{1-b}.$$

The hidden constants depend only on the source exponents.

Lemma K.11 (Empirical spectrum under power law). *Assume $\mu_j(\Sigma) \asymp j^{-a}$ for $j \in [M]$. Then there exist a-dependent constants $c, c_1, c_2 > 0$ such that, conditioned on S, with probability at least $1 - \exp(-\Omega(N))$ over the sample,*

$$c_1 j^{-a} \leq \mu_j(\widehat{\Sigma}) \leq c_2 j^{-a}, \qquad j \leq \min\{M, N/c\},$$

and

$$\mu_j(\widehat{\Sigma}) \lesssim j^{-a}, \qquad j \leq \min\{M, N\}.$$

L Experimental setup

The main-text experiments use ambient dimension $d = 10^4$, sketch dimension $M = 64$, spectral exponent $a = 2$, source exponent $b = 1.5$, peak stepsize $\gamma = 0.05$, and noise standard deviation $\sigma = 1$. The one-pass update horizon is $T = 64$; the multi-pass experiments use $N = L = 512$. Each stochastic comparison uses $R = 300$ repetitions, with base random seed 123. The sample-processing budgets for the two sweeps are specified below. Experiment 4 uses $D_{\text{grad}} = 4096$.

L.1 Synthetic sketched-regression model

The ambient covariance is diagonal with $\lambda_i = i^{-a}$. Independently for each coordinate, we draw

$$w_i^* \sim \mathcal{N}\left(0, \frac{i^{-b}}{\lambda_i}\right),$$

and independently draw the Gaussian sketch entries from $\mathcal{N}(0, 1/M)$. We then form the sketched covariance $\Sigma = SHS^\top$ and the joint Gaussian law of the sketched feature $z = Sx$ and clean signal $\langle x, w^* \rangle$. Each dataset is sampled directly from this joint law with

$$y = \langle x, w^* \rangle + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

The sketch and target are fixed using seed 123. The dataset and optimization randomness are refreshed over $R = 300$ repetitions, and all schedules at a common budget use the same dataset within each repetition.

L.2 WSD influence and batch schedules

For a horizon J, we set the warmup endpoint to $\lfloor 0.1J \rfloor$, the stable endpoint to $\lfloor 0.5J \rfloor$, and use the WSD schedule in Section 2. From the eigenvalues of the simulated Σ, we compute

$$\kappa_t^{(J)} = \gamma_t^2 \sum_{j=1}^{M} \mu_j^2 \prod_{s=t+1}^{J} (1 - \gamma_s \mu_j)^2.$$

This is done once for $T = 64$ and once for $L = 512$.

At each total budget, four integer schedules are constructed from the weights

$$\begin{array}{llll} \sqrt{\kappa_t^{(J)}} & \text{(oracle)}, & 1 & \text{(constant)}, \\ J - t + 1 & \text{(early-large/late-small)}, & t & \text{(early-small/late-large)}. \end{array}$$

Each update receives at least one sample. The remaining budget is allocated proportionally to the corresponding weights and rounded by a largest-remainder rule that preserves the exact total; multi-pass batches are additionally capped at N. The oracle weights are evaluated as $\sqrt{\kappa_t + \epsilon_\kappa}$, with $\epsilon_\kappa = 10^{-12} \max_s \kappa_s + 10^{-30}$, to avoid zero numerical weights. This positive integer allocation approximates the continuous square-root rule; it is not asserted to be the exact solution of the constrained integer optimization problem.

L.3 Experiments 1 and 2: dynamic-schedule scaling

For one pass, we fix $T = 64$ and vary

$$N \in \{96, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448, 480, 512\}.$$

Every schedule satisfies $\sum_{t=1}^{T} B_t = N$, and every sample is used once. If $u_{T,r}^{\mathrm{op}}$ is the output on repetition r, the reported centered variance is

$$\widehat{\mathrm{Var}} = \frac{1}{R-1} \sum_{r=1}^{R} \left\| \Sigma^{1/2} \left(u_{T,r}^{\mathrm{op}} - \frac{1}{R} \sum_{q=1}^{R} u_{T,q}^{\mathrm{op}} \right) \right\|_2^2.$$

Figure 1(a) indexes this quantity by

$$\frac{1}{B_{\mathrm{eff}}} = \frac{\sum_t \kappa_t^{(T)} / B_t}{\sum_t \kappa_t^{(T)}}.$$

The dotted reference is a single fitted multiple of B_{eff}^{-1} across all schedules and budgets. For measured values y_p and theoretical factors $f_p > 0$, the fitted multiplier is $\exp(|\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} \log(y_p / f_p))$, where $\mathcal{P}$ contains the points with $y_p > 0$. This fits the vertical scale, not the predicted exponent.

For multi-pass WSD-SGD, we fix $N = L = 512$ and use

$$D_{\mathrm{grad}} \in \{512, 640, 768, 1024, 1280, 1536, 2048, 2560, 3072, 3584, 4096\}.$$

Each mini-batch is sampled uniformly without replacement from the fixed dataset, independently across updates. On the same dataset and WSD trajectory, we compute the full-batch reference θ_L and report

$$\widehat{\mathrm{Fluc}} = \frac{1}{R}\sum_{r=1}^{R} \left\| \Sigma^{1/2}\left(u_{L,r}^{\mathrm{mp}} - \theta_{L,r}\right)\right\|_2^2.$$

Figure 1(b) indexes this quantity by

$$\rho_{\mathrm{eff}} = \frac{\sum_t \rho_t \kappa_t^{(L)}}{\sum_t \kappa_t^{(L)}}, \qquad \rho_t = \frac{N - B_t}{B_t(N-1)}.$$

The dotted reference is one fitted multiple of ρ_{eff}, using the same fitting rule as for one pass. The displayed stochastic errors are point estimates over the repetitions; no confidence intervals are shown.

L.4 Experiment 3: relative allocation comparison

At $N = 512$ and $D_{\mathrm{grad}} = 4096$, we normalize both the measured stochastic errors and their influence-weighted predictions by the constant schedule. The resulting ratios are displayed in Figure 1(c). This comparison includes the integer approximation to the oracle rule in Corollary 3.1.

L.5 Experiment 4: WSD influence profile

For $L = 512$ and $D_{\mathrm{grad}} = 4096$, the warmup, stable, and decay phases are iterations 1:51, 52:256, and 257:512. We normalize γ_t, κ_t, and $\sqrt{\kappa_t}$ by their respective maxima and plot them with the oracle integer schedule and the constant batch size $D_{\mathrm{grad}}/L = 8$. The oracle schedule sums to 4096, ranges from 1 to 23, and reaches its maximum at iteration 256, where κ_t also peaks, as shown in Figure 1(d).

M Limitations and Broader Effects

Limitations. The results concern sketched linear regression with Gaussian features and sketching, independent homoscedastic label noise, power-law covariance decay, and an average-case source condition. They do not establish analogous laws for general model misspecification, heavy-tailed or dependent observations, non-Gaussian sketches, or feature learning in nonlinear networks. The two-horizon conclusions rely on the specified WSD schedule and phase-length conditions. The multi-pass method and its GD reference use a common, feature-dependent stepsize safeguard; the analysis does not establish the same unconditional bounds for an unguarded method. The result also imposes restrictions on the horizon relative to sample size and model dimension; its fluctuation bounds match only up to a $\log N$ factor. The simplified pointwise expression for final-risk influence in Appendix A.8 applies before the spectral cutoff reaches the model dimension.

The batch schedules are deterministic. The square-root allocation optimizes the nominal influence-weighted stochastic proxy at a fixed update horizon and sample-processing budget, using the population sketched covariance. It is therefore an oracle prescription, not a data-adaptive procedure with an estimated spectrum or a joint optimization of horizon, learning rate, and batch size. The experiments use an integer approximation to this rule. Moreover, changing batch size under a fixed budget can change the number of updates, so the fixed-horizon invariance of the deterministic terms does not extend to every compute-constrained comparison. Multi-pass sampling is without replacement within each batch and independent across updates; it does not describe random reshuffling across epochs.

The four experiments are synthetic illustrations at a single pair of spectral and source exponents and a fixed sketch and target. Repetitions vary the dataset and optimization randomness, not the sketch or target. Thus they do not test variability across those latter draws or validate all asymptotic regimes. The fitted reference curves assess scaling shape rather than absolute constants, and no confidence intervals are displayed. The experiments do not measure wall-clock time, communication cost, memory use, energy consumption, or performance on real-world tasks.

Broader effects. The separation of learning and noise-memory horizons may help motivate training strategies that allocate samples to updates where noise reduction matters most. Such improvements could reduce wasted computation, but the present results do not establish practical speedups or environmental benefits. A principal risk is applying these stylized laws as universal training rules without accounting for distribution shift, hardware costs, or application-specific fairness and safety requirements. More efficient optimization is also a general-purpose capability whose downstream effects depend on its use. The work should therefore be treated as a theoretical foundation for further validation, not as evidence of deployment safety or reliability.