LINEAR RNN SCALING LAWS: WHEN LONGER SEQUENCES BEAT MORE SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Empirical scaling laws for autoregressive language models relate prediction loss to model size, data size, and optimization compute, but their theoretical origin is still poorly understood in sequential pretraining settings. We study this question in a tractable teacher–student model where a stable latent linear RNN generates trajectories and a sketched linear recurrent student is trained by safeguarded full-batch WSD gradient descent on next-token prediction. The sketch dimension M plays the role of model size, while N independent trajectories of length P provide the training tokens. We allow the innovation and initialization covariances to have different power-law exponents α and θ. The induced design spectrum produces explicit approximation, optimization, and statistical scaling laws separated by spectral crossovers. When $\theta \geq \alpha$, the original one-scale rates $M^{1-\beta_\alpha}$, $R^{(1-\beta_\alpha)/\alpha}$, and $(NP)^{-1}\min\{M, R^{1/\alpha}\}$ are recovered. When $\alpha - 2r \leq \theta < \alpha$, the heavier initialization tail changes the rates beyond P-dependent model and optimization crossovers. The proof uses a covariance event only internally and a globally safeguarded step size on its complement. The variance retains the factor $(NP)^{-1}$, while sequence length also suppresses the initialization transient, so N and P cease to be fully interchangeable in the two-scale regime.

1 INTRODUCTION

Empirical language-model loss follows predictable power laws in model size, data, and compute, making scaling laws central to pretraining decisions (Hestness et al., 2017; Kaplan et al., 2020; Henighan et al., 2020; Rae et al., 2021; Hoffmann et al., 2022; Rosenfeld et al., 2020). Their theoretical origin is less clear. Existing analyses explain such laws in regression, kernel, random-feature, and data-reuse models (Sharma & Kaplan, 2022; Bahri et al., 2024; Bordelon et al., 2024; Lin et al., 2024; 2025), but largely abstract away the sequential structure of autoregressive pretraining.

We treat autoregressive pretraining (AP) and in-context learning (ICL) as complementary but distinct abstractions. From our own interpretation, AP learns reusable token-transition dynamics that iteratively compress tokens into context representations and later generate outputs token by token; ICL maps encoded demonstrations and an encoded query to an answer-sequence seed, which AP-trained decoding dynamics then expand. Figure 1 illustrates this view.

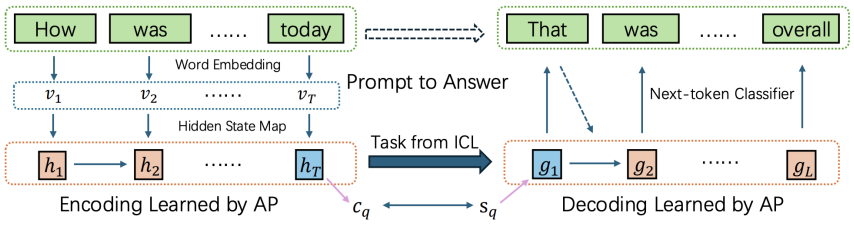


Figure 1: Our conceptual interpretation of AP and ICL. AP learns reusable dynamics for iteratively encoding token sequences and autoregressively decoding outputs. ICL operates on the resulting representations, mapping demonstrations and a query context to an answer-sequence seed.

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.

Theoretical ICL has been studied extensively, including interpretations as implicit Bayesian inference, function-class learning, regression or gradient descent, algorithm selection, induction-head formation, learning from structured pretraining sequences, and ICL-specific scaling laws (Xie et al., 2022; Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023; Bai et al., 2023; Olsson et al., 2022; Chan et al., 2022; Arora et al., 2025; Lyu et al., 2025). By comparison, the provable theory of AP and next-token prediction is less developed: existing work studies attention-based token retrieval, dependent-sample generalization, the emergence of ICL from autoregressive pretraining and linear dynamical sources, and scaling in linear bigram or synthetic hierarchical-language models, but does not jointly track representation dimension, trajectory count, trajectory length, optimization time, and sampling error when learning recurrent latent dynamics (Li et al., 2024; 2025; Gong et al., 2025; Riechers et al., 2025; Kunstner & Bach, 2025; Cagnetta et al., 2025).

We formalize this upstream learning problem with a linear teacher–student model. A stable latent RNN teacher generates unlabeled trajectories, a Gaussian sketch restricts the observable representation to dimension M, and a linear recurrent student is trained by safeguarded empirical WSD gradient descent. The data contain N trajectories of length P, and R denotes the effective optimization horizon. Distinct power laws for the initialization and innovation covariances induce two spectral regimes. In both, the prediction error separates into representation approximation, finite-optimization bias, and trajectory sampling variance, allowing the roles of M, N, P, and R to be read explicitly.

Our main contributions are as follows.

1. **A recurrent AP scaling framework.** We give an end-to-end analysis of a sketched linear recurrent student trained on dependent trajectories. Its expected population risk decomposes into representation approximation, finite-optimization bias, statistical variance, and a controlled cross term, exposing the separate roles of M, N, P, and R.
2. **Innovation-dominated scaling.** When $\theta \geq \alpha$, the innovation covariance has the heavier spectral tail. We recover approximation $M^{1-\beta_\alpha}$, bias $R^{(1-\beta_\alpha)/\alpha}$, and variance $(NP)^{-1}\min\{M, R^{1/\alpha}\}$, up to stability constants. Thus N and P are interchangeable at leading order, paralleling data-reuse and many-trajectory results (Lin et al., 2025; Tu et al., 2024).
3. **Initialization-dominated scaling.** When $\alpha - 2r \leq \theta < \alpha$, the initialization has the heavier tail. We identify P-dependent representation and optimization crossovers: beyond them, longer trajectories reduce approximation and bias in addition to variance. Hence trajectory length and trajectory count are no longer interchangeable in Regime II.

Notation. We collect the important notations used across the paper. For two positive-valued functions f and g, we write $f \lesssim g$ ($f = \mathcal{O}(g)$) and $f \gtrsim g$ when the inequality holds up to an absolute constant, and $f \asymp g$ (($f = \Theta(g)$)) when both comparisons hold. For vectors, $\langle u, v\rangle = u^\top v$. For matrices, $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ denote the Frobenius and operator norms, $\mathrm{Tr}(\cdot)$ denotes the trace, and $\preceq$ denotes the Loewner order on symmetric matrices. For any positive semidefinite matrix G, we use the G-weighted matrix inner product $\langle D, E\rangle_G := \mathrm{Tr}(DGE^\top)$ and norm $\|D\|_G^2 := \langle D, D\rangle_G = \|DG^{1/2}\|_F^2$. For a symmetric matrix A, $\mu_j(A)$ denotes its j-th eigenvalue in nonincreasing order.

2 Preliminary

2.1 Problem Setup

Latent linear RNN teacher. For each independent trajectory $i = 1, \ldots, N$, the latent state $x_{i,p} \in \mathbb{R}^d$ follows the linear Gaussian autoregressive model

$$x_{i,p+1} = A_* x_{i,p} + \xi_{i,p+1}, \qquad p = 0, \ldots, P-1.$$

Mathematically, this recursion is a stable Gaussian vector autoregression of order one (VAR(1); Hamilton, 1994; Lütkepohl, 2005), with i.i.d. innovations $\xi_{i,p} \sim \mathcal{N}(0, \Sigma_\xi)$, independent of the possibly nonstationary initial states $x_{i,0} \sim \mathcal{N}(0, \Sigma_0)$. Here p indexes token position and each trajectory represents an independent text sequence of length $P \geq 2$. In our language-model interpretation, $x_{i,p}$ represents latent context, $A_* \in \mathbb{R}^{d\times d}$ captures predictable recurrent dynamics, and $\xi_{i,p}$ represents fresh stochastic innovation.

Sketched observations. The learner observes only the sketched states

$$u_{i,p} := Sx_{i,p} \in \mathbb{R}^M, \qquad y_{i,p} := u_{i,p+1},$$

where $S \in \mathbb{R}^{M\times d}$ is a Gaussian sketch with independent entries $S_{\ell j} \sim \mathcal{N}(0, 1/M)$. Throughout the paper, the ambient and sketch dimensions satisfy $d \geq C_{\mathrm{amb}} M$, where $C_{\mathrm{amb}} > 1$ is a sufficiently large fixed constant, independent of M, N, P, L. We use the sketch to model the finite embedding bottleneck in language models. Actual linguistic dynamics may involve a much richer finite collection of latent semantic and syntactic factors, while a model maps each token or context into an M-dimensional embedding before processing it. The sketch dimension M therefore plays the role of the embedding/model width available to represent these latent directions. Because the trajectories are identically distributed, their positionwise covariances do not depend on the trajectory index. Define

$$\Sigma_p := \mathbb{E}[x_{i,p}x_{i,p}^\top], \qquad H := \frac{1}{P}\sum_{p=0}^{P-1} \Sigma_p,$$

and retain the exact positionwise sketched covariance and cross-covariance

$$G_p := \mathbb{E}[u_{i,p}u_{i,p}^\top] = S\Sigma_p S^\top, \qquad C_p := \mathbb{E}[y_{i,p}u_{i,p}^\top].$$

Their position averages are

$$G := SHS^\top = \frac{1}{P}\sum_{p=0}^{P-1} G_p, \qquad C := \frac{1}{P}\sum_{p=0}^{P-1} C_p.$$

Population and empirical regression. The observable population and empirical objectives are

$$\mathcal{L}_M(B) := \frac{1}{2P}\sum_{p=0}^{P-1} \mathbb{E}\|y_{i,p} - Bu_{i,p}\|_2^2,$$

$$\widehat{\mathcal{L}}_M(B) := \frac{1}{2NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} \|y_{i,p} - Bu_{i,p}\|_2^2, \qquad B \in \mathbb{R}^{M\times M}.$$

Define the empirical covariance and cross-covariance

$$\widehat{G} := \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} u_{i,p}u_{i,p}^\top, \qquad \widehat{C} := \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} y_{i,p}u_{i,p}^\top.$$

Then calculating gradients of the objectives and define B_* to be the optimizer of population,

$$\nabla\mathcal{L}_M(B) = BG - C, \qquad \nabla\widehat{\mathcal{L}}_M(B) = B\widehat{G} - \widehat{C}, \qquad B_* := CG^{-1}.$$

Thus B_* is the sketched population predictor, and the quadratic excess risk has the exact form

$$\mathcal{L}_M(B) - \mathcal{L}_M(B_*) = \frac{1}{2}\|B - B_*\|_G^2.$$

Safeguarded WSD gradient descent. For L gradient steps, choose warmup and stable-phase endpoints $1 \leq L_{\mathrm{w}} < L_{\mathrm{s}} < L$. For fixed constants $c_{\mathrm{s}}, c_{\mathrm{d}} > 0$, we use the warmup–stable–decay (WSD) schedule (Wen et al., 2025) with stability update safeguard$L_{\mathrm{s}} - L_{\mathrm{w}} \geq c_{\mathrm{s}}L,\ L - L_{\mathrm{s}} \geq c_{\mathrm{d}}L$.

$$\bar{\gamma}_t := \frac{\bar{\gamma}}{\gamma}\gamma_t, \quad \bar{\gamma} := \min\left\{\gamma, \frac{1}{2\|\widehat{G}\|_{\mathrm{op}}}\right\}, \quad \gamma_t := \begin{cases} \gamma t/L_{\mathrm{w}}, & 1 \leq t \leq L_{\mathrm{w}}, \\ \gamma, & L_{\mathrm{w}} < t \leq L_{\mathrm{s}}, \\ \gamma \exp\left(-\dfrac{(t-L_{\mathrm{s}})\log L}{L - L_{\mathrm{s}}}\right), & L_{\mathrm{s}} < t \leq L. \end{cases}$$

We define $R = \gamma L$ as the cumulative optimization horizon. Starting from $B_0 = 0$, the full-batch WSD-GD updates the learnable parameter with

$$B_t = B_{t-1} - \bar{\gamma}_t(B_{t-1}\widehat{G} - \widehat{C}), \qquad t = 1, \ldots, L.$$

The safeguard clips only the peak WSD learning rate and ensures $0 \preceq I - \bar{\gamma}_t\widehat{G} \preceq I$ for every sample. On the covariance event used in the proofs it is inactive, so $\bar{\gamma}_t = \gamma_t$; its only purpose is to control the rare complement event when taking an unconditional expectation over the training trajectories.

Figure 2 provides a schematic demonstration of the setup. Initial latent states generate N trajectories through the shared transition A_*, the Gaussian sketch produces the observed sequences $\{u_{i,p}\}$, and safeguarded WSD-GD on the empirical next-token objective returns the iterate B_L; the displayed words are illustrative rather than an additional modeling assumption.

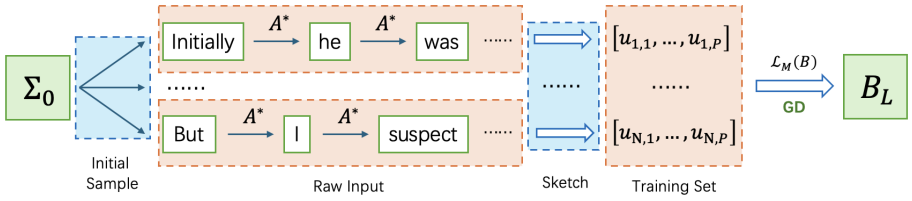


Figure 2: Schematic demonstration of the learning setup, from random initialization and latent autoregressive trajectories to sketched training sequences and the empirical WSD-GD iterate B_L.

2.2 Standing Assumptions

Assumption 1 (Positive stable teacher and source decay)**.** *The teacher transition matrix A_* is symmetric positive definite with orthonormal eigenbasis $\{e_j\}_{j=1}^d$ and respective eigenvalues $a_j \in \mathbb{R}$:*

$$A_* e_j = a_j e_j, \qquad 1 \le j \le d.$$

The teacher is uniformly stable and its eigenvalues have power-law decay:

$$0 < a_j \le \rho < 1, \qquad a_j \asymp j^{-r}, \qquad 1 \le j \le d,$$

for an exponent satisfying $0 < 2r < 1$.

This assumption says that the latent dynamics are stable and not equally strong in every eigendirection. Stability, $0 < a_j < 1$, makes old tokens gradually lose influence, while the power law for a_j^2 says that the teacher has a graded source spectrum: leading modes carry most of the predictable transition energy, and higher-index modes become progressively weaker. In language points of view, the leading modes represent dominant contextual factors that are easier to learn, while weaker modes represent finer linguistic distinctions that require more representation dimension or optimization time. Such power-law source or target spectra are standard to encode task regularity in theoretical scaling-law analyses (Bahri et al., 2024; Lin et al., 2024; 2025).

Assumption 2 (Primitive covariance power laws)**.** *In the basis of Assumption 1:*

$$\Sigma_\xi = \sum_{j=1}^d \sigma_{\xi,j} e_j e_j^\top, \qquad \Sigma_0 = \sum_{j=1}^d \sigma_{0,j} e_j e_j^\top.$$

For exponents $\alpha > 1$ and $\theta > 1$, their eigenvalues satisfy

$$\sigma_{\xi,j} \asymp j^{-\alpha}, \qquad \sigma_{0,j} \asymp j^{-\theta}, \qquad 1 \le j \le d.$$

Together with the source exponent r, assume $\theta \ge \alpha - 2r$ and define $\beta_\alpha := \alpha + 2r$, $\beta_\theta := \theta + 2r$.

In particular, $\Sigma_\xi \succ 0$ for every finite d. Upper bounds may compare this finite sum with the corresponding infinite power-law tail. Matching lower bounds use only fixed annuli of order M, which exist because $d \ge C_{\mathrm{amb}} M$. The restriction $\theta \ge \alpha - 2r$ is the one-step smoothing condition needed by the nonstationary variance lower bound. Coordinatewise,

$$a_j^2 \sigma_{0,j} \asymp j^{-(\theta+2r)} \lesssim j^{-\alpha} \asymp \sigma_{\xi,j},$$

and hence $A_* \Sigma_0 A_*^\top \preceq C_{\mathrm{sm}} \Sigma_\xi$.

When $\theta \ge \alpha$, the innovation covariance determines the tail of the averaged design covariance and the original one-scale law is recovered. When $\alpha - 2r \le \theta < \alpha$, the initialization has the heavier tail, but one application of A_* smooths it to the innovation scale. Note that $0 < 2r < 1$ implies $1 < \beta_\alpha < \alpha + 1$ and $1 < \beta_\theta < \theta + 1$. We impose the covariance power laws in the teacher eigenbasis because the Gaussian sketch is rotationally invariant: for any fixed orthogonal change of basis, the sketch distribution is unchanged. Thus only the relative spectral alignment matters, and the aligned formulation in the basis of A_* is the natural representative case for the sketching analysis (Halko et al., 2011; Woodruff, 2014).

Assumption 3 (WSD step size and effective horizon)**.** *The deterministic peak step size satisfies* $0 < \gamma \le c_\gamma$*, and the effective horizon* $R = \gamma L$ *satisfies* $R \gtrsim 1$*. Here* $c_\gamma > 0$ *is sufficiently small and depends only on the fixed primitive spectral constants, not on* M, N, P, L*.*

The constant c_γ is chosen small enough that the safeguard is inactive on the covariance and sketch events. The assumption also excludes the vacuous regime in which the nominal optimization horizon is too small for the bias and variance filters to enter their asymptotic regime.

3 Main result

We now state the scaling law imitating the preceding setup Lin et al. (2024; 2025): first separate the population risk into approximation, GD bias, GD variance, and a cross term, then bound these terms under the power-law source condition. One notable additional feature is that the samples come from stable trajectories, where the positions of input samples matter.

Let the empirical residual and the empirical WSD-GD filters be

$$\widehat{\mathcal{B}}_L(D) := D\prod_{t=1}^{L}(I - \bar{\gamma}_t \widehat{G}), \qquad \widehat{\mathcal{V}}_L(E) := \sum_{t=1}^{L} \bar{\gamma}_t E \prod_{s=t+1}^{L} (I - \bar{\gamma}_s \widehat{G}),$$

where we define an empty product to be the identity. We use the G-weighted inner product from the notation paragraph in the introduction. Recall that $\mathcal{L}(A)$ is the original full-state loss with $A \in \mathbb{R}^{d\times d}$. The GD analysis uses the observable sketched model loss $\mathcal{L}_M(B)$. The expectation inside the definition of $\mathcal{L}_M(B)$ is a population expectation over an independent test trajectory. When $B = B_L$, the additional outer expectation $\mathbb{E}_{\mathcal{D}}[\cdot \mid S]$ averages over the training trajectories $\mathcal{D}$, conditional on the Gaussian sketch S. The same convention is used for Bias, Var, and Cross.

Proposition 3.1 (Sketched-model risk decomposition for empirical WSD-GD)**.** Let B_L be the L-step empirical WSD-GD iterate initialized at $B_0 = 0$. Then the sketched-model risk decomposes as

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}_M(B_L) \mid S] = \underbrace{\frac{1}{2}\operatorname{Tr}(S\Sigma_\xi S^\top)}_{\text{Irreducible}} + \underbrace{\inf_{B\in\mathbb{R}^{M\times M}} \frac{1}{2}\left\|(BS - SA_*)H^{1/2}\right\|_F^2}_{\text{Approx}} + \underbrace{\frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{B}}_L(B_*)\|_G^2 \mid S\right]}_{\text{Bias}}$$
$$+ \underbrace{\frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{V}}_L(\widehat{E})\|_G^2 \mid S\right]}_{\text{Var}} + \underbrace{\left(-\mathbb{E}_{\mathcal{D}}\left[\left\langle \widehat{\mathcal{B}}_L(B_*), \widehat{\mathcal{V}}_L(\widehat{E})\right\rangle_G \mid S\right]\right)}_{\text{Cross}}.$$

where we use $\widehat{E} := \widehat{C} - B_*\widehat{G}$ to represent the cross-covariance residual.

Note that in this proposition, $|\,\text{Cross}\,| \le 2\sqrt{\text{Bias}\,\text{Var}} \le \text{Bias} + \text{Var}$. Consequently

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}_M(B_L) \mid S] - \frac{1}{2}\operatorname{Tr}(S\Sigma_\xi S^\top) \lesssim \text{Approx} + \text{Bias} + \text{Var}\,.$$

This proposition is the matrix-regression analogue of the population risk decomposition in Lin et al. (2025). The approximation term measures the price of restricting the ambient transition to an M-dimensional sketch, the GD-bias term measures finite optimization time, and the GD-variance term measures empirical fluctuations from the observed trajectories.

Theorem 3.1 (Two-regime scaling law for empirical WSD-GD)**.** *Suppose Assumptions 1, 2, and 3 hold, and assume the unsaturated optimization regime* $R \le c_{\text{unsat}}\left(M^{-\alpha} + P^{-1}M^{-\theta}\right)^{-1}$*, where* $c_{\text{unsat}} > 0$ *is a sufficiently small fixed constant, and define*

$$C_\rho := \frac{1+\rho}{1-\rho}, \quad \varepsilon_M := (M + R^2)\exp(-cM/2).$$

For Regime II, define $P_M := M^{\alpha-\theta}$ *and* $P_R := R^{(\alpha-\theta)/\alpha}$*. Then, with probability at least* $1 - \exp(-\Omega(M))$ *over the Gaussian sketch alone, reducible terms in Proposition 3.1 satisfy*

$$\text{Approx} = \mathcal{A}_{M,P}, \quad \text{Bias} = \mathcal{B}_{R,P} + \mathcal{O}(\varepsilon_M), \quad \text{Var} = \mathcal{V}_{R,M,P} + \mathcal{O}(\varepsilon_M).$$

From the above arguments on the cross term, we have $\text{Cross} = \mathcal{O}(\sqrt{\mathcal{B}_{R,P}\mathcal{V}_{R,M,P}} + \varepsilon_M)$*. We specify* $\mathcal{A}_{M,P}$*,* $\mathcal{B}_{R,P}$*,* $\mathcal{V}_{R,M,P}$ *in the two regimes below.*

Regime I: $\theta \geq \alpha$. *Assume* $NP \gtrsim C_\rho RM$. *Then we have*

$$\mathcal{A}_{M,P} \asymp M^{1-\beta_\alpha}, \qquad \mathcal{B}_{R,P} \asymp R^{(1-\beta_\alpha)/\alpha}, \qquad \mathcal{V}_{R,M,P} \asymp \frac{C_\rho}{NP} \min\{M, R^{1/\alpha}\}.$$

Regime II: $\alpha - 2r \leq \theta < \alpha$. *Assume* $NP \gtrsim C_\rho RM,\ N \gtrsim RM$. *Then we have*

$$\begin{cases} \mathcal{A}_{M,P} \asymp \begin{cases} P^{-1}M^{1-\beta_\theta}, & P \lesssim P_M, \\ M^{1-\beta_\alpha}, & P \gtrsim P_M, \end{cases} \\ \mathcal{B}_{R,P} \asymp \begin{cases} P^{-1}(R/P)^{(1-\beta_\theta)/\theta}, & P \lesssim P_R, \\ R^{(1-\beta_\alpha)/\alpha}, & P \gtrsim P_R, \end{cases} \\ \mathcal{V}_{R,M,P} \asymp \frac{C_\rho}{NP} \begin{cases} \min\{M, (R/P)^{1/\theta}\}, & P \lesssim P_R, \\ \min\{M, R^{1/\alpha}\}, & P \gtrsim P_R. \end{cases} \end{cases}$$

The expectation is over the full distribution of the sampled training trajectories $\mathcal{D}$, conditional only on S; in particular, it is not conditioned on the covariance event, which is used only inside the proof. The constants $c, C > 0$ depend only on the primitive covariance exponents, teacher source exponent, and constants in the standing assumptions.

Proof sketch. The proof starts from the sketched-model risk $\mathcal{L}_M(B_L)$. Expanding $Sx_{i,p+1} = SA_*x_{i,p} + S\xi_{i,p+1}$ separates the irreducible innovation loss from the noiseless transition error. The noiseless transition error is a quadratic whose minimum is Approx, and its excess around B_* is the G-norm squared. The empirical WSD-GD iterate then decomposes into a bias filter applied to B_* and a variance filter applied to the centered residual $\widehat{E} = \widehat{C} - B_*\widehat{G}$, which gives the bias–variance–cross terms and controls the cross term by Cauchy–Schwarz. The approximation bound is a Gaussian-sketch projection argument: a head–tail split plus the source condition gives the upper bound, and a non-degenerate tail projection gives the matching lower bound. For the WSD-GD bias and variance, a regularized-resolvent argument transfers the scalar filters to the noncommuting pair $(G, \widehat{G})$; the remaining population sums split at $j \asymp \kappa_{R,P}$. For the variance, the covariance and residual-score proofs split the process into initialization and innovation components. The scalar variance filter then yields $d_{\mathrm{eff}}(R, M, P) \asymp \min\{M, \kappa_{R,P}\}$.

3.1 Interpretations

Overall interpretation. The three terms describe distinct limitations of the learned recurrent predictor. The approximation term is the information lost when the latent transition is represented through only M sketched coordinates, so a larger model dimension exposes more predictive directions and decreases this error. The GD-bias term is the part of the population predictor that has not yet been reached after optimization horizon R, so additional optimization reduces bias by progressively learning lower-curvature directions. The variance term comes from fitting empirical trajectory fluctuations; increasing N or P supplies more transitions, whereas increasing R activates more weakly observed directions and can increase the effective statistical dimension. These qualitative trends agree with the formulas above.

The two regimes distinguish which primitive covariance has the heavier spectral tail. In Regime I, $\theta \geq \alpha$, the innovations decay no faster than the initial-state covariance, which may be viewed as the sequence-level topic carried by the first token. Fresh noise therefore occupies more high-index directions and transition learning must filter a spectrally broad disturbance. In Regime II, $\theta < \alpha$, the initial topic has the heavier tail while fresh innovations decay faster; the within-sequence token relationship is correspondingly more informative, but the initial transient remains visible until it is diluted by a sufficiently long trajectory.

The $\mathcal{O}(\sqrt{\mathcal{B}_{R,P}\mathcal{V}_{R,M,P}})$ term comes from the cross term, which is the interaction between the residual bias filter and the empirical variance filter, whose magnitude is controlled by Cauchy–Schwarz. The $\mathcal{O}(\varepsilon_M)$ term accounts for the exponentially rare failure of the covariance event, on which the covariance replacement holds with controlled gap.

Regime I: innovation-dominated trajectories. Regime I recovers the sketched data-reuse scaling of Lin et al. (2025), with the independent sample size replaced by the total number NP of observed transitions, up to the stability factor C_ρ. This agrees with the many-trajectory least-squares result of Tu et al. (2024), where the leading risk also behaves like an independent-sample rate based on NP. Thus approximation and bias are independent of N and P at leading order, while variance scales as $(NP)^{-1}$, making trajectory count and length interchangeable up to constants.

Regime II: initialization-dominated trajectories. In Regime II, trajectory length enters the population spectrum itself: averaging over positions attenuates the heavy initialization component, whereas adding independent trajectories primarily averages empirical fluctuations. This creates the M- and R-dependent crossovers in Theorem 3.1. The mechanism is consistent in spirit with controlled long-context experiments showing that preserving longer documents can expose more within-document dependency pairs than splitting the same tokens across shorter documents, although those experiments do not directly test our spectral regimes (Gao et al., 2025).

3.2 Compute-optimal allocation

Corollary 3.1 (Compute-optimal token and model allocation)**.** *In this subsection, $\mathcal{D} := NP \le \mathcal{T}$ denotes the number of used transitions rather than the random training dataset. Fix the compute budget $\mathcal{C} = \mathcal{D}M^2L$, and suppress fixed constants such as C_ρ. The following allocations are optimal within the theorem-controlled asymptotic regime, with sufficiently small fixed constants chosen so that the unsaturated condition of Theorem 3.1 holds.*

Regime I: $\theta \ge \alpha$. *Define $\mathcal{T}_{\mathrm{crit},I} \asymp (\gamma\mathcal{C})^{\frac{\alpha+1}{2\alpha+3}}$. If $\mathcal{T} \lesssim \mathcal{T}_{\mathrm{crit},I}$, namely when token budget is limited*

$$\mathcal{D}^\star \asymp \mathcal{T}, \quad M^\star \asymp \frac{\gamma\mathcal{C}}{\mathcal{T}^2}, \quad L^\star \asymp \frac{\mathcal{T}^3}{\gamma^2\mathcal{C}}.$$

If $\mathcal{T} \gtrsim \mathcal{T}_{\mathrm{crit},I}$, the token budget saturates,

$$\mathcal{D}^\star \asymp (\gamma\mathcal{C})^{\frac{\alpha+1}{2\alpha+3}}, \quad M^\star \asymp (\gamma\mathcal{C})^{\frac{1}{2\alpha+3}}, \quad L^\star \asymp \gamma^{-1}(\gamma\mathcal{C})^{\frac{\alpha}{2\alpha+3}}.$$

There is no leading-order preference for $P^\star$, and $N^\star = \frac{\mathcal{D}^\star}{P^\star}$.

Regime II: $\alpha - 2r \le \theta < \alpha$. *Define $\mathcal{T}_{\mathrm{crit},II} \asymp (\gamma\mathcal{C})^{\frac{2\alpha-\theta+1}{3\alpha-\theta+3}}$. If $\mathcal{T} \lesssim \mathcal{T}_{\mathrm{crit},II}$, namely when token budget is limited*

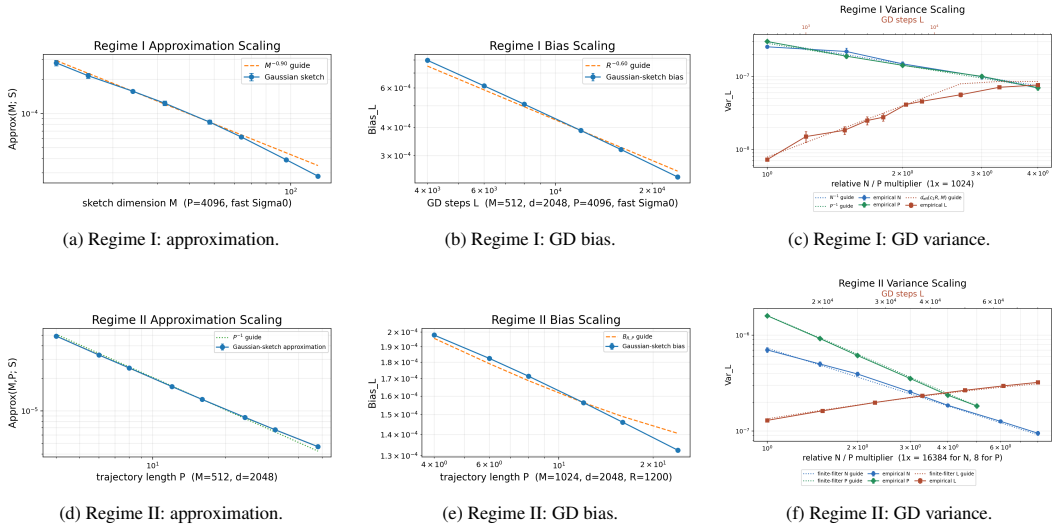
$$M^\star \asymp \left(\frac{(\gamma\mathcal{C})^2}{\mathcal{T}^3}\right)^{\frac{1}{\theta+3}}, \; L^\star = \frac{\left((\gamma\mathcal{C})^{\theta-1}\mathcal{T}^{3-\theta}\right)^{\frac{1}{\theta+3}}}{\gamma}, \; P^\star \asymp \left(\frac{\mathcal{T}^{2\theta+3}}{(\gamma\mathcal{C})^{\theta+1}}\right)^{\frac{1}{\theta+3}}, \; N^\star \asymp \left(\frac{(\gamma\mathcal{C})^{\theta+1}}{\mathcal{T}^\theta}\right)^{\frac{1}{\theta+3}}.$$

If $\mathcal{T} \gtrsim \mathcal{T}_{\mathrm{crit},II}$, the token budget saturates,

$$M^\star \asymp (\gamma\mathcal{C})^{\frac{1}{3\alpha-\theta+3}}, \quad L^\star = \frac{(\gamma\mathcal{C})^{\frac{\alpha}{3\alpha-\theta+3}}}{\gamma}, \quad P^\star \asymp (\gamma\mathcal{C})^{\frac{\alpha-\theta}{3\alpha-\theta+3}}, \quad N^\star \asymp (\gamma\mathcal{C})^{\frac{\alpha+1}{3\alpha-\theta+3}}.$$

The corollary separates a token-limited phase from a token-rich phase. Below the critical scale, every available token should be used; above it, the leading-order optimum instead increases model width and optimization horizon, while the number of used tokens saturates. **Thus, at fixed compute, more available data improve the optimal scaling only up to an explicit critical token budget; beyond that point, additional tokens need not be consumed at leading order.**

The allocation within the used token budget depends on the spectral regime. In Regime I, only the product NP matters at leading order, so trajectories may be made more numerous or longer without changing the rate. In Regime II, the initialization transient makes these allocations inequivalent. **Sequence length itself becomes a compute-optimal design variable: longer coherent trajectories are preferred within our position-averaged nonstationary objective until the initialization-induced benefit saturates.** This is not a universal training prescription under a fixed evaluation distribution. In the token-rich Regime-II phase, the optimum lies at the spectral crossovers $P^\star \asymp (M^\star)^{\alpha-\theta}$ and $R^\star \asymp (M^\star)^\alpha$.



(a) Regime I: approximation. (b) Regime I: GD bias. (c) Regime I: GD variance.

(d) Regime II: approximation. (e) Regime II: GD bias. (f) Regime II: GD variance.

Figure 3: Synthetic verification of the two-regime scaling law. The first row tests Regime I, where the initialization tail decays at least as fast as the innovation tail: approximation follows the M-power law, bias decreases with the optimization horizon, and variance scales inversely with N and P while increasing with the active dimension. The second row tests Regime II, where the heavier initialization tail produces an explicit P^{-1} transient in approximation and bias and changes the P- and R-dependence of variance. Solid curves are measured values; dashed or dotted curves are asymptotic or finite-filter guides.

4 Experiments

We run synthetic experiments for the nonstationary Gaussian-sketch model in Theorem 3.1. The diagonal latent recursion and primitive covariances are specified in Appendix F. For every sequence length P, the implementation recomputes the averaged latent covariance H_P, the sketched population matrices G_P and C_P, and the population minimizer $B_{*,P} = C_P G_P^{-1}$. Empirical variance is evaluated by directly sampling complete latent trajectories, forming $\widehat{G}$ and $\widehat{C}$, and applying safeguarded one-sided WSD-GD filters with $R = \gamma L$, as defined in Section 2. The warmup and decay phases each occupy 2% of the steps; Appendix F specifies the schedule details. We use Gaussian sketches $S_{\ell j} \sim \mathcal{N}(0, 1/M)$.

The first configuration represents Regime I:

$$\alpha = 1.5, \quad \theta = 2.5 \geq \alpha, \quad r = 0.2, \quad \rho = 0.55, \quad c_\xi = 0.05, \quad c_0 = 0.01.$$

The second represents Regime II:

$$\alpha = 1.8, \quad \theta = 1.25, \quad r = 0.35, \quad \rho = 0.55, \quad c_\xi = 0.03, \quad c_0 = 0.30,$$

where $\alpha - 2r = 1.1 \leq \theta < \alpha$. Both configurations use $\gamma = 0.1$. Additional formulas, grids, and checkpoint diagnostics are given in Appendix F.

Regime I. Panel (a) varies $M \in \{12, 16, 24, 32, 48, 64, 96, 128\}$ at $P = 4096$ and averages over eight Gaussian sketches. The fitted slope is -0.962, compared with the predicted exponent $1 - \beta_\alpha = -(\alpha + 2r - 1) = -0.90$. Panel (b) fixes $(M, d, P) = (512, 2048, 4096)$, averages over three sketches, and varies $L \in \{4000, 6000, 8000, 12000, 16000, 24000\}$. The fitted slope against R is -0.666, with reference exponent $(1 - \beta_\alpha)/\alpha = -0.60$.

Panel (c) uses $(M, d) = (20, 128)$ and directly sampled trajectories. Its fitted variance slopes are -0.972 in N, -1.026 in P, and 0.482 in R, compared with the plotted guide slopes -1, -1, and 0.530, respectively. The horizon sweep reaches effective-dimension saturation, so its guide slope is below the unsaturated exponent $1/\alpha = 0.667$. Thus increasing trajectory count or trajectory length has the same leading inverse-token effect in this regime, while longer optimization activates additional noisy directions until saturation.

Regime II. Panel (d) fixes $(M, d) = (512, 2048)$ and varies $P \in \{4, 6, 8, 12, 16, 24, 32, 48\}$. The measured approximation has slope -0.949; the two-scale guide $M^{1-\beta_\alpha} + P^{-1}M^{1-\beta_\theta}$ has slope -0.950 on the same grid. Panel (e) fixes $(M, d, R) = (1024, 2048, 1200)$ and varies $P \in \{4, 6, 8, 12, 16, 24\}$. The empirical bias slope is -0.224, following the $\mathcal{B}_{R,P}$ guide slope -0.185.

Panel (f) uses $(M, d) = (40, 128)$ and compares trajectory-sampled variance with finite-filter guides. The measured slopes are -0.979 in N, -1.359 in P, and 0.524 in R, while the corresponding guide slopes are -1.000, -1.353, and 0.484. In particular, the stronger-than-P^{-1} decay confirms that trajectory length both enlarges the token budget and suppresses the initialization transient; hence N and P are no longer interchangeable in Regime II.

5 Related work

Solvable scaling-law models. Our closest methodological inspiration is Lin et al. (2024; 2025), who derive sketched linear-regression scaling laws in model size, data, and optimization time; related multi-epoch analysis shows that the useful amount of data reuse itself grows with dataset size (Yan et al., 2026). For sequence models, Lyu et al. (2025) characterize training-, model-, data-, and context-length scaling in a solvable self-attention model. We instead study a recurrent matrix predictor whose training trajectory length P changes the population spectrum through nonstationary initialization, while explicitly separating N from P.

Learning from multiple trajectories. Tu et al. (2024) study supervised learning from multiple dependent trajectories, including linear-system identification. Their many-trajectory rate depends on the total observation count, matching our Regime-I dependence on NP, while their few- versus many-trajectory transition shows more generally that trajectory count and length need not always be interchangeable. Y"uksel et al. (2025) obtain near-i.i.d. parametric sample complexity for long-context linear autoregressive systems. Our setting adds a sketched representation, finite-time GD, two primitive spectra, and separate dependence on trajectory count and trajectory length.

Sequence length in language-model pretraining. Empirically, fewer document truncations improves language modeling (Ding et al., 2024), while variable-length dataset decomposition shows the sequence-length distribution can materially change performance even under controlled token budgets (Pouransari et al., 2024). These results motivate treating sequence allocation as consequential, but do not establish theoretical mechanism. Related theories study next-token prediction and ICL (Li et al., 2024; Gong et al., 2025; Kunstner & Bach, 2025; Arora et al., 2025). In comparison, our focus is the joint M, N, P, R scaling of dynamic learning.

6 Conclusion

We developed a tractable scaling-law model for autoregressive pretraining with a stable latent linear RNN, a sketched representation, and an empirically trained recurrent student. The expected prediction risk separates into irreducible loss, approximation, optimization bias, statistical variance, and a controlled cross term, with explicit dependence on model dimension M, token budget NP, and optimization horizon R. Distinct initialization and innovation spectra produce two regimes: trajectory count and length are interchangeable at leading order in Regime I, whereas longer trajectories additionally suppress the initialization transient in Regime II.

Important extensions include learned nonlinear representations, state-space and attention-based architectures, stochastic optimization, and more realistic sequence dependence. A broader theory should also connect pretraining dynamics to in-context adaptation and relate the spectral exponents in this stylized model to measurable properties of real language data and embedding spaces.

AI USE STATEMENT.

The authors used ChatGPT 5.6-Sol to assist with literature review, manuscript editing, mathematical exposition, and implementation and debugging of the synthetic experiments. All AI-assisted mathematical arguments, code, experimental outputs, and final manuscript content were independently checked and edited by the authors.

ETHICS AND REPRODUCIBILITY STATEMENT

Ethics. This work studies an idealized linear autoregressive model through theoretical analysis and synthetic Gaussian experiments. The experiments involve no human participants, personal data, or real-world text corpora. The results describe scaling within the stated spectral and stability assumptions; they should not be interpreted as universal prescriptions for language-model training or as evidence of improved fairness or safety in deployed systems.

Reproducibility. The model assumptions and optimization procedure are specified in Section 2, and the appendices provide proofs of the theoretical results. Section 4 and Appendix F document the synthetic data-generating process, parameter settings, experimental grids, sampling repetitions, error bars, and evaluation formulas. The main theorem uses safeguarded WSD-GD with $R = \gamma L$, and the reported experiments use the same WSD schedule. The phase lengths and experimental settings are documented in Appendix F.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D. Goodman. Bayesian scaling laws for in-context learning. In *Second Conference on Language Modeling*, 2025.

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. doi: 10.1073/pnas.2311878121.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, 2023.

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.

Francesco Cagnetta, Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Scaling laws and representation learning in simple hierarchical languages: Transformers versus convolutional architectures. *Physical Review E*, 112:065312, 2025. doi: 10.1103/qtd6-nl8p.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022.

Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. Fewer truncations improve language modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11030–11048, 2024.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, 2025. doi: 10.18653/v1/2025.acl-long.366.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

Zixuan Gong, Xiaolin Hu, Huayi Tang, and Yong Liu. Towards auto-regressive next-token prediction: In-context learning emerges from generalization. In *International Conference on Learning Representations*, 2025.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

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, Nick Ryder, Daniel M. Ziegler, John Schulman, 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. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30032, 2022.

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.

Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. In *Advances in Neural Information Processing Systems*, volume 38, pp. 160282–160314, 2025. doi: 10.52202/085713-4835.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 685–693, 2024.

Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 34943–34975, 2025.

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, pp. 60556–60606, 2024. doi: 10.52202/079017-1937.

Licong Lin, Jingfeng Wu, and Peter L. Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems*, volume 38, pp. 100719–100765, 2025. doi: 10.52202/085713-3034.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005. doi: 10.1007/978-3-540-27752-1.

Bochen Lyu, Di Wang, and Zhanxing Zhu. A solvable attention for neural scaling laws. In *International Conference on Learning Representations*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Hadi Pouransari, Chun-Liang Li, Jen-Hao Rick Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. Dataset decomposition: Faster llm training with variable sequence length curriculum. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1139.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lars Martens, Xiang Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolay Grigorev, Doug Fritz, Thibault Sottiaux, Arnas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Timo Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeffrey Stanway, Liza Bennett, Demis Hassabis, Koray Kavukcuoglu, Geoffrey Irving, and Shane Legg. Scaling language models: Methods, analysis and insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Paul M. Riechers, Henry R. Bigelow, Eric A. Alt, and Adam Shai. Next-token pretraining implies in-context learning. *arXiv preprint arXiv:2505.18373*, 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.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109, 2024.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, 2025.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. In *International Conference on Learning Representations*, 2026.

Oğuz Kaan Y"uksel, Mathieu Even, and Nicolas Flammarion. Long-context linear system identification. In *International Conference on Learning Representations*, 2025.

Linear RNN Scaling Law

Supplementary Material

A Preliminaries: sketched autoregressive prediction

A.1 Sketched regression identities

Sketched autoregressive regression. For each trajectory $i = 1, \dots, N$, the learner observes

$$u_{i,p} := Sx_{i,p} \in \mathbb{R}^M, \qquad y_{i,p} := u_{i,p+1}, \qquad p = 0, \dots, P-1.$$

The population covariance and cross-covariance are, for any fixed trajectory index because the trajectories are identically distributed,

$$G := \frac{1}{P} \sum_{p=0}^{P-1} \mathbb{E}[u_{i,p} u_{i,p}^\top], \qquad C := \frac{1}{P} \sum_{p=0}^{P-1} \mathbb{E}[y_{i,p} u_{i,p}^\top],$$

and their empirical versions are

$$\widehat{G} := \frac{1}{NP} \sum_{i=1}^{N} \sum_{p=0}^{P-1} u_{i,p} u_{i,p}^\top, \qquad \widehat{C} := \frac{1}{NP} \sum_{i=1}^{N} \sum_{p=0}^{P-1} y_{i,p} u_{i,p}^\top.$$

The original population objective is defined before imposing the sketch. For any ambient transition $A \in \mathbb{R}^{d \times d}$, define

$$\mathcal{L}(A) := \frac{1}{2P} \sum_{p=0}^{P-1} \mathbb{E} \|x_{i,p+1} - A x_{i,p}\|_2^2.$$

Using $x_{i,p+1} = A_* x_{i,p} + \xi_{i,p+1}$ and independence of the mean-zero innovation from $x_{i,p}$,

$$\mathcal{L}(A) = \mathcal{L}(A_*) + \frac{1}{2}\left\|(A - A_*)H^{1/2}\right\|_F^2, \qquad \mathcal{L}(A_*) = \frac{1}{2}\operatorname{Tr}(\Sigma_\xi).$$

Thus A_* is the original population minimizer and $\mathcal{L}(A_*)$ is the full-state irreducible innovation loss.

The trained sketched model uses only the observed variables $u_{i,p} = Sx_{i,p}$ and $y_{i,p} = Sx_{i,p+1}$. Its observable population objective is

$$\mathcal{L}_M(B) := \frac{1}{2P}\sum_{p=0}^{P-1}\mathbb{E}\|y_{i,p} - Bu_{i,p}\|_2^2, \qquad B \in \mathbb{R}^{M\times M}.$$

Equivalently,

$$\mathcal{L}_M(B) = \frac{1}{2}\operatorname{Tr}(S\Sigma_\xi S^\top) + \frac{1}{2}\left\|(BS - SA_*)H^{1/2}\right\|_F^2.$$

This is not the full-state risk of any lifted ambient matrix; using such a lifted risk would change the approximation, bias, and variance terms analyzed below. The empirical sketched objective is

$$\widehat{\mathcal{L}}_M(B) := \frac{1}{2NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1}\|y_{i,p} - Bu_{i,p}\|_2^2.$$

The gradients are

$$\nabla\widehat{\mathcal{L}}_M(B) = B\widehat{G} - \widehat{C}, \qquad \nabla\mathcal{L}_M(B) = BG - C.$$

Thus

$$B_* := CG^{-1}$$

is the sketched population minimizer. Using the G-weighted norm from the notation paragraph, we have

$$\mathcal{L}_M(B) - \mathcal{L}_M(B_*) = \frac{1}{2}\|B - B_*\|_G^2.$$

The population approximation term is

$$\mathrm{Approx}_M := \mathcal{L}_M(B_*) - \frac{1}{2}\operatorname{Tr}(S\Sigma_\xi S^\top) = \inf_{B\in\mathbb{R}^{M\times M}}\frac{1}{2}\left\|(BS - SA_*)H^{1/2}\right\|_F^2.$$

Empirical WSD gradient descent. Full-batch WSD-GD on $\widehat{\mathcal{L}}_M$ is

$$B_t = B_{t-1} - \bar{\gamma}_t(B_{t-1}\widehat{G} - \widehat{C}), \qquad t = 1,\dots,L, \qquad B_0 = 0.$$

Let

$$\Delta_t := B_t - B_*, \qquad \widehat{E} := \widehat{C} - B_*\widehat{G}.$$

Then

$$\Delta_t = \Delta_{t-1}(I - \bar{\gamma}_t\widehat{G}) + \bar{\gamma}_t\widehat{E}.$$

Iterating the recursion gives

$$B_L - B_* = -B_*\prod_{t=1}^{L}(I - \bar{\gamma}_t\widehat{G}) + \sum_{t=1}^{L}\bar{\gamma}_t\widehat{E}\prod_{s=t+1}^{L}(I - \bar{\gamma}_s\widehat{G}).$$

Accordingly define

$$\widehat{\mathcal{B}}_L(B_*) := B_*\prod_{t=1}^{L}(I - \bar{\gamma}_t\widehat{G})$$

and

$$\widehat{\mathcal{V}}_L(\widehat{E}) := \sum_{t=1}^{L}\bar{\gamma}_t\widehat{E}\prod_{s=t+1}^{L}(I - \bar{\gamma}_s\widehat{G}).$$

Then

$$B_L - B_* = -\widehat{\mathcal{B}}_L(B_*) + \widehat{\mathcal{V}}_L(\widehat{E}).$$

Consequently, taking expectation over the sampled training trajectories (and hence over $\widehat{G}, \widehat{C}$ and B_L, conditional on the sketch),

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}_M(B_L) - \mathcal{L}_M(B_*) \mid S] = \mathrm{Bias}_L + \mathrm{Var}_L + \mathrm{Cross}_L,$$

where

$$\mathrm{Bias}_L := \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{B}}_L(B_*)\|_G^2 \mid S\right],$$
$$\mathrm{Var}_L := \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{V}}_L(\widehat{E})\|_G^2 \mid S\right],$$

and

$$\mathrm{Cross}_L := -\mathbb{E}_{\mathcal{D}}\left[\left\langle \widehat{\mathcal{B}}_L(B_*), \widehat{\mathcal{V}}_L(\widehat{E})\right\rangle_G \mid S\right].$$

By Cauchy–Schwarz and $2ab \le a^2 + b^2$,

$$|\,\mathrm{Cross}_L\,| \le 2\sqrt{\mathrm{Bias}_L\,\mathrm{Var}_L} \le \mathrm{Bias}_L + \mathrm{Var}_L\,.$$

Covariance event used in proofs. The GD-bias and GD-variance estimates use the following auxiliary event:

$$\mathcal{E}_{\mathrm{cov}}(R) := \left\{\left\|G^{-1/2}\widehat{G}G^{-1/2} - I\right\| \le c_{\mathrm{rel}}R^{-1/2}\right\},$$

where $c_{\mathrm{rel}} > 0$ is a sufficiently small absolute constant. This event is used only inside the proofs; Lemma E.7 turns the effective-sample condition into this event with high probability. On the Gaussian-sketch event, let $\|G\|_{\mathrm{op}} \le C_G$. If $\gamma \le (4C_G)^{-1}$, then on $\mathcal{E}_{\mathrm{cov}}(R)$,

$$\widehat{G} \preceq 2G, \qquad \gamma \le \frac{1}{2\|\widehat{G}\|_{\mathrm{op}}}, \qquad \bar{\gamma} = \gamma.$$

Thus all good-event filter calculations may use the nominal schedule γ_t, while the safeguarded schedule is stable for every sample.

A.2 Proof of Proposition 3.1

Proof. Start from the sketched-model loss and substitute $y_{i,p} = Sx_{i,p+1} = SA_*x_{i,p} + S\xi_{i,p+1}$. For any $B \in \mathbb{R}^{M\times M}$,

$$y_{i,p} - Bu_{i,p} = (SA_* - BS)x_{i,p} + S\xi_{i,p+1}.$$

The innovation is mean zero and independent of $x_{i,p}$, so the cross term vanishes after expectation. Averaging over positions gives

$$\mathcal{L}_M(B) = \frac{1}{2}\,\mathrm{Tr}(S\Sigma_\xi S^\top) + \frac{1}{2}\left\|(BS - SA_*)H^{1/2}\right\|_F^2.$$

Therefore the best sketched population predictor B_* satisfies

$$\mathrm{Approx}_M := \mathcal{L}_M(B_*) - \frac{1}{2}\,\mathrm{Tr}(S\Sigma_\xi S^\top) = \inf_{B\in\mathbb{R}^{M\times M}} \frac{1}{2}\left\|(BS - SA_*)H^{1/2}\right\|_F^2.$$

The quadratic expansion of $\mathcal{L}_M$ around its minimizer gives

$$\mathcal{L}_M(B_L) = \frac{1}{2}\,\mathrm{Tr}(S\Sigma_\xi S^\top) + \mathrm{Approx}_M + \frac{1}{2}\|B_L - B_*\|_G^2.$$

It remains to decompose the final excess term. Subtracting B_* from the GD recursion gives

$$B_t - B_* = (B_{t-1} - B_*)(I - \bar{\gamma}_t\widehat{G}) + \bar{\gamma}_t\widehat{E}.$$

Iterating this recursion from $B_0 = 0$ yields

$$B_L - B_* = -\widehat{\mathcal{B}}_L(B_*) + \widehat{\mathcal{V}}_L(\widehat{E}).$$

Therefore

$$\frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|B_L - B_*\|_G^2 \mid S\right] = \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{B}}_L(B_*)\|_G^2 \mid S\right] + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathcal{V}}_L(\widehat{E})\|_G^2 \mid S\right] - \mathbb{E}_{\mathcal{D}}\left[\left\langle \widehat{\mathcal{B}}_L(B_*), \widehat{\mathcal{V}}_L(\widehat{E})\right\rangle_G \mid S\right].$$

Combining the preceding displays proves the underbraced decomposition in Proposition 3.1. The cross-term bound follows from Cauchy–Schwarz and $2ab \le a^2 + b^2$. □

A.3 Proof of Theorem 3.1

Proof. Fix a sketch in the intersection of the high-probability events from Lemmas E.11 and E.8, together with the appropriate specific approximation theorem, Theorem B.1 or Theorem B.2. This intersection has probability at least $1 - \exp(-\Omega(M))$ over S. On it,

$$\mathrm{Approx}_M \asymp \mathcal{A}_{M,P}, \qquad \|G\|_{\mathrm{op}} \le C_G, \qquad \|B_*\|_F^2 \lesssim M.$$

All remaining expectations and probabilities in the proof are conditional on this fixed sketch.

Let $\mathcal{E} := \mathcal{E}_{\mathrm{cov}}(R)$. By Lemma E.7 and the regime-dependent sample conditions in the theorem,

$$\mathbb{P}_{\mathcal{D}}(\mathcal{E}^c \mid S) \le \exp(-cM).$$

Lemma E.8 shows that $\bar{\gamma} = \gamma$ on $\mathcal{E}$. Thus the appropriate eventwise bias estimate from Theorem C.1 or Theorem C.2, together with the strictly unsaturated condition, gives

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{bias}_L(\mathcal{D})\mathbf{1}_{\mathcal{E}} \mid S\right] = \Theta\left(\mathcal{B}_{R,P}\right)$$

where the exponentially small loss of probability is absorbed into the constants. The appropriate variance theorem and its covariance-event lower bound similarly give

$$\mathrm{Var}_L^{\mathrm{good}} = \Theta(\mathcal{V}_{R,M,P}).$$

On $\mathcal{E}^c$, the safeguarded products are contractive. Therefore, using $\|G\|_{\mathrm{op}} \le C_G$, $\|B_*\|_F^2 \lesssim M$, and the variance complement estimate from Proposition D.1,

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{bias}_L(\mathcal{D})\mathbf{1}_{\mathcal{E}^c} \mid S\right] \lesssim Me^{-cM},$$
$$\mathrm{Var}_L - \mathrm{Var}_L^{\mathrm{good}} \lesssim R^2 e^{-cM/2}.$$

Consequently,

$$\mathrm{Bias}_L = \Theta(\mathcal{B}_{R,P}) + \mathcal{O}(\varepsilon_M), \qquad \mathrm{Var}_L = \Theta(\mathcal{V}_{R,M,P}) + \mathcal{O}(\varepsilon_M).$$

Applying Cauchy–Schwarz separately on $\mathcal{E}$ and $\mathcal{E}^c$ yields

$$\mathrm{Cross}_L = \mathcal{O}\left(\sqrt{\mathcal{B}_{R,P}\mathcal{V}_{R,M,P}} + \varepsilon_M\right).$$

Substituting these component estimates and $\mathrm{Approx}_M = \Theta(\mathcal{A}_{M,P})$ into the exact decomposition of Proposition 3.1 proves the stated rate equality. □

A.4 Proof of Corollary 3.1

Proof. Write

$$\mathcal{K} := \gamma\mathcal{C} = \mathcal{D}M^2R.$$

We optimize the polynomial terms in Theorem 3.1; the exponentially small remainder is negligible under the polynomial allocations below. Fixed constants, including C_ρ, are absorbed into $\asymp$.

Regime I. The leading approximation and bias terms balance when

$$M^{1-\beta_\alpha} \asymp R^{(1-\beta_\alpha)/\alpha},$$

or equivalently

$$R \asymp M^\alpha.$$

In the token-rich phase, the sample requirement binds at

$$\mathcal{D} \asymp RM \asymp M^{\alpha+1}.$$

Consequently,

$$\mathcal{K} = \mathcal{D}M^2R \asymp M^{2\alpha+3}.$$

Solving gives

$$M^\star \asymp \mathcal{K}^{1/(2\alpha+3)}, \qquad R^\star \asymp \mathcal{K}^{\alpha/(2\alpha+3)},$$
$$\mathcal{D}^\star \asymp \mathcal{K}^{(\alpha+1)/(2\alpha+3)}, \quad L^\star \asymp R^\star/\gamma.$$

Thus the token-rich allocation is feasible precisely when $\mathcal{T} \gtrsim \mathcal{T}_{\mathrm{crit},I}$.

In the token-limited phase, set $\mathcal{D} = \mathcal{T}$. The compute constraint gives $R = \mathcal{K}/(\mathcal{T}M^2)$. The unconstrained balance $R \asymp M^\alpha$ would give

$$M_0 \asymp (\mathcal{K}/\mathcal{T})^{1/(\alpha+2)}.$$

The sample condition $\mathcal{T} \gtrsim RM$ holds at this point exactly when

$$\mathcal{T} \gtrsim \mathcal{K}^{(\alpha+1)/(2\alpha+3)}.$$

Below this scale the sample condition binds, so

$$RM \asymp \mathcal{T}, \qquad \mathcal{K} = \mathcal{T}M^2R.$$

Therefore

$$M^\star \asymp \frac{\mathcal{K}}{\mathcal{T}^2}, \qquad R^\star \asymp \frac{\mathcal{T}^3}{\mathcal{K}}, \qquad L^\star \asymp \frac{\mathcal{T}^3}{\gamma\mathcal{K}} = \frac{\mathcal{T}^3}{\gamma^2\mathcal{C}}.$$

The Regime-I theorem depends on N and P only through $\mathcal{D} = NP$, so any leading-order split with $N^\star = \mathcal{D}^\star/P^\star$ is equivalent.

It remains to justify neglecting variance. In the token-rich phase,

$$\frac{\mathcal{V}}{\mathcal{A}} \asymp M^{2r-1} = o(1).$$

In the token-limited phase, $R \lesssim M^\alpha$, so bias is at least as large as approximation and

$$\frac{\mathcal{V}}{\mathcal{B}} \lesssim \frac{R^{2r/\alpha}}{M} \le M^{2r-1} = o(1).$$

Regime II. In the token-rich phase, place the representation and optimization crossovers at the learned model scale:

$$R \asymp M^\alpha, \qquad P \asymp M^{\alpha-\theta}.$$

The additional trajectory condition binds at

$$N \asymp RM \asymp M^{\alpha+1},$$

and hence

$$\mathcal{D} = NP \asymp M^{2\alpha-\theta+1}.$$

The compute identity becomes

$$\mathcal{K} \asymp M^{3\alpha-\theta+3}.$$

Solving gives

$$M^\star \asymp \mathcal{K}^{1/(3\alpha-\theta+3)}, \qquad R^\star \asymp \mathcal{K}^{\alpha/(3\alpha-\theta+3)},$$
$$P^\star \asymp \mathcal{K}^{(\alpha-\theta)/(3\alpha-\theta+3)}, \quad N^\star \asymp \mathcal{K}^{(\alpha+1)/(3\alpha-\theta+3)},$$

and

$$\mathcal{D}^\star \asymp \mathcal{K}^{(2\alpha-\theta+1)/(3\alpha-\theta+3)}.$$

This gives $\mathcal{T}_{\mathrm{crit},II}$ and the token-rich formulas.

In the token-limited phase, set $\mathcal{D} = \mathcal{T}$. The solution lies in the initialization-dominated branch. Balancing approximation and bias gives

$$P^{-1}M^{1-\beta_\theta} \asymp P^{-1}(R/P)^{(1-\beta_\theta)/\theta},$$

and therefore

$$R \asymp PM^\theta.$$

Increasing P improves the leading terms until the trajectory condition binds:

$$N = \frac{\mathcal{T}}{P} \asymp RM, \qquad \mathcal{T} \asymp PRM.$$

Together with $R \asymp PM^\theta$ and $\mathcal{K} = \mathcal{T}M^2R$, this yields

$$\mathcal{T} \asymp P^2M^{\theta+1}, \qquad \mathcal{K} \asymp \mathcal{T}PM^{\theta+2}.$$

Solving gives

$$M^\star \asymp \left(\frac{\mathcal{K}^2}{\mathcal{T}^3}\right)^{1/(\theta+3)}, \qquad R^\star \asymp \left(\mathcal{K}^{\theta-1}\mathcal{T}^{3-\theta}\right)^{1/(\theta+3)},$$
$$P^\star \asymp \left(\frac{\mathcal{T}^{2\theta+3}}{\mathcal{K}^{\theta+1}}\right)^{1/(\theta+3)}, \quad N^\star \asymp \left(\frac{\mathcal{K}^{\theta+1}}{\mathcal{T}^{\theta}}\right)^{1/(\theta+3)}.$$

Moreover,

$$\frac{P^\star}{(M^\star)^{\alpha-\theta}} \asymp \left(\frac{\mathcal{T}^{3\alpha-\theta+3}}{\mathcal{K}^{2\alpha-\theta+1}}\right)^{1/(\theta+3)}.$$

Thus $P^\star \lesssim (M^\star)^{\alpha-\theta}$ exactly when

$$\mathcal{T} \lesssim \mathcal{K}^{(2\alpha-\theta+1)/(3\alpha-\theta+3)} = \mathcal{T}_{\mathrm{crit},II}.$$

Direct substitution at equality shows that the token-limited and token-rich allocations match continuously up to constants.

Finally, in the token-limited phase $(R/P)^{1/\theta} \asymp M$, and therefore

$$\frac{\mathcal{V}}{\mathcal{A}} \asymp \frac{M^{2r-1}}{P} = o(1).$$

In the token-rich phase,

$$\frac{\mathcal{V}}{\mathcal{A}} \asymp M^{\theta-\alpha+2r-1} = o(1),$$

because $\theta < \alpha$ and $2r < 1$. Thus variance is lower order in both regimes, and Cauchy–Schwarz makes the bias–variance cross term lower order as well. Choosing sufficiently small fixed constants in the displayed balances keeps all allocations in the unsaturated regime of Theorem 3.1. □

B Approximation error

Define

$$K_* := A_* H^{1/2}, \qquad \widetilde{K}_* := SK_*, \qquad h_{P,j} := e_j^\top H e_j, \qquad \tau_{P,j} := \|K_* e_j\|_2^2 = h_{P,j} a_j^2.$$

For an index set I, let Π_I be the coordinate projection onto $\mathrm{span}\{e_j : j \in I\}$, and write $\Pi_{\le m} := \Pi_{\{1,\dots,m\}}$ and $\Pi_{>m} := I - \Pi_{\le m}$.

B.1 General approximation upper bound

Lemma B.1 (General approximation upper bound)**.** *Suppose Assumptions 1 and 2 hold. There is a sufficiently small absolute constant $c_0 > 0$ such that, for every integer $m \le c_0 M$, with probability at least $1 - \exp(-cM)$ over the Gaussian sketch,*

$$\mathrm{Approx}_M \lesssim \sum_{j=m+1}^{d} \tau_{P,j} + h_{P,m+1} \sum_{j \le m} \frac{\tau_{P,j}}{h_{P,j}}.$$

Proof. Let

$$U := G^{-1/2} S H^{1/2}, \qquad P_S := U^\top U, \qquad Q_S := I - P_S.$$

Because $UU^\top = I_M$, P_S is the orthogonal projection onto the row space of $SH^{1/2}$. Projecting the rows of $SA_*H^{1/2}$ onto this row space gives

$$\mathrm{Approx}_M = \frac{1}{2}\|SK_* Q_S\|_F^2.$$

Splitting at m and using that Q_S is a contraction,

$$\|SK_* Q_S\|_F^2 \le 2\|SK_* \Pi_{\le m} Q_S\|_F^2 + 2\|SK_* \Pi_{>m}\|_F^2.$$

Gaussian quadratic-form concentration gives

$$\|SK_*\Pi_{>m}\|_F^2 \lesssim \sum_{j=m+1}^{d} \tau_{P,j}$$

with probability at least $1 - \exp(-cM)$.

For the head term, let $H_0 := \Pi_{\le m} H \Pi_{\le m}$. Then

$$\|SK_*\Pi_{\le m} Q_S\|_F^2 \le \|SK_*\Pi_{\le m} H_0^{-1/2}\|_F^2 \|H_0^{1/2}\Pi_{\le m} Q_S\|^2,$$
$$\|SK_*\Pi_{\le m} H_0^{-1/2}\|_F^2 \lesssim \sum_{j\le m} \frac{\tau_{P,j}}{h_{P,j}}.$$

It remains to control the projection leakage. Write $S = (S_0, S_1)$, $H = H_0 \oplus H_1$, and $\Sigma_1 := S_1 H_1 S_1^\top$. When $m \le c_0 M$, standard Gaussian singular-value concentration gives $S_0^\top S_0 \succeq cI_m$. Moreover, the mixed sequence satisfies

$$\sum_{j>m} h_{P,j} \lesssim m h_{P,m}, \qquad h_{P,m} \asymp h_{P,m+1},$$

by Lemma E.9 and integral comparison. A weighted chi-square net argument therefore yields $\|\Sigma_1\| \lesssim h_{P,m+1}$. The Woodbury identity now gives

$$\|H_0^{1/2}\Pi_{\le m} Q_S\|^2 = \left\| \left(H_0^{-1} + S_0^\top \Sigma_1^{-1} S_0\right)^{-1} \right\|$$
$$\lesssim h_{P,m+1}.$$

Combining the head and tail estimates proves the result. □

B.2 General approximation lower bound

Lemma B.2 (General approximation lower bound)**.** *Suppose Assumptions 1 and 2 hold. With probability at least* $1 - \exp(-cM)$ *over the Gaussian sketch, for every* m *satisfying* $c_1 M \le m \le c_2 M$*,*

$$\mathrm{Approx}_M \gtrsim \sum_{j=m+1}^{2m} \tau_{P,j}.$$

Proof. Write $s_j := Se_j$. Then

$$2\ \mathrm{Approx}_M = \inf_{B \in \mathbb{R}^{M\times M}} \sum_{j=1}^{d} h_{P,j} \|(B - a_j I) s_j\|_2^2.$$

Thus every sketched direction s_j asks the same matrix B to act approximately as multiplication by the corresponding teacher eigenvalue a_j.

Fix an admissible $m \asymp M$, let $n = \lceil 2M \rceil$, and define

$$I_1 := \{m+1, \dots, m+n\}, \qquad I_2 := \{\lceil Km \rceil + 1, \dots, \lceil Km \rceil + n\}.$$

Here $K > 1$ is a sufficiently large fixed constant. We choose K so the blocks are disjoint for every $c_1 M \le m \le c_2 M$, and choose C_{amb} larger than the endpoint multiplier of I_2, so both blocks lie in $\{1, \dots, d\}$. Since the indices in each block are fixed constant multiples of m, Assumption 1 gives

$$a_j \asymp m^{-r}, \qquad j \in I_1,$$

and

$$a_j \asymp K^{-r} m^{-r}, \qquad j \in I_2.$$

Thus the two blocks require incompatible transition scales when K is sufficiently large.

Consider the Gaussian frame event

$$cI_M \preceq \sum_{j\in I_\ell} s_j s_j^\top \preceq CI_M, \qquad \ell = 1, 2,$$

which holds with probability at least $1 - \exp(-cM)$ by Gaussian singular-value concentration. On this event, for every deterministic matrix $D \in \mathbb{R}^{M\times M}$,

$$\sum_{j\in I_\ell} \|Ds_j\|_2^2 = \mathrm{Tr}\left(D^\top D \sum_{j\in I_\ell} s_j s_j^\top\right),$$
$$c\|D\|_F^2 \le \sum_{j\in I_\ell} \|Ds_j\|_2^2 \le C\|D\|_F^2, \qquad \ell = 1, 2.$$

Because these inequalities follow from a matrix event, they hold simultaneously for every D, including a matrix chosen after observing the sketch. Thus each block probes the whole M-dimensional observable space.

Define

$$E_1(B) := \sum_{j\in I_1} \|(B - a_j I)s_j\|_2^2, \qquad E_2(B) := \sum_{j\in I_2} \|(B - a_j I)s_j\|_2^2.$$

We now prove, uniformly over B, that

$$E_1(B) + E_2(B) \gtrsim Mm^{-2r}.$$

If $E_1(B) \gtrsim Mm^{-2r}$, the desired bound is immediate. For the complementary case, suppose

$$E_1(B) \le \eta Mm^{-2r}$$

for a sufficiently small fixed $\eta > 0$. Using

$$a_j s_j = Bs_j - (B - a_j I)s_j$$

and $\|x - y\|_2^2 \le 2\|x\|_2^2 + 2\|y\|_2^2$, we obtain

$$\sum_{j\in I_1} \|Bs_j\|_2^2 \ge \frac{1}{2}\sum_{j\in I_1} a_j^2\|s_j\|_2^2 - E_1(B).$$

The frame event also gives

$$\sum_{j\in I_1} \|s_j\|_2^2 = \mathrm{Tr}\left(\sum_{j\in I_1} s_j s_j^\top\right) \asymp M.$$

Since $a_j^2 \asymp m^{-2r}$ on I_1, choosing η sufficiently small yields

$$\sum_{j\in I_1} \|Bs_j\|_2^2 \gtrsim Mm^{-2r}.$$

The upper frame inequality on I_1 therefore implies

$$\|B\|_F^2 \gtrsim Mm^{-2r}.$$

Applying the lower frame inequality on I_2,

$$\sum_{j\in I_2} \|Bs_j\|_2^2 \gtrsim \|B\|_F^2 \gtrsim Mm^{-2r}.$$

In contrast, the teacher energy on the second block satisfies

$$\sum_{j\in I_2} a_j^2\|s_j\|_2^2 \lesssim K^{-2r}Mm^{-2r}.$$

Using $\|x - y\|_2^2 \ge \frac{1}{2}\|x\|_2^2 - \|y\|_2^2$ with $x = Bs_j$ and $y = a_j s_j$, we get

$$E_2(B) \ge \frac{1}{2}\sum_{j\in I_2} \|Bs_j\|_2^2 - \sum_{j\in I_2} a_j^2\|s_j\|_2^2.$$

Choose K sufficiently large so that the second term is at most a fixed small fraction of the first. It follows that

$$E_2(B) \gtrsim Mm^{-2r}.$$

We have shown the dichotomy explicitly: either I_1 already incurs error of order Mm^{-2r}, or fitting I_1 well forces $\|B\|_F^2 \gtrsim Mm^{-2r}$, in which case I_2 incurs error of the same order. Hence

$$\inf_{B\in\mathbb{R}^{M\times M}} \{E_1(B) + E_2(B)\} \gtrsim Mm^{-2r}.$$

Finally, restore the covariance weights. By Lemma E.9,

$$h_{P,j} \asymp j^{-\alpha} + P^{-1}j^{-\theta}.$$

All indices in $I_1 \cup I_2$ are fixed constant multiples of m, so fixed-scale regularity gives

$$h_{P,j} \gtrsim h_{P,m}, \qquad j \in I_1 \cup I_2,$$

up to constants depending only on the fixed choice of K. Therefore

$$\mathrm{Approx}_M \gtrsim Mh_{P,m}m^{-2r} \asymp mh_{P,m}m^{-2r},$$

because $m \asymp M$. On the annulus $m < j \le 2m$,

$$h_{P,j} \asymp h_{P,m}, \qquad a_j^2 \asymp m^{-2r}.$$

Consequently,

$$mh_{P,m}m^{-2r} \asymp \sum_{j=m+1}^{2m} h_{P,j}a_j^2 = \sum_{j=m+1}^{2m} \tau_{P,j}.$$

This proves the claim for a fixed m. There are only $O(M)$ admissible integer values of m, so a union bound makes the frame events simultaneous for every such m while preserving probability at least $1 - \exp(-cM)$. □

B.3 Specific approximation bound for $\theta \ge \alpha$

Theorem B.1 (Approximation bound for $\theta \ge \alpha$)**.** *If $\theta \ge \alpha$, then, with probability at least $1 - \exp(-cM)$ over the Gaussian sketch,*

$$\mathrm{Approx}_M \asymp M^{1-\beta_\alpha}.$$

The boundary case $\theta = \alpha$ is included.

Proof. Lemma E.9 gives $h_{P,j} \asymp j^{-\alpha}$ and $\tau_{P,j} \asymp j^{-\beta_\alpha}$. Choose $m \asymp M$ in the two general lemmas. The finite tail up to d has the same order as the infinite power-law tail because it contains the fixed annulus $(m, 2m]$. Since $\tau_{P,j}/h_{P,j} = a_j^2 \asymp j^{-2r}$,

$$\sum_{j=m+1}^{d} \tau_{P,j} \asymp m^{1-\beta_\alpha}, \qquad h_{P,m}\sum_{j\le m} \frac{\tau_{P,j}}{h_{P,j}} \asymp m^{-\alpha}m^{1-2r} = m^{1-\beta_\alpha}.$$

The annulus lower bound has the same order. □

B.4 Specific approximation bound for $\alpha - 2r \le \theta < \alpha$

Theorem B.2 (Approximation bound for heavier initialization)**.** *If $\alpha - 2r \le \theta < \alpha$, then, with probability at least $1 - \exp(-cM)$ over the Gaussian sketch,*

$$\mathrm{Approx}_M \asymp M^{1-\beta_\alpha} + P^{-1}M^{1-\beta_\theta}.$$

Equivalently, with $m_P = P^{1/(\alpha-\theta)}$,

$$\mathrm{Approx}_M \asymp \begin{cases} M^{1-\beta_\alpha}, & M \lesssim m_P, \\ P^{-1}M^{1-\beta_\theta}, & M \gtrsim m_P. \end{cases}$$

Proof. By Lemma E.9,

$$h_{P,j} \asymp j^{-\alpha} + P^{-1}j^{-\theta}, \qquad \tau_{P,j} \asymp j^{-\beta_\alpha} + P^{-1}j^{-\beta_\theta}.$$

For $m \asymp M$, integral comparison gives

$$\sum_{j=m+1}^{d} \tau_{P,j} \asymp m^{1-\beta_\alpha} + P^{-1}m^{1-\beta_\theta}.$$

Here the upper bound follows from the infinite power-law tail, while the lower bound follows from the annulus $(m, 2m] \subseteq \{1, \dots, d\}$. Also, because $0 < 2r < 1$,

$$\begin{aligned} h_{P,m} \sum_{j\le m} \frac{\tau_{P,j}}{h_{P,j}} &\asymp \left(m^{-\alpha} + P^{-1}m^{-\theta}\right) \sum_{j\le m} j^{-2r} \\ &\asymp m^{1-\beta_\alpha} + P^{-1}m^{1-\beta_\theta}. \end{aligned}$$

The annulus lower bound matches this order. Comparing the two summands gives the crossover m_P. □

C BIAS BOUND

Throughout this section we work on the Gaussian-sketch event of Lemma E.11 and, for eventwise filter comparisons, on

$$\mathcal{E}_{\mathrm{cov}}(R) := \left\{ \left\| G^{-1/2}\widehat{G}G^{-1/2} - I \right\| \le c_{\mathrm{rel}} R^{-1/2} \right\}.$$

The safeguard is inactive on this event. Write

$$G = \sum_{j=1}^{M} \mu_j v_j v_j^\top, \qquad q_j := \mu_j \|B_* v_j\|_2^2,$$

and define

$$\widehat{\Psi}_L := \prod_{t=1}^{L} (I - \gamma_t \widehat{G}), \qquad \psi_L(s) := \prod_{t=1}^{L} (1 - \gamma_t s).$$

The pointwise bias integrand is

$$\mathrm{bias}_L(\mathcal{D}) := \frac{1}{2}\|B_* \widehat{\Psi}_L\|_G^2.$$

The quantity Bias_L denotes the expectation of $\mathrm{bias}_L(\mathcal{D})$ over trajectories.

C.1 GENERAL GD-BIAS UPPER BOUND

Lemma C.1 (General GD-bias upper bound)**.** *On the events specified above,*

$$\mathrm{bias}_L(\mathcal{D}) \lesssim \sum_{j=1}^{M} q_j \min\{1, (R\mu_j)^{-1}\}.$$

Proof. Set $\lambda = R^{-1}$. The deterministic noncommutative comparison in Lemma E.13 gives

$$\widehat{\Psi}_L G \widehat{\Psi}_L \preceq C\lambda G(G + \lambda I)^{-1}.$$

Therefore

$$\begin{aligned} \|B_* \widehat{\Psi}_L\|_G^2 &\lesssim \mathrm{Tr}\left[B_* \lambda G(G + \lambda I)^{-1} B_*^\top\right] \\ &= \sum_{j=1}^{M} q_j \frac{\lambda}{\mu_j + \lambda} \\ &\asymp \sum_{j=1}^{M} q_j \min\{1, (R\mu_j)^{-1}\}. \end{aligned}$$

□

C.2 General GD-bias lower bound

Lemma C.2 (General GD-bias lower bound)**.** *On the covariance and Gaussian-sketch events, let $k \asymp \kappa_{R,P}$. If $k \le cM$, then*

$$\mathrm{bias}_L(\mathcal{D}) \gtrsim k^{1-\beta_\alpha} + P^{-1}k^{1-\beta_\theta}.$$

Proof. This is the final conclusion of Lemma E.14, with the factor $1/2$ in the definition of $\mathrm{bias}_L(\mathcal{D})$ absorbed into the constant. □

C.3 Specific GD-bias bound for $\theta \ge \alpha$

Theorem C.1 (GD-bias bound for $\theta \ge \alpha$)**.** *Suppose $\theta \ge \alpha$ and*

$$1 \lesssim R \lesssim M^\alpha.$$

On the covariance and Gaussian-sketch events,

$$\mathrm{bias}_L(\mathcal{D}) \lesssim R^{(1-\beta_\alpha)/\alpha}.$$

If $R^{1/\alpha} \le cM$, the matching empirical lower bound also holds:

$$\mathrm{bias}_L(\mathcal{D}) \gtrsim R^{(1-\beta_\alpha)/\alpha}.$$

The case $\theta = \alpha$ is included.

Proof. Let $k \asymp R^{1/\alpha}$, and set $\widetilde{k} := \min\{k, \lfloor M/3 \rfloor\}$. The general upper bound and Lemma E.12 give

$$\begin{aligned}\mathrm{bias}_L(\mathcal{D}) &\lesssim \frac{1}{R}\widetilde{k}^{1-2r} + \widetilde{k}^{1-\beta_\alpha} \\ &\lesssim R^{(1-\beta_\alpha)/\alpha}.\end{aligned}$$

When $k > M/3$, the unsaturated condition implies $k \asymp M$, so replacing k by $\widetilde{k}$ changes only constants. For the lower bound, Lemma C.2 gives

$$\mathrm{bias}_L(\mathcal{D}) \gtrsim k^{1-\beta_\alpha} \asymp R^{(1-\beta_\alpha)/\alpha}$$

whenever $k \le cM$. □

C.4 Specific GD-bias bound for $\alpha - 2r \le \theta < \alpha$

Theorem C.2 (GD-bias bound for heavier initialization)**.** *Suppose $\alpha - 2r \le \theta < \alpha$ and*

$$1 \lesssim R \lesssim \left(M^{-\alpha} + P^{-1}M^{-\theta}\right)^{-1}.$$

On the covariance and Gaussian-sketch events,

$$\mathrm{bias}_L(\mathcal{D}) \asymp \kappa_{R,P}^{1-\beta_\alpha} + P^{-1}\kappa_{R,P}^{1-\beta_\theta},$$

provided $\kappa_{R,P} \le cM$ for the lower bound. Equivalently, with $R_P = P^{\alpha/(\alpha-\theta)}$,

$$\mathrm{bias}_L(\mathcal{D}) \asymp \begin{cases} R^{(1-\beta_\alpha)/\alpha}, & R \lesssim R_P, \\ P^{-1}(R/P)^{(1-\beta_\theta)/\theta}, & R \gtrsim R_P. \end{cases}$$

Proof. Let $k \asymp \kappa_{R,P}$, and set $\widetilde{k} := \min\{k, \lfloor M/3 \rfloor\}$. By the general upper bound and Lemma E.12,

$$\mathrm{bias}_L(\mathcal{D}) \lesssim \frac{1}{R}\widetilde{k}^{1-2r} + \widetilde{k}^{1-\beta_\alpha} + P^{-1}\widetilde{k}^{1-\beta_\theta}.$$

Since $R^{-1} \asymp k^{-\alpha} + P^{-1}k^{-\theta}$, the first term has the same order as the last two. If $k > M/3$, the unsaturated condition gives $k \asymp M$, so $\widetilde{k} \asymp k$. Hence

$$\mathrm{bias}_L(\mathcal{D}) \lesssim k^{1-\beta_\alpha} + P^{-1}k^{1-\beta_\theta}.$$

For the lower inequality, apply Lemma C.2. This gives the same two-scale lower bound whenever $k \le cM$. The piecewise form follows from Lemma E.10. □

D VARIANCE BOUND

All expectations are over the sampled trajectories conditional on a fixed Gaussian sketch. Recall

$$Q := \widehat{E} = \widehat{C} - B_*\widehat{G}, \qquad \mathrm{Var}_L := \frac{1}{2}\mathbb{E}\|Q\widehat{g}_L(\widehat{G})\|_G^2,$$

where the empirical filter uses the safeguarded steps. Define its covariance-event contribution by

$$\mathrm{Var}_L^{\mathrm{good}} := \frac{1}{2}\mathbb{E}\left[\|Q\widehat{g}_L(\widehat{G})\|_G^2 \mathbf{1}_{\mathcal{E}_{\mathrm{cov}}(R)}\right].$$

Explicitly,

$$\widehat{g}_L(\widehat{G}) := \sum_{t=1}^{L} \bar{\gamma}_t \prod_{q=t+1}^{L} (I - \bar{\gamma}_q \widehat{G}).$$

On $\mathcal{E}_{\mathrm{cov}}(R)$, the safeguard is inactive. Write

$$G = \sum_{j=1}^{M} \mu_j v_j v_j^\top, \qquad g_L(s) := \sum_{t=1}^{L} \gamma_t \prod_{q=t+1}^{L} (1 - \gamma_q s),$$

and define the general effective dimension

$$d_{\mathrm{eff}}(R, M, P) := \sum_{j=1}^{M} \min\{1, (R\mu_j)^2\}.$$

D.1 GENERAL GD-VARIANCE UPPER BOUND

Proposition D.1 (General GD-variance upper bound)**.** Suppose the sample conditions in Lemma E.7 hold. On the Gaussian-sketch event,

$$\mathrm{Var}_L \lesssim \frac{C_\rho}{NP}\left\{d_{\mathrm{eff}}(R, M, P) + R^2 e^{-cM/2}\right\}.$$

Proof. Set $\lambda = R^{-1}$ and $K_\lambda(G) := G(G + \lambda I)^{-2}$. On $\mathcal{E}_{\mathrm{cov}}(R)$, the noncommutative response-filter comparison in Lemma E.13 gives

$$\|Q\widehat{g}_L(\widehat{G})\|_G^2 \lesssim \mathrm{Tr}\left[Q^\top Q K_\lambda(G)\right].$$

Here $\widehat{g}_L(\widehat{G}) = g_L(\widehat{G})$ because the safeguard is inactive on the event. Using Lemma E.5,

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}\left[Q^\top Q K_\lambda(G)\right] &= \mathrm{Tr}\left[K_\lambda(G)\mathbb{E}(Q^\top Q)\right] \\
&\lesssim \frac{C_\rho}{NP}\sum_{j=1}^{M}\frac{\mu_j^2}{(\mu_j + \lambda)^2} \\
&\asymp \frac{C_\rho}{NP} d_{\mathrm{eff}}(R, M, P),
\end{aligned}$$

because $\mu_j^2/(\mu_j + \lambda)^2 \asymp \min\{1, (R\mu_j)^2\}$.

On the complement event, global safeguarding gives $\|\widehat{g}_L(\widehat{G})\| \le R$. Therefore

$$\|Q\widehat{g}_L(\widehat{G})\|_G^2 \lesssim R^2\|Q\|_F^2.$$

Lemma E.6, Cauchy–Schwarz, and Lemma E.7 yield

$$\begin{aligned}
\mathbb{E}\left[\|Q\|_F^2 \mathbf{1}_{\mathcal{E}_{\mathrm{cov}}(R)^c}\right] &\lesssim \mathbb{E}\|Q\|_F^2 e^{-cM/2} \\
&\lesssim \frac{C_\rho}{NP} e^{-cM/2},
\end{aligned}$$

because $\mathrm{Tr}(G) \asymp 1$ on the sketch event. Adding the good- and bad-event contributions proves the proposition. □

D.2 General GD-variance lower bound

Proposition D.2 (General GD-variance lower bound)**.** Under the same sketch and sample conditions, for sufficiently large M,

$$\mathrm{Var}_L \gtrsim \frac{c_{\rho,C_{\mathrm{sm}}}}{NP} d_{\mathrm{eff}}(R,M,P).$$

Proof. The one-step smoothing condition in Assumption 2 allows Lemma E.4 to be applied without stationarity. Set $\lambda = R^{-1}$, $K_\lambda(G) := G(G+\lambda I)^{-2}$, and

$$Y_\lambda := \mathrm{Tr}\left[Q^\top Q K_\lambda(G)\right] = \|QK_\lambda(G)^{1/2}\|_F^2.$$

Then

$$\begin{aligned}\mathbb{E}Y_\lambda &\gtrsim \frac{c_{\rho,C_{\mathrm{sm}}}}{NP}\,\mathrm{Tr}\left[GK_\lambda(G)\right]\\ &\asymp \frac{c_{\rho,C_{\mathrm{sm}}}}{NP} d_{\mathrm{eff}}(R,M,P).\end{aligned}$$

On the covariance event, the deterministic lower comparison in Lemma E.13 gives

$$\mathrm{Var}_L \gtrsim \mathbb{E}\left[Y_\lambda \mathbf{1}_{\mathcal{E}_{\mathrm{cov}}(R)}\right].$$

By Lemma E.6,

$$\mathbb{E}\left[Y_\lambda \mathbf{1}_{\mathcal{E}_{\mathrm{cov}}(R)^c}\right] \lesssim e^{-cM/2}\mathbb{E}Y_\lambda.$$

Thus restriction to the covariance event removes only an exponentially small fraction of $\mathbb{E}Y_\lambda$, proving the lower bound. □

D.3 Specific GD-variance bound for $\theta \geq \alpha$

Theorem D.1 (GD-variance bound for $\theta \geq \alpha$)**.** *Suppose $\theta \geq \alpha$,*

$$1 \lesssim R \lesssim M^\alpha, \qquad NP \gtrsim C_\rho RM.$$

Then, on the Gaussian-sketch event,

$$\frac{c_{\rho,C_{\mathrm{sm}}}}{NP}\min\{M, R^{1/\alpha}\} \lesssim \mathrm{Var}_L \lesssim \frac{C_\rho}{NP}\left\{\min\{M, R^{1/\alpha}\} + R^2 e^{-cM/2}\right\}.$$

The boundary case $\theta = \alpha$ is included.

Proof. Lemma E.11 gives $\mu_j \asymp j^{-\alpha}$. Splitting the effective-dimension sum at $j \asymp R^{1/\alpha}$ yields

$$d_{\mathrm{eff}}(R,M,P) \asymp \min\{M, R^{1/\alpha}\}.$$

Apply Propositions D.1 and D.2. □

D.4 Specific GD-variance bound for $\alpha - 2r \leq \theta < \alpha$

Theorem D.2 (GD-variance bound for heavier initialization)**.** *Suppose $\alpha - 2r \leq \theta < \alpha$,*

$$1 \lesssim R \lesssim \left(M^{-\alpha} + P^{-1}M^{-\theta}\right)^{-1},$$

and

$$NP \gtrsim C_\rho RM, \qquad N \gtrsim RM.$$

Then, on the Gaussian-sketch event,

$$\frac{c_{\rho,C_{\mathrm{sm}}}}{NP}\min\{M, \kappa_{R,P}\} \lesssim \mathrm{Var}_L \lesssim \frac{C_\rho}{NP}\left\{\min\{M, \kappa_{R,P}\} + R^2 e^{-cM/2}\right\}.$$

Equivalently, with $R_P = P^{\alpha/(\alpha-\theta)}$, the polynomial term is

$$\mathrm{Var}_L \asymp \frac{1}{NP}\begin{cases}\min\{M, R^{1/\alpha}\}, & R \lesssim R_P,\\ \min\{M, (R/P)^{1/\theta}\}, & R \gtrsim R_P.\end{cases}$$

In the initialization-dominated unsaturated regime,

$$\mathrm{Var}_L \asymp \frac{R^{1/\theta}}{NP^{1+1/\theta}}.$$

Proof. The mixed-spectrum transfer lemma and Lemma E.10 give

$$d_{\text{eff}}(R, M, P) \asymp \min\{M, \kappa_{R,P}\}.$$

The two general propositions give the matching bounds. Substituting the two forms of $\kappa_{R,P}$ yields the piecewise display. □

E AUXILIARY LEMMAS

Lemma E.1 (Primitive dynamics imply a two-scale covariance decomposition)**.** *Suppose Assumptions 1 and 2 hold and $P \geq 2$. Write*

$$\Sigma_p = \underbrace{A_*^p \Sigma_0 (A_*^\top)^p}_{\Sigma_p^{(0)}} + \underbrace{\sum_{q=0}^{p-1} A_*^q \Sigma_\xi (A_*^\top)^q}_{\Sigma_p^{(\xi)}},$$

and let H_0 and H_ξ be the respective position averages. Set

$$G_0 := S H_0 S^\top, \qquad G_\xi := S H_\xi S^\top, \qquad G = G_0 + G_\xi.$$

Then

$$h_{0,P,j} := e_j^\top H_0 e_j \asymp P^{-1} j^{-\theta}, \qquad h_{\xi,P,j} := e_j^\top H_\xi e_j \asymp j^{-\alpha},$$

so

$$h_{P,j} := e_j^\top H e_j \asymp j^{-\alpha} + P^{-1} j^{-\theta}.$$

Moreover,

$$A_* \Sigma_0 A_*^\top \preceq C_{\text{sm}} \Sigma_\xi, \qquad A_* \Sigma_p A_*^\top \preceq C_{\rho, C_{\text{sm}}} \Sigma_\xi$$

uniformly in p, and $\Sigma_p^{(\xi)} \preceq C_\rho H_\xi$. The exact identities

$$C_p = S A_* \Sigma_p S^\top, \qquad C = S A_* H S^\top, \qquad B_* = S A_* H S^\top (S H S^\top)^{-1}$$

remain valid. In Regime II, no uniform comparison $\Sigma_p \preceq CH$ is asserted.

Proof. In the common eigenbasis,

$$\sigma_{p,j} = a_j^{2p} \sigma_{0,j} + \sum_{q=0}^{p-1} a_j^{2q} \sigma_{\xi,j}.$$

Therefore

$$h_{0,P,j} = \sigma_{0,j} \frac{1}{P} \sum_{p=0}^{P-1} a_j^{2p} \asymp P^{-1} j^{-\theta},$$

because the geometric sum is between 1 and $(1-\rho^2)^{-1}$. Also,

$$h_{\xi,P,j} = \sigma_{\xi,j} \frac{1}{P} \sum_{q=0}^{P-2} (P-1-q) a_j^{2q} \asymp j^{-\alpha},$$

because the weighted sum lies between $(P-1)/P$ and $(1-\rho^2)^{-1}$. The smoothing condition follows from

$$a_j^2 \sigma_{0,j} \asymp j^{-(\theta+2r)} \lesssim j^{-\alpha} \asymp \sigma_{\xi,j}.$$

Applying the recursion once more gives

$$A_* \Sigma_p A_*^\top \preceq \left(C_{\text{sm}} + \frac{\rho^2}{1-\rho^2} \right) \Sigma_\xi.$$

Finally, $\Sigma_p^{(\xi)} \preceq (1-\rho^2)^{-1} \Sigma_\xi \preceq C_\rho H_\xi$. The cross-covariance identities follow from innovation independence and position averaging. □

Lemma E.2 (Componentwise geometric dependence)**.** *Under the conditions of Lemma E.1, write* $x_{i,p} = x_{i,p}^{(0)} + x_{i,p}^{(\xi)}$*, where*

$$x_{i,p}^{(0)} := A_*^p x_{i,0}, \qquad x_{i,p}^{(\xi)} := \sum_{q=0}^{p-1} A_*^q \xi_{i,p-q}.$$

For $q \geq p$*,*

$$\mathbb{E}[x_{i,q}^{(\xi)} (x_{i,p}^{(\xi)})^\top] = A_*^{q-p} \Sigma_p^{(\xi)},$$

and, with $G_\xi := S H_\xi S^\top$*,*

$$\left\| G_\xi^{-1/2} \mathbb{E}[u_{i,q}^{(\xi)} (u_{i,p}^{(\xi)})^\top] G_\xi^{-1/2} \right\|_{\mathrm{op}} \lesssim \rho^{q-p}.$$

The initial component satisfies

$$\mathbb{E}[x_{i,q}^{(0)} (x_{i,p}^{(0)})^\top] = A_*^q \Sigma_0 (A_*^\top)^p,$$

whose norm decays geometrically in $p + q$*. Thus innovation-driven quadratic forms have effective sample size* NP*, whereas the initialization component is an average of* N *independent Gaussian quadratic forms.*

Proof. The identities follow from the two independent components of the state recursion. By Lemma E.1,

$$A_*^{q-p} \Sigma_p^{(\xi)} (A_*^\top)^{q-p} \preceq C \rho^{2(q-p)} H_\xi, \qquad \Sigma_p^{(\xi)} \preceq C H_\xi.$$

The same factorization and Cauchy–Schwarz argument used for normalized cross-covariances gives the displayed sketched bound. The initial-state claim follows from $\|A_*^p\|_{\mathrm{op}} \leq \rho^p$. □

Lemma E.3 (Innovation–corrector coercivity)**.** *Let* $\Sigma \succeq 0$ *and* $\Sigma_\xi \succ 0$*, and suppose*

$$A_* \Sigma A_*^\top \preceq \Lambda^2 \Sigma_\xi.$$

Then, for every symmetric matrix K *and vectors* r, c*,*

$$\begin{aligned} &\left\| \Sigma_\xi^{1/2} \left(r c^\top + 2 K A_* \right) \Sigma^{1/2} \right\|_F^2 + 2 \left\| \Sigma_\xi^{1/2} K \Sigma_\xi^{1/2} \right\|_F^2 \\ &\qquad \geq \frac{1}{1 + 2\Lambda^2} (r^\top \Sigma_\xi r)(c^\top \Sigma c). \end{aligned}$$

Proof. Set

$$X := \Sigma_\xi^{1/2} r c^\top \Sigma^{1/2}, \qquad U := 2 \Sigma_\xi^{1/2} K A_* \Sigma^{1/2}, \qquad W := \Sigma_\xi^{1/2} K \Sigma_\xi^{1/2}.$$

The covariance comparison implies

$$\left\| \Sigma_\xi^{-1/2} A_* \Sigma^{1/2} \right\|_{\mathrm{op}} \leq \Lambda,$$

and therefore $\|U\|_F \leq 2\Lambda \|W\|_F$. If $\Lambda = 0$, the claim is immediate. Otherwise,

$$2\|W\|_F^2 \geq \frac{1}{2\Lambda^2} \|U\|_F^2.$$

For any vectors in a Hilbert space and any $a > 0$, completing the square gives

$$\|X + U\|^2 + a\|U\|^2 \geq \frac{a}{1 + a} \|X\|^2.$$

Taking $a = (2\Lambda^2)^{-1}$ proves the result because $\|X\|_F^2 = (r^\top \Sigma_\xi r)(c^\top \Sigma c)$. □

Lemma E.4 (Nonstationary trajectory-score lower bound). *Suppose $\|A_*\|_{\mathrm{op}} \le \rho < 1$, the matrices A_*, Σ_0, and Σ_ξ are diagonal in a common basis, and the one-step smoothing condition*

$$A_*\Sigma_0 A_*^\top \preceq C_{\mathrm{sm}}\Sigma_\xi$$

holds. Let

$$D := SA_* - B_*S, \qquad \Omega_\xi := S\Sigma_\xi S^\top,$$

and

$$Q := \widehat{C} - B_*\widehat{G} = \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} e_{i,p}u_{i,p}^\top.$$

There is a constant $c_{\rho,C_{\mathrm{sm}}} > 0$, independent of M, N, P, L, such that

$$\mathbb{E}_{\mathcal{D}}[Q^\top Q \mid S] \succeq \frac{c_{\rho,C_{\mathrm{sm}}}\,\mathrm{Tr}(\Omega_\xi)}{NP} G.$$

On the Gaussian-sketch event of Lemma E.11, $\mathrm{Tr}(\Omega_\xi) \asymp 1$, and hence

$$\mathbb{E}_{\mathcal{D}}[Q^\top Q \mid S] \succeq \frac{c_{\rho,C_{\mathrm{sm}}}}{NP} G.$$

No stationarity assumption is used.

Proof. Fix $v \in \mathbb{R}^M$ and one trajectory, suppressing its trajectory index. Define

$$T(v) := \sum_{p=0}^{P-1} e_p(v^\top u_p).$$

The averaged normal equation gives

$$DHS^\top = SA_*HS^\top - B_*SHS^\top = C - B_*G = 0.$$

Therefore

$$\mathbb{E}[T(v)] = \sum_{p=0}^{P-1} D\Sigma_p S^\top v = PDHS^\top v = 0.$$

Let f_m denote the m-th standard basis vector of $\mathbb{R}^M$, and set

$$r_m := S^\top f_m, \qquad d_m := D^\top f_m, \qquad c := S^\top v, \qquad F_m := \frac{1}{2}(d_m c^\top + c d_m^\top).$$

Write $T_m(v) := f_m^\top T(v)$. Its position-p contribution is

$$f_m^\top e_p(v^\top u_p) = x_p^\top F_m x_p + (r_m^\top \xi_{p+1})(c^\top x_p).$$

Define

$$q_{p,m}(x) := x^\top F_m x - \mathrm{Tr}(F_m\Sigma_p).$$

Since

$$\sum_{p=0}^{P-1} \mathrm{Tr}(F_m\Sigma_p) = Pd_m^\top Hc = Pf_m^\top DHS^\top v = 0,$$

the full trajectory score is unchanged if each quadratic term is replaced by $q_{p,m}$.

For each output coordinate, define the finite-horizon backward sequence

$$K_{P,m} := 0, \qquad K_{p,m} := F_m + A_*^\top K_{p+1,m}A_*, \qquad p = P-1, \dots, 0,$$

and

$$h_{p,m}(x) := x^\top K_{p,m}x - \mathrm{Tr}(K_{p,m}\Sigma_p).$$

Let $\mathcal{F}_p := \sigma(x_0, \xi_1, \dots, \xi_p)$. Using $x_{p+1} = A_*x_p + \xi_{p+1}$ and $\Sigma_{p+1} = A_*\Sigma_p A_*^\top + \Sigma_\xi$, direct substitution gives the finite-horizon Poisson identity

$$q_{p,m}(x_p) = h_{p,m}(x_p) - \mathbb{E}[h_{p+1,m}(x_{p+1}) \mid \mathcal{F}_p].$$

Define

$$M_{p+1,m} := (r_m^\top \xi_{p+1})(c^\top x_p) + h_{p+1,m}(x_{p+1}) - \mathbb{E}[h_{p+1,m}(x_{p+1}) \mid \mathcal{F}_p].$$

Then $\mathbb{E}[M_{p+1,m} \mid \mathcal{F}_p] = 0$, and summing the martingale–coboundary identity over positions yields

$$T_m(v) = h_{0,m}(x_0) + \sum_{p=0}^{P-1} M_{p+1,m},$$

because $h_{P,m} = 0$. The initial term and the martingale differences are mutually orthogonal in L^2, so

$$\mathbb{E}[T_m(v)^2] \geq \sum_{p=0}^{P-1} \mathbb{E}[M_{p+1,m}^2].$$

Expanding the conditional-centering term gives

$$M_{p+1,m} = \xi_{p+1}^\top \left(r_m c^\top + 2K_{p+1,m}A_*\right) x_p + \xi_{p+1}^\top K_{p+1,m}\xi_{p+1} - \operatorname{Tr}(K_{p+1,m}\Sigma_\xi).$$

The bilinear term and the centered quadratic term are orthogonal by Gaussian symmetry. Therefore

$$\mathbb{E}[M_{p+1,m}^2] = \left\| \Sigma_\xi^{1/2} \left(r_m c^\top + 2K_{p+1,m}A_*\right) \Sigma_p^{1/2} \right\|_F^2 + 2 \left\| \Sigma_\xi^{1/2} K_{p+1,m} \Sigma_\xi^{1/2} \right\|_F^2.$$

It remains to verify the uniform covariance comparison needed for Lemma E.3. In the common eigenbasis,

$$\sigma_{p,j} = a_j^{2p}\sigma_{0,j} + \sum_{q=0}^{p-1} a_j^{2q}\sigma_{\xi,j},$$

and hence

$$A_*\Sigma_p A_*^\top = A_*^{p+1}\Sigma_0 (A_*^\top)^{p+1} + \sum_{q=1}^{p} A_*^q \Sigma_\xi (A_*^\top)^q \preceq \Lambda_{\rho,C_{\mathrm{sm}}}^2 \Sigma_\xi, \qquad \Lambda_{\rho,C_{\mathrm{sm}}}^2 := C_{\mathrm{sm}} + \frac{\rho^2}{1-\rho^2},$$

uniformly in p. Lemma E.3 now gives

$$\mathbb{E}[M_{p+1,m}^2] \geq c_{\rho,C_{\mathrm{sm}}} (r_m^\top \Sigma_\xi r_m)(c^\top \Sigma_p c).$$

Summing over output coordinates and positions,

$$\mathbb{E}\|T(v)\|_2^2 \geq c_{\rho,C_{\mathrm{sm}}} \operatorname{Tr}(\Omega_\xi) \sum_{p=0}^{P-1} v^\top S \Sigma_p S^\top v = c_{\rho,C_{\mathrm{sm}}} P \operatorname{Tr}(\Omega_\xi) v^\top G v.$$

The trajectory scores are independent and centered across i. Since $Qv = (NP)^{-1}\sum_{i=1}^N T_i(v)$,

$$\mathbb{E}\|Qv\|_2^2 = \frac{1}{NP^2}\mathbb{E}\|T(v)\|_2^2 \geq \frac{c_{\rho,C_{\mathrm{sm}}} \operatorname{Tr}(\Omega_\xi)}{NP} v^\top G v.$$

This proves the Loewner lower bound. Finally, $\operatorname{Tr}(\Sigma_\xi) \asymp 1$ because $\alpha > 1$, and standard Gaussian quadratic-form concentration gives $\operatorname{Tr}(S\Sigma_\xi S^\top) \asymp 1$ on the sketch event. □

Lemma E.5 (Residual trajectory-score covariance upper bound)**.** *Under Assumptions 1 and 2, there is a Gaussian-sketch event with probability at least* $1 - \exp(-cM)$ *on which the residual score*

$$Q := \widehat{C} - B_*\widehat{G} = \frac{1}{NP}\sum_{i=1}^N \sum_{p=0}^{P-1} \left\{ e_{i,p}u_{i,p}^\top - \mathbb{E}[e_{i,p}u_{i,p}^\top] \right\}$$

satisfies

$$\mathbb{E}_{\mathcal{D}}[Q^\top Q \mid S] \preceq \frac{C_{\rho,C_{\mathrm{sm}}}}{NP} G.$$

On the same event,

$$\|B_*\|_{\mathrm{op}} \le C, \qquad \mathrm{Tr}(D\Sigma_0 D^\top) \le C, \qquad D := SA_* - B_*S.$$

Proof. We first record two sketch estimates used to control the initial boundary term in the finite-horizon decomposition. Fix a sufficiently small constant $c_0 \in (0, 1/3)$, set $k = \lfloor c_0 M \rfloor$, and split the primitive eigenbasis into a head and a tail. Write

$$G = G_{\mathrm{h}} + G_{\mathrm{t}}, \qquad C = C_{\mathrm{h}} + C_{\mathrm{t}},$$

where

$$G_{\mathrm{h}} := S_{\mathrm{h}} H_{\mathrm{h}} S_{\mathrm{h}}^\top, \qquad C_{\mathrm{h}} := S_{\mathrm{h}} H_{\mathrm{h}} A_{\mathrm{h}} S_{\mathrm{h}}^\top,$$
$$G_{\mathrm{t}} := S_{\mathrm{t}} H_{\mathrm{t}} S_{\mathrm{t}}^\top, \qquad C_{\mathrm{t}} := S_{\mathrm{t}} H_{\mathrm{t}} A_{\mathrm{t}} S_{\mathrm{t}}^\top.$$

The finite mixed sequence $(h_{P,j})_{j=1}^d$ is decreasing, doubling on the index range used below, and satisfies

$$\frac{1}{M} \sum_{j>k} h_{P,j} \asymp h_{P,M}$$

because $k \asymp M$; the finite sum has the same order because it contains a fixed annulus beyond k. The condition $d \ge C_{\mathrm{amb}} M$ ensures that the next $2M$ tail columns exist. Weighted Gaussian tail concentration, together with the lower bound obtained from those columns, therefore gives

$$c h_{P,M} I \preceq G_{\mathrm{t}} \preceq C h_{P,M} I.$$

On the same event, rectangular Gaussian concentration gives

$$\|S_{\mathrm{h}}\|_{\mathrm{op}} \le C, \qquad S_{\mathrm{h}}^\top S_{\mathrm{h}} \succeq cI.$$

These events have joint probability at least $1 - \exp(-cM)$, uniformly in P.

Woodbury's identity yields

$$G^{-1} S_{\mathrm{h}} H_{\mathrm{h}} = G_{\mathrm{t}}^{-1} S_{\mathrm{h}} \left(H_{\mathrm{h}}^{-1} + S_{\mathrm{h}}^\top G_{\mathrm{t}}^{-1} S_{\mathrm{h}} \right)^{-1}.$$

By the preceding two displays,

$$S_{\mathrm{h}}^\top G_{\mathrm{t}}^{-1} S_{\mathrm{h}} \succeq c h_{P,M}^{-1} I,$$

and hence

$$\left\| G^{-1} S_{\mathrm{h}} H_{\mathrm{h}} \right\|_{\mathrm{op}} \le C.$$

It follows that

$$\|G^{-1} C_{\mathrm{h}}\|_{\mathrm{op}} \le \left\| G^{-1} S_{\mathrm{h}} H_{\mathrm{h}} \right\|_{\mathrm{op}} \|A_{\mathrm{h}} S_{\mathrm{h}}^\top\|_{\mathrm{op}} \le C.$$

Let $\bar{a}_k := \sup_{j>k} |a_j|$. Since A_* and H are diagonal in the same basis,

$$\left\| G_{\mathrm{t}}^{-1/2} C_{\mathrm{t}} G_{\mathrm{t}}^{-1/2} \right\|_{\mathrm{op}} \le \bar{a}_k \le \rho.$$

Furthermore, $G \succeq G_{\mathrm{t}}$, so

$$\|G^{-1} G_{\mathrm{t}}\|_{\mathrm{op}} \le \sqrt{\mathrm{cond}(G_{\mathrm{t}})} \le C,$$
$$\|G_{\mathrm{t}}^{-1} C_{\mathrm{t}}\|_{\mathrm{op}} \le \bar{a}_k \sqrt{\mathrm{cond}(G_{\mathrm{t}})} \le C.$$

Consequently,

$$\|G^{-1} C_{\mathrm{t}}\|_{\mathrm{op}} \le \|G^{-1} G_{\mathrm{t}}\|_{\mathrm{op}} \|G_{\mathrm{t}}^{-1} C_{\mathrm{t}}\|_{\mathrm{op}} \le C.$$

Because $B_*^\top = G^{-1} C$, the head and tail estimates prove

$$\|B_*\|_{\mathrm{op}} \le C.$$

The primitive trace sequences are summable:

$$\mathrm{Tr}(\Sigma_0) \lesssim 1, \qquad \mathrm{Tr}(A_* \Sigma_0 A_*^\top) \lesssim 1, \qquad \mathrm{Tr}(\Sigma_\xi) \asymp 1, \qquad \mathrm{Tr}(A_* H A_*^\top) \lesssim 1,$$

because $\alpha, \theta > 1$ and $\beta_\alpha, \beta_\theta > 1$. Gaussian quadratic-form concentration and the operator bound on B_* imply

$$\begin{aligned}\operatorname{Tr}(S\Sigma_0 S^\top) &\le C,\\ \operatorname{Tr}(SA_*\Sigma_0 A_*^\top S^\top) &\le C,\\ \operatorname{Tr}(\Omega_\xi) &\asymp 1,\\ \operatorname{Tr}(SA_*HA_*^\top S^\top) &\le C,\\ \operatorname{Tr}(B_*S\Sigma_0 S^\top B_*^\top) &\le C.\end{aligned}$$

Therefore, for $D := SA_* - B_*S$,

$$\operatorname{Tr}(D\Sigma_0 D^\top) \le 2\operatorname{Tr}(SA_*\Sigma_0 A_*^\top S^\top) + 2\operatorname{Tr}(B_*S\Sigma_0 S^\top B_*^\top) \le C.$$

We now apply the finite-horizon martingale–coboundary decomposition from Lemma E.4. For one trajectory and a deterministic v, write

$$T(v) := \sum_{p=0}^{P-1} \left\{ e_p(v^\top u_p) - \mathbb{E}[e_p(v^\top u_p)] \right\}.$$

The averaged normal equation implies that the sum of the subtracted expectations is zero. Thus this is the same trajectory score used in Lemma E.4. Retain its notation

$$r_m = S^\top f_m, \qquad d_m = D^\top f_m, \qquad c = S^\top v, \qquad F_m = \frac{1}{2}(d_m c^\top + c d_m^\top),$$

and its backward recursion

$$K_{P,m} = 0, \qquad K_{p,m} = F_m + A_*^\top K_{p+1,m} A_*.$$

Then

$$T_m(v) = h_{0,m}(x_0) + \sum_{p=0}^{P-1} M_{p+1,m},$$

where the initial term and all martingale differences are mutually orthogonal in L^2. Hence

$$\mathbb{E}\|T(v)\|_2^2 = \sum_{m=1}^{M} \mathbb{E}[h_{0,m}(x_0)^2] + \sum_{p=0}^{P-1}\sum_{m=1}^{M} \mathbb{E}[M_{p+1,m}^2].$$

Because A_* is diagonal, the backward recursion gives entrywise

$$(K_{p,m})_{j\ell} = (F_m)_{j\ell} \sum_{q=0}^{P-1-p} (a_j a_\ell)^q.$$

Thus, for every diagonal covariance $\Gamma \succeq 0$,

$$\left\| \Gamma^{1/2} K_{p,m} \Gamma^{1/2} \right\|_F^2 \le \frac{C}{(1-\rho^2)^2} (d_m^\top \Gamma d_m)(c^\top \Gamma c).$$

Indeed, the geometric multiplier is bounded by $(1-\rho^2)^{-1}$, and the claimed estimate follows by expanding the two rank-one terms in F_m.

Since x_0 is Gaussian,

$$\mathbb{E}[h_{0,m}(x_0)^2] = 2 \left\| \Sigma_0^{1/2} K_{0,m} \Sigma_0^{1/2} \right\|_F^2.$$

Summing this estimate with $\Gamma = \Sigma_0$ over output coordinates and using the preceding bound on $\operatorname{Tr}(D\Sigma_0 D^\top)$ gives

$$\begin{aligned}\sum_{m=1}^{M} \mathbb{E}[h_{0,m}(x_0)^2] &\le C \operatorname{Tr}(D\Sigma_0 D^\top) v^\top S\Sigma_0 S^\top v\\ &\le CP\, v^\top G_0 v \le CP\, v^\top G v.\end{aligned}$$

where the second inequality uses $H_0 \succeq P^{-1}\Sigma_0$.

For the martingale terms, the exact variance formula from Lemma E.4 is

$$\mathbb{E}[M_{p+1,m}^2] = \left\| \Sigma_\xi^{1/2} \left(r_m c^\top + 2K_{p+1,m} A_* \right) \Sigma_p^{1/2} \right\|_F^2 + 2 \left\| \Sigma_\xi^{1/2} K_{p+1,m} \Sigma_\xi^{1/2} \right\|_F^2 .$$

The direct innovation contribution satisfies

$$\sum_{m=1}^{M} \left\| \Sigma_\xi^{1/2} r_m c^\top \Sigma_p^{1/2} \right\|_F^2 = \mathrm{Tr}(\Omega_\xi) v^\top S \Sigma_p S^\top v.$$

The one-step smoothing estimate gives

$$A_* \Sigma_p A_*^\top \preceq \Lambda_{\rho, C_{\mathrm{sm}}}^2 \Sigma_\xi.$$

Consequently,

$$\left\| \Sigma_\xi^{1/2} K_{p+1,m} A_* \Sigma_p^{1/2} \right\|_F \le \Lambda_{\rho, C_{\mathrm{sm}}} \left\| \Sigma_\xi^{1/2} K_{p+1,m} \Sigma_\xi^{1/2} \right\|_F .$$

Applying the resolvent estimate with $\Gamma = \Sigma_\xi$, summing over m, and using $\Sigma_\xi \preceq CH$, yields

$$\sum_{m=1}^{M} \left\| \Sigma_\xi^{1/2} K_{p+1,m} \Sigma_\xi^{1/2} \right\|_F^2 \le C \,\mathrm{Tr}(D \Sigma_\xi D^\top) v^\top \Omega_\xi v \le C v^\top G v.$$

For the final inequality, the normal equation gives

$$DHD^\top = S A_* H A_*^\top S^\top - B_* G B_*^\top \preceq S A_* H A_*^\top S^\top,$$

whose trace is constant order on the sketch event because $\sum_j a_j^2 h_{P,j} < \infty$. Combining the direct and corrector estimates with $\|X+Y\|_F^2 \le 2\|X\|_F^2 + 2\|Y\|_F^2$ gives

$$\sum_{m=1}^{M} \mathbb{E}[M_{p+1,m}^2] \le C_{\rho, C_{\mathrm{sm}}} \left\{ \mathrm{Tr}(\Omega_\xi) v^\top S \Sigma_p S^\top v + v^\top G v \right\}.$$

Summing over positions and using $\sum_p S \Sigma_p S^\top = PG$ and $\mathrm{Tr}(\Omega_\xi) \asymp 1$,

$$\sum_{p=0}^{P-1} \sum_{m=1}^{M} \mathbb{E}[M_{p+1,m}^2] \le C_{\rho, C_{\mathrm{sm}}} P \, v^\top G v.$$

The orthogonal decomposition, initial-boundary estimate, and martingale estimate prove

$$\mathbb{E}\|T(v)\|_2^2 \le C_{\rho, C_{\mathrm{sm}}} P \, v^\top G v.$$

Finally, the complete trajectory scores are centered and independent across i. Since $Qv = (NP)^{-1} \sum_{i=1}^{N} T_i(v)$,

$$\mathbb{E}\|Qv\|_2^2 = \frac{1}{NP^2} \mathbb{E}\|T(v)\|_2^2 \le \frac{C_{\rho, C_{\mathrm{sm}}}}{NP} v^\top G v.$$

Since this holds for every $v \in \mathbb{R}^M$, the claimed Loewner bound follows. □

Lemma E.6 (Gaussian-chaos fourth-moment control)**.** *Conditional on a fixed sketch S, let W be any deterministic matrix of compatible dimension. Then*

$$\mathbb{E}_{\mathcal{D}} \|QW\|_F^4 \le C \left(\mathbb{E}_{\mathcal{D}} \|QW\|_F^2 \right)^2,$$

where C is an absolute constant. In particular, for $Y := \|Q g_L(G)\|_G^2$,

$$\mathbb{E}_{\mathcal{D}}[Y^2 \mid S] \le C \left(\mathbb{E}_{\mathcal{D}}[Y \mid S] \right)^2 .$$

Proof. Every state is a linear function of the independent Gaussian initial state and innovations. Consequently, every coordinate of QW is a polynomial of total degree at most two in independent Gaussian variables. Scalar Gaussian hypercontractivity gives

$$\mathbb{E}Z^4 \le C(\mathbb{E}Z^2)^2$$

for every such scalar polynomial Z. Writing the squared Frobenius norm as a sum over coordinates and applying Cauchy–Schwarz to every pair gives

$$\begin{aligned}\mathbb{E}\|QW\|_F^4 &= \sum_{k,\ell} \mathbb{E}[Z_k^2 Z_\ell^2] \\ &\le \sum_{k,\ell} (\mathbb{E}Z_k^4)^{1/2} (\mathbb{E}Z_\ell^4)^{1/2} \\ &\lesssim \left(\sum_k \mathbb{E}Z_k^2\right)^2.\end{aligned}$$

Taking $W = g_L(G)G^{1/2}$ proves the final claim. □

Lemma E.7 (High-probability covariance event for stable trajectories)**.** *Suppose Assumptions 1 and 2 hold and $P \ge 2$. Let*

$$\widehat{G} := \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} u_{i,p}u_{i,p}^\top, \qquad G = \frac{1}{P}\sum_{p=0}^{P-1} G_p.$$

Let $R = \gamma L$, and define

$$\omega_{0,M,P} := \sup_{1\le j\le 2M} \frac{P^{-1}j^{-\theta}}{j^{-\alpha} + P^{-1}j^{-\theta}} \asymp \begin{cases} P^{-1}, & \theta \ge \alpha, \\ \min\{1, M^{\alpha-\theta}/P\}, & \theta < \alpha. \end{cases}$$

There is an event $\mathcal{E}_S$, depending only on the Gaussian sketch and satisfying $\mathbb{P}_S(\mathcal{E}_S) \ge 1 - \exp(-cM)$, such that, conditional on every $S \in \mathcal{E}_S$, the following holds for every $t \ge 0$:

$$\begin{aligned}\left\|G^{-1/2}\widehat{G}G^{-1/2} - I\right\| \le C\Bigg\{ &\sqrt{\frac{C_\rho(M+t)}{NP}} + \frac{C_\rho(M+t)}{NP} \\ &+ \omega_{0,M,P}\sqrt{\frac{M+t}{N}} + \omega_{0,M,P}\frac{M+t}{N}\Bigg\},\end{aligned}$$

with conditional probability at least $1 - 2\exp(-ct)$, where $C_\rho \asymp (1+\rho)/(1-\rho)$. Consequently, $\mathcal{E}_{\mathrm{cov}}(R)$ holds with probability at least $1 - \exp(-\Omega(M))$ under either sufficient condition

$$\begin{aligned}&\theta \ge \alpha: && NP \gtrsim C_\rho RM, \\ &\alpha - 2r \le \theta < \alpha: && NP \gtrsim C_\rho RM \quad \textit{and} \quad N \gtrsim RM.\end{aligned}$$

Proof. We first record the relative size of the initialization component after sketching. Write

$$H_0 = \mathrm{diag}(h_{0,P,j}), \qquad H = \mathrm{diag}(h_{P,j}).$$

By Lemma E.1, uniformly in P,

$$h_{0,P,j} \asymp P^{-1}j^{-\theta}, \qquad h_{P,j} \asymp j^{-\alpha} + P^{-1}j^{-\theta}.$$

Let $S_{\le 2M}$ and $S_{>2M}$ denote the corresponding column blocks. The ambient condition $d \ge C_{\mathrm{amb}}M$ ensures that the first block and a further constant-width tail block are available. Standard rectangular-Gaussian singular-value concentration and a weighted Gaussian tail bound imply, simultaneously with probability at least $1 - \exp(-cM)$,

$$S_{\le 2M}H_{\le 2M}S_{\le 2M}^\top \succeq ch_{P,M}I, \qquad \left\|S_{>2M}(H_0)_{>2M}S_{>2M}^\top\right\|_{\mathrm{op}} \le Ch_{0,P,M}.$$

For completeness, the second estimate follows by applying scalar Bernstein to $\sum_{j>2M} h_{0,P,j}(a^\top s_j)^2$ on a $1/4$-net: its mean is $M^{-1}\sum_{j>2M} h_{0,P,j} \lesssim h_{0,P,M}$, and its largest summand scale is $h_{0,P,2M} \lesssim$

$h_{0,P,M}$. The first estimate follows from $H_{\le 2M} \succeq h_{P,2M} I$, $h_{P,2M} \asymp h_{P,M}$, and a constant lower bound on the smallest singular value of the M-by-$2M$ Gaussian block. These are the same head–tail estimates used in the proof of the Gaussian-sketch spectral lemma of Lin et al. (2024).

For $j \le 2M$, the definition of $\omega_{0,M,P}$ gives $h_{0,P,j} \le C\omega_{0,M,P} h_{P,j}$, while $h_{0,P,M}/h_{P,M} \le C\omega_{0,M,P}$. Therefore, on the preceding sketch event,

$$\begin{aligned} G_0 &\preceq C\omega_{0,M,P} S_{\le 2M} H_{\le 2M} S_{\le 2M}^\top + C h_{0,P,M} I \\ &\preceq C\omega_{0,M,P} G, \end{aligned}$$

and hence

$$\left\| G^{-1/2} G_0 G^{-1/2} \right\|_{\mathrm{op}} \le C\omega_{0,M,P}. \tag{1}$$

Fix a unit vector $a \in \mathbb{R}^M$ and, for one trajectory, define the Gaussian vector $z_i(a) \in \mathbb{R}^P$ by

$$z_{i,p}(a) := a^\top G^{-1/2} u_{i,p}, \qquad p = 0, \dots, P-1.$$

Let $K_a := \mathrm{Cov}(z_i(a))$, and decompose $K_a = K_{a,\xi} + K_{a,0}$ according to the independent innovation and initialization components. Since $G = P^{-1} \sum_p G_p$,

$$\mathrm{Tr}(K_a) = P.$$

The normalized geometric covariance estimate in Lemma E.2, together with $G_\xi \preceq G$, gives

$$|(K_{a,\xi})_{pq}| \le C\rho^{|p-q|}, \qquad \|K_{a,\xi}\|_{\mathrm{op}} \le C_\rho, \qquad \|K_{a,\xi}\|_F^2 \le C_\rho P. \tag{2}$$

Moreover, equation 1 implies

$$\mathrm{Tr}(K_{a,0}) = P a^\top G^{-1/2} G_0 G^{-1/2} a \le CP\omega_{0,M,P}.$$

Since $K_{a,0} \succeq 0$,

$$\|K_{a,0}\|_{\mathrm{op}} \le CP\omega_{0,M,P}, \qquad \|K_{a,0}\|_F \le CP\omega_{0,M,P}. \tag{3}$$

We now use an exact Gaussian quadratic-form inequality. If $g \sim \mathcal{N}(0, K)$, then diagonalizing K and applying the exponential moment bound for centered chi-square variables gives, for every $s \ge 0$,

$$\mathbb{P}\left(|\|g\|_2^2 - \mathrm{Tr}(K)| > 2\|K\|_F \sqrt{s} + 2\|K\|_{\mathrm{op}} s \right) \le 2e^{-s}. \tag{4}$$

Apply equation 4 to the block Gaussian vector formed by the N independent copies of $z_i(a)$. Using equation 2–equation 3, the triangle inequality for the operator and Frobenius norms, and dividing by NP, we obtain

$$\begin{aligned} &\left| a^\top G^{-1/2} (\widehat{G} - G) G^{-1/2} a \right| \\ &\quad \le C \left\{ \sqrt{\frac{C_\rho s}{NP}} + \frac{C_\rho s}{NP} + \omega_{0,M,P} \sqrt{\frac{s}{N}} + \omega_{0,M,P} \frac{s}{N} \right\} \end{aligned}$$

with probability at least $1 - 2e^{-s}$. Notice that the mixed initialization–innovation terms are already included in $\|z_i(a)\|_2^2$; no separate independence or decoupling claim is needed.

Finally, apply this fixed-direction estimate on a $1/4$-net of the unit sphere, whose cardinality is at most 9^M, with $s = C_1(M + t)$. The standard self-adjoint net inequality

$$\|A\|_{\mathrm{op}} \le 2 \sup_{a \in \mathcal{N}} |a^\top A a|$$

gives the asserted operator-norm bound after increasing the constants.

Take $t \asymp M$. The condition $NP \gtrsim C_\rho RM$ controls the innovation terms. If $\theta < \alpha$, then $\omega_{0,M,P} \le 1$, and $N \gtrsim RM$ controls the initialization terms. If $\theta \ge \alpha$, then $\omega_{0,M,P} \asymp P^{-1}$; the condition on NP implies

$$P^{-1} \sqrt{\frac{M}{N}} \lesssim R^{-1/2}, \qquad P^{-1} \frac{M}{N} \lesssim R^{-1},$$

so no separate condition on N is needed. This proves the two event statements. □

Lemma E.8 (Safeguarded stability and empirical cross-covariance moments)**.** *Suppose Assumptions 1 and 2 hold and $P \geq 2$. On the Gaussian-sketch event of Lemma E.11, there are constants $C_G, C > 0$ such that*

$$\|G\|_{\mathrm{op}} \leq C_G, \qquad \mathbb{E}_{\mathcal{D}}[\|\widehat{C}\|_F^4 \mid S] \leq C, \qquad \|B_*\|_F^2 \leq CM.$$

For every trajectory sample, the safeguarded GD schedule satisfies

$$0 \preceq I - \bar{\gamma}_t \widehat{G} \preceq I, \qquad \sum_{t=1}^{L} \bar{\gamma}_t \leq R, \qquad \|B_L\|_F \leq R\|\widehat{C}\|_F.$$

Moreover, if $\gamma \leq (4C_G)^{-1}$, then $\bar{\gamma} = \gamma$ on $\mathcal{E}_{\mathrm{cov}}(R)$.

Proof. The sketch-transfer event gives $\mu_j(G) \asymp j^{-\alpha} + P^{-1}j^{-\theta}$. Hence $\|G\|_{\mathrm{op}} \leq C_G$ and, because $\alpha, \theta > 1$,

$$\mathrm{Tr}(G) \asymp \sum_{j=1}^{M} \left(j^{-\alpha} + P^{-1}j^{-\theta}\right) \asymp 1.$$

Write $Z_{i,p} := y_{i,p}u_{i,p}^\top$, so that $\widehat{C} = (NP)^{-1}\sum_{i,p} Z_{i,p}$. Convexity of $A \mapsto \|A\|_F^4$ gives

$$\mathbb{E}_{\mathcal{D}}[\|\widehat{C}\|_F^4 \mid S] \leq \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} \mathbb{E}[\|Z_{i,p}\|_F^4 \mid S].$$

Since $\|Z_{i,p}\|_F^4 = \|y_{i,p}\|_2^4\|u_{i,p}\|_2^4$, Cauchy–Schwarz yields

$$\mathbb{E}\|Z_{i,p}\|_F^4 \leq \left(\mathbb{E}\|y_{i,p}\|_2^8 \mathbb{E}\|u_{i,p}\|_2^8\right)^{1/2}.$$

The vectors are Gaussian. The exact covariance recursion and $\alpha, \theta > 1$ give uniformly bounded traces for their positionwise covariances on the sketch event. The standard Gaussian moment bound $\mathbb{E}\|z\|_2^8 \lesssim \mathrm{Tr}(\mathrm{Cov}(z))^4$ therefore makes the preceding display uniformly bounded, proving the fourth moment claim.

Lemma E.5 gives $\|B_*\|_{\mathrm{op}} \leq C$ on the joint sketch event. Consequently,

$$\|B_*\|_F^2 \leq M\|B_*\|_{\mathrm{op}}^2 \lesssim M.$$

By construction, $\bar{\gamma}_t\|\widehat{G}\|_{\mathrm{op}} \leq 1/2$, which proves the contraction inequalities. Since every WSD multiplier is at most one, $\sum_t \bar{\gamma}_t \leq L\bar{\gamma} \leq R$. Since $B_0 = 0$, iteration of GD gives

$$B_L = \widehat{C}\sum_{t=1}^{L} \bar{\gamma}_t \prod_{s=t+1}^{L} (I - \bar{\gamma}_s \widehat{G}).$$

Every product has operator norm at most one, so $\|B_L\|_F \leq R\|\widehat{C}\|_F$.

Finally, on $\mathcal{E}_{\mathrm{cov}}(R)$, $\widehat{G} \preceq 2G$, and hence $\|\widehat{G}\|_{\mathrm{op}} \leq 2C_G$. Thus $\gamma \leq (4C_G)^{-1} \leq (2\|\widehat{G}\|_{\mathrm{op}})^{-1}$, proving that the safeguard is inactive on the good event. □

Lemma E.9 (Two-scale induced covariance and source)**.** *Suppose Assumptions 1 and 2 hold and $P \geq 2$. Then*

$$h_{P,j} := e_j^\top H e_j \asymp j^{-\alpha} + P^{-1}j^{-\theta},$$

and the source energy

$$\tau_{P,j} := h_{P,j}a_j^2 \asymp j^{-\beta_\alpha} + P^{-1}j^{-\beta_\theta}.$$

If $\theta < \alpha$, define

$$m_P := P^{1/(\alpha-\theta)}.$$

Then $h_{P,j} \asymp j^{-\alpha}$ for $j \lesssim m_P$, while $h_{P,j} \asymp P^{-1}j^{-\theta}$ for $j \gtrsim m_P$.

Proof. By Lemma E.1,

$$h_{P,j} \asymp P^{-1}j^{-\theta} + j^{-\alpha}.$$

Multiplication by $a_j^2 \asymp j^{-2r}$ gives the source formula. In the heavier-initialization regime, the two covariance terms are equal when $j^{\alpha-\theta} \asymp P$, which gives m_P and the two cases. □

Lemma E.10 (Mixed-spectrum optimization cutoff)**.** *For $R \gtrsim 1$, define $\kappa_{R,P}$ by*

$$R\left(\kappa_{R,P}^{-\alpha} + P^{-1}\kappa_{R,P}^{-\theta}\right) \asymp 1.$$

If $\theta < \alpha$, let $R_P := P^{\alpha/(\alpha-\theta)}$. Then

$$\kappa_{R,P} \asymp \begin{cases} R^{1/\alpha}, & R \lesssim R_P, \\ (R/P)^{1/\theta}, & R \gtrsim R_P. \end{cases}$$

Moreover, for the mixed sequence $\mu_j \asymp j^{-\alpha} + P^{-1}j^{-\theta}$,

$$\sum_{j=1}^{M} \min\{1, (R\mu_j)^2\} \asymp \min\{M, \kappa_{R,P}\}.$$

Proof. If $\theta \geq \alpha$, the initialization term is dominated uniformly by the innovation term, and hence $\kappa_{R,P} \asymp R^{1/\alpha}$. If $\theta < \alpha$, the cutoff lies before the spectral crossover when $Rm_P^{-\alpha} \lesssim 1$, equivalently $R \lesssim R_P$, and after it when $R \gtrsim R_P$. Substitution gives the two displayed formulas. For the effective dimension, split the sum at $j \asymp \kappa_{R,P}$. The head contributes $\min\{M, \kappa_{R,P}\}$; the tail has the same order by integral comparison, using $\alpha, \theta > 1$ and the doubling property of the mixed sequence. □

Lemma E.11 (Joint Gaussian-sketch spectral transfer)**.** *Suppose Assumptions 1 and 2 hold and S is a Gaussian sketch with entries $N(0, 1/M)$. Let*

$$G := SHS^\top, \qquad C := SA_*HS^\top,$$

and denote their eigenvalues in non-increasing order by $\mu_j(G)$ and $\nu_j(C)$, respectively. Then, with probability at least $1 - \exp(-cM)$, simultaneously for $j = 1, \ldots, M$,

$$\mu_j(G) \asymp j^{-\alpha} + P^{-1}j^{-\theta},$$

$$\nu_j(C) \asymp j^{-(\alpha+r)} + P^{-1}j^{-(\theta+r)}.$$

Moreover, for every integer $0 \leq k \leq M/3$, the tail covariance

$$G_{>k} := S_{>k}H_{>k}S_{>k}^\top$$

satisfies

$$\frac{\mu_{M/2}(G_{>k})}{\mu_M(G_{>k})} \leq C.$$

The event may be chosen so that the primitive weighted trace estimates used below hold simultaneously. In particular, for every $1 \leq k \leq M/3$,

$$\mathrm{Tr}\left(S_{\leq k}A_{*,\leq k}^2 S_{\leq k}^\top\right) \leq C\sum_{j\leq k} a_j^2,$$

$$\mathrm{Tr}\left(S_{>k}A_{*,>k}^2 H_{>k}S_{>k}^\top\right) \leq C\sum_{j>k} a_j^2 h_{P,j},$$

and $\mathrm{Tr}(S\Sigma_\xi S^\top) \asymp 1$. The event may also be chosen to be contained in the sketch event of Lemma E.5. All constants are uniform over $P \geq 2$.

Proof. We use the general Gaussian head–tail argument of Lin et al. (2024, Appendix G). For a decreasing finite diagonal sequence $D = \mathrm{diag}(d_1, \ldots, d_d)$, split at an index $m \leq cM$:

$$SDS^\top = S_{\leq m}D_{\leq m}S_{\leq m}^\top + \frac{1}{M}\sum_{j>m} d_j I + \left(S_{>m}D_{>m}S_{>m}^\top - \frac{1}{M}\sum_{j>m} d_j I\right).$$

Rectangular Gaussian concentration controls the head, while a weighted Bernstein net argument bounds the final centered tail by

$$C\left(d_{m+1} + \sqrt{\frac{1}{M}\sum_{j>m} d_j^2}\right).$$

The minimum eigenvalue is bounded below by retaining a block of $2M$ tail columns whose weights are comparable. Such a block exists by $d \geq C_{\mathrm{amb}} M$. Consequently, on the range $j \leq cM$, if

$$d_{2j} \asymp d_j, \qquad \sum_{\ell=j+1}^{d} d_\ell \asymp j d_j, \qquad \sum_{\ell=j+1}^{d} d_\ell^2 \lesssim j d_j^2,$$

then

$$\mu_j(SDS^\top) \asymp d_j, \qquad j = 1, \ldots, M.$$

Lemma E.9 gives

$$h_{P,j} \asymp j^{-\alpha} + P^{-1} j^{-\theta}.$$

This mixed sequence satisfies the three preceding conditions uniformly in P, by applying integral comparison to each power separately. Hence the first spectral claim follows.

Because A_* is positive definite and $a_j \asymp j^{-r}$, the diagonal sequence of $A_* H$ satisfies

$$a_j h_{P,j} \asymp j^{-(\alpha+r)} + P^{-1} j^{-(\theta+r)}.$$

It obeys the same doubling and tail conditions. Applying the same Gaussian spectral argument to $C = S(A_* H)S^\top$ proves the second claim. If the primitive sequence is not exactly monotone, reorder it; Gaussian column exchangeability leaves the sketch distribution unchanged, and the reordered sequence has the same two-power order.

For the tail-ratio statement, apply the same head–tail estimates to the finite shifted sequence $(h_{P,k+j})_{1 \leq j \leq d-k}$. When $k \leq M/3$, the indices $k + M/2$, $k + M$, and $k + 2M$ are comparable up to absolute constants. The ambient-dimension condition guarantees that these indices do not exceed d. The doubling and tail estimates therefore place both $\mu_{M/2}(G_{>k})$ and $\mu_M(G_{>k})$ at the same scale, proving the uniform ratio. Finally, the trace statements follow from scalar Gaussian quadratic-form concentration for the summable primitive sequences. A union bound over the two spectra, the $O(M)$ tail splits, and the finitely many trace events preserves probability $1 - \exp(-cM)$. □

Lemma E.12 (Aggregate Gaussian-sketch transfer for GD bias)**.** *On the event of Lemma E.11, let*

$$G = \sum_{j=1}^{M} \mu_j v_j v_j^\top, \qquad q_j := \mu_j \|B_* v_j\|_2^2.$$

For every $1 \leq k \leq M/3$ *and* $R \gtrsim 1$*,*

$$\sum_{j=1}^{M} q_j \min\{1, (R\mu_j)^{-1}\} \lesssim \frac{1}{R} k^{1-2r} + k^{1-\beta_\alpha} + P^{-1} k^{1-\beta_\theta}.$$

Let $k \asymp \kappa_{R,P}$*, so that* $R h_{P,k} \asymp 1$*. If* $k \leq cM$*, then, for a sufficiently small constant* $c_0 > 0$*,*

$$\sum_{j : R\mu_j \leq c_0} q_j \gtrsim k^{1-\beta_\alpha} + P^{-1} k^{1-\beta_\theta}.$$

Proof. For the upper bound, define the spectral filter

$$\phi_R(s)^2 := \min\{1, (Rs)^{-1}\}.$$

Then

$$\sum_{j=1}^{M} q_j \phi_R(\mu_j)^2 = \|B_* \phi_R(G)\|_G^2.$$

Let f_m be the m-th standard basis vector of $\mathbb{R}^M$, and define the ambient scalar target

$$w_m := A_* S^\top f_m.$$

Its population sketched coefficient is

$$B_*^\top f_m = G^{-1} S H w_m.$$

Split $w_m = w_{m,\le k} + w_{m,>k}$. Since $0 \preceq G\phi_R(G)^2 \preceq R^{-1}I$, the head contribution is bounded by

$$\frac{C}{R}\|w_{m,\le k}\|_2^2$$

provided $\|G^{-1}S_{\le k}H_{\le k}\|_{\mathrm{op}} \lesssim 1$. To verify this uniformly, set $T_k := G_{>k}$ and use Woodbury's identity:

$$G^{-1}S_{\le k}H_{\le k} = T_k^{-1}S_{\le k}\left(H_{\le k}^{-1} + S_{\le k}^\top T_k^{-1}S_{\le k}\right)^{-1}.$$

Conditionally on T_k, rotational invariance and rectangular Gaussian concentration on an $M/2$-dimensional eigenspace give

$$S_{\le k}^\top T_k^{-1}S_{\le k} \succeq \frac{c}{\mu_{M/2}(T_k)}I.$$

Thus

$$\|G^{-1}S_{\le k}H_{\le k}\|_{\mathrm{op}} \lesssim \frac{\mu_{M/2}(T_k)}{\mu_M(T_k)} \lesssim 1$$

by Lemma E.11.

For the tail, use $0 \preceq G\phi_R(G)^2 \preceq G$ and

$$\left\|H_{>k}^{1/2}S_{>k}^\top G^{-1}S_{>k}H_{>k}^{1/2}\right\|_{\mathrm{op}} \le 1.$$

This bounds its contribution by $\|w_{m,>k}\|_{H_{>k}}^2$. Summing over output coordinates gives

$$\begin{aligned}\|B_*\phi_R(G)\|_G^2 \lesssim \frac{1}{R}\,&\mathrm{Tr}\left(S_{\le k}A_{*,\le k}^2S_{\le k}^\top\right)\\ &+ \mathrm{Tr}\left(S_{>k}A_{*,>k}^2H_{>k}S_{>k}^\top\right).\end{aligned}$$

The simultaneous weighted trace estimates on the sketch event yield

$$\mathrm{Tr}\left(S_{\le k}A_{*,\le k}^2S_{\le k}^\top\right) \lesssim \sum_{j\le k} a_j^2 \asymp k^{1-2r},$$

and

$$\begin{aligned}\mathrm{Tr}\left(S_{>k}A_{*,>k}^2H_{>k}S_{>k}^\top\right) &\lesssim \sum_{j>k} a_j^2h_{P,j}\\ &\asymp k^{1-\beta_\alpha} + P^{-1}k^{1-\beta_\theta}.\end{aligned}$$

This proves the upper bound.

For the lower bound, let

$$\Pi_R := \mathbf{1}_{\{G\preceq c_0R^{-1}I\}}.$$

Since $B_* = CG^{-1}$,

$$\sum_{j:R\mu_j\le c_0} q_j = \mathrm{Tr}(\Pi_RG^{-1}C^2).$$

On the range of Π_R, $G^{-1} \succeq cRI$, and therefore

$$\mathrm{Tr}(\Pi_RG^{-1}C^2) = \mathrm{Tr}(C\Pi_RG^{-1}\Pi_RC) \gtrsim R\,\mathrm{Tr}(\Pi_RC^2).$$

The eigenvalue estimate for G shows that Π_R has rank at least $M - C_0k$. Ky Fan's minimum principle and the spectral estimate for C then imply

$$\begin{aligned}\mathrm{Tr}(\Pi_RC^2) &\ge \sum_{j>C_0k}\nu_j(C)^2\\ &\gtrsim k\,(a_kh_{P,k})^2,\end{aligned}$$

where $k \le cM$ ensures that the retained annulus has length comparable to k. Finally, $Rh_{P,k} \asymp 1$ gives

$$\begin{aligned}Rk(a_kh_{P,k})^2 &\asymp ka_k^2h_{P,k}\\ &\asymp k^{1-\beta_\alpha} + P^{-1}k^{1-\beta_\theta}.\end{aligned}$$

This proves the lower bound. □

Lemma E.13 (Scalar GD filters and noncommutative empirical transfer)**.** *Suppose Assumption 3 holds, $G \succ 0$, and $\|G\| \lesssim 1$. For the WSD step-size schedule in Section 2, let*

$$\psi_L(s) := \prod_{t=1}^{L}(1-\gamma_t s), \qquad g_L(s) := \sum_{t=1}^{L}\gamma_t \prod_{q=t+1}^{L}(1-\gamma_q s).$$

Then, for the effective horizon $R = \gamma L$ and every $s \geq 0$ satisfying $\gamma s \leq 1/2$,

$$\psi_L(s)^2 \lesssim \min\{1,(Rs)^{-1}\}, \qquad s^2 g_L(s)^2 \asymp \min\{1,(Rs)^2\}.$$

Let $\widehat{G} \succ 0$ satisfy

$$\delta := \left\| G^{-1/2}\widehat{G}G^{-1/2} - I \right\|_{\mathrm{op}} \leq c_{\mathrm{rel}} R^{-1/2},$$

with c_{rel} sufficiently small, and set $\lambda := R^{-1}$. Define

$$\widehat{\Psi}_L := \psi_L(\widehat{G}), \qquad \widetilde{g}_L := g_L(\widehat{G}), \qquad K_\lambda(G) := G(G+\lambda I)^{-2}.$$

Then the following deterministic Loewner comparisons hold:

$$\widehat{\Psi}_L G \widehat{\Psi}_L \preceq C\lambda G(G+\lambda I)^{-1},$$

and

$$cK_\lambda(G) \preceq \widetilde{g}_L G \widetilde{g}_L \preceq CK_\lambda(G).$$

In addition, if

$$\Pi_{\mathrm{low}} := \mathbf{1}_{[0,c_0\lambda]}(G),$$

then

$$\left\|(I-\widehat{\Psi}_L)\Pi_{\mathrm{low}}\right\|_{\mathrm{op}} \leq C\sqrt{c_0}.$$

Proof. Let $H_L := \sum_{t=1}^{L}\gamma_t$. The linear-length stable phase and the pointwise bound $0 \leq \gamma_t \leq \gamma$ give

$$c_{\mathrm{s}} R \leq H_L \leq R.$$

When $\gamma s \leq 1/2$, every scalar factor lies in $[0,1]$, and

$$\psi_L(s) \leq e^{-H_L s}.$$

Therefore

$$\psi_L(s)^2 \lesssim \min\{1,(Rs)^{-1}\}.$$

The telescoping identity gives $s g_L(s) = 1 - \psi_L(s)$. Moreover,

$$1 - e^{-H_L s} \leq 1 - \psi_L(s) \leq \min\{1, H_L s\},$$

and hence

$$s^2 g_L(s)^2 \asymp \min\{1,(Rs)^2\}.$$

Equivalently, up to constants,

$$s\psi_L(s)^2 \lesssim \frac{\lambda s}{s+\lambda}, \qquad s g_L(s)^2 \asymp \frac{s}{(s+\lambda)^2}.$$

We now prove the matrix transfer without assuming that G and $\widehat{G}$ commute. The relative event gives

$$(1-\delta)G \preceq \widehat{G} \preceq (1+\delta)G.$$

Since $G \preceq (1-\delta)^{-1}\widehat{G}$, congruence by $\widehat{\Psi}_L$ and the first scalar estimate above give

$$\begin{aligned}\widehat{\Psi}_L G \widehat{\Psi}_L &\preceq C\widehat{\Psi}_L \widehat{G} \widehat{\Psi}_L \\ &\preceq C\lambda \widehat{G}(\widehat{G}+\lambda I)^{-1}.\end{aligned}$$

The map $X \mapsto \lambda X(X+\lambda I)^{-1}$ is operator monotone on positive semidefinite matrices. Applying the upper comparison in the relative covariance display, followed by scalar comparison of $f((1+\delta)G)$ and $f(G)$, proves

$$\widehat{\Psi}_L G \widehat{\Psi}_L \preceq C\lambda G(G+\lambda I)^{-1}.$$

For the response filter, congruence in the relative covariance display and the second scalar estimate above first give

$$\widetilde{g}_L G \widetilde{g}_L \asymp \widetilde{g}_L \widehat{G} \widetilde{g}_L \asymp K_\lambda(\widehat{G}).$$

It remains to compare the two ridge-response kernels. Set

$$A := G + \lambda I, \qquad \widehat{A} := \widehat{G} + \lambda I, \qquad D := A^{-1/2} G A^{-1/2}.$$

Writing

$$\Delta := G^{-1/2}(\widehat{G} - G)G^{-1/2}, \qquad E := A^{-1/2}(\widehat{G} - G)A^{-1/2} = D^{1/2}\Delta D^{1/2},$$

we have $\|E\| \le \delta$. Let $T := (I + E)^{-1}$. Then

$$\widehat{A}^{-1} G \widehat{A}^{-1} = A^{-1/2} T D T A^{-1/2}.$$

The weighted similarity transform

$$L := D^{1/2} T D^{-1/2}$$

satisfies

$$L = I - D^{1/2} T D^{1/2} \Delta, \qquad L^{-1} = I + D\Delta.$$

Hence $\|L\| + \|L^{-1}\| \le C$, and therefore

$$TDT = D^{1/2} L^\top L D^{1/2}, \qquad cD \preceq TDT \preceq CD.$$

Consequently,

$$\widehat{A}^{-1} G \widehat{A}^{-1} \asymp A^{-1} G A^{-1} = K_\lambda(G).$$

Using the relative covariance comparison once more inside congruence by $\widehat{A}^{-1}$ yields

$$K_\lambda(\widehat{G}) = \widehat{A}^{-1} \widehat{G} \widehat{A}^{-1} \asymp \widehat{A}^{-1} G \widehat{A}^{-1} \asymp K_\lambda(G).$$

Together with the preceding response-filter display, this proves the two-sided response-filter comparison.

Finally, the telescoping identity $I - \widehat{\Psi}_L = \widehat{G}\widetilde{g}_L$, contraction of every GD factor, and $\sum_t \gamma_t \le R$ imply

$$0 \preceq I - \widehat{\Psi}_L \preceq R\widehat{G}, \qquad (I - \widehat{\Psi}_L)^2 \preceq I - \widehat{\Psi}_L.$$

Therefore

$$\begin{aligned} &\Pi_{\mathrm{low}}(I - \widehat{\Psi}_L)^2 \Pi_{\mathrm{low}} \\ &\quad \preceq R \Pi_{\mathrm{low}} \widehat{G} \Pi_{\mathrm{low}} \preceq C c_0 \Pi_{\mathrm{low}}, \end{aligned}$$

which proves the final norm estimate. This estimate alone is not used to claim a global empirical bias lower comparison; the source-aware trace argument in the next lemma supplies the required lower bound. □

Lemma E.14 (Source-aware empirical bias transfer)**.** *Suppose $G, \widehat{G} \succ 0$, $C = C^\top$, and*

$$B_* := CG^{-1}, \qquad \delta := \left\| G^{-1/2} \widehat{G} G^{-1/2} - I \right\|_{\mathrm{op}} \le \frac{1}{2}.$$

Set $\lambda := R^{-1}$, let $\widehat{\Psi}_L := \psi_L(\widehat{G})$, and define

$$\widehat{\Pi}_{\mathrm{low}} := \mathbf{1}_{[0, c_0 \lambda]}(\widehat{G}),$$

where $c_0 > 0$ is sufficiently small. Then

$$\|B_* \widehat{\Psi}_L\|_G^2 \ge cR \operatorname{Tr}\left(\widehat{\Pi}_{\mathrm{low}} C^2\right) - C\delta^2 \operatorname{Tr}\left(CG^{-1}C\right).$$

On the joint Gaussian-sketch event of Lemmas E.11 and E.5, let $k \asymp \kappa_{R,P}$ and suppose $k \le cM$. If, in addition, $\delta \le c_{\mathrm{rel}} R^{-1/2}$ with c_{rel} sufficiently small, then

$$\|B_* \widehat{\Psi}_L\|_G^2 \gtrsim k^{1-\beta_\alpha} + P^{-1} k^{1-\beta_\theta}.$$

Proof. The relative covariance comparison gives

$$(1-\delta)G \preceq \widehat{G} \preceq (1+\delta)G.$$

Define

$$T := \widehat{G}^{1/2} G^{-1} \widehat{G}^{1/2}, \qquad Z := \widehat{G}^{-1/2} C.$$

The eigenvalues of T lie in $[1-\delta, 1+\delta]$, and hence

$$\|T - I\|_{\mathrm{op}} \le \delta.$$

Because $G \succeq (1+\delta)^{-1}\widehat{G}$ and $B_*^\top = G^{-1}C$,

$$\begin{aligned} \|B_* \widehat{\Psi}_L\|_G^2 &\gtrsim \|B_* \widehat{\Psi}_L\|_{\widehat{G}}^2 \\ &= \left\| \widehat{\Psi}_L \widehat{G}^{1/2} G^{-1} C \right\|_F^2 \\ &= \|\widehat{\Psi}_L T Z\|_F^2. \end{aligned}$$

Every GD factor is positive semidefinite on the covariance event. Moreover, for $0 \le s \le c_0 R^{-1}$,

$$\psi_L(s) \ge 1 - s\sum_{t=1}^{L} \gamma_t \ge 1 - c_0 \ge \frac{1}{2},$$

after decreasing c_0. Since $\widehat{\Pi}_{\mathrm{low}}$ commutes with $\widehat{\Psi}_L$,

$$\begin{aligned} \|\widehat{\Psi}_L T Z\|_F^2 &\gtrsim \|\widehat{\Pi}_{\mathrm{low}} T Z\|_F^2 \\ &\ge \frac{1}{2}\|\widehat{\Pi}_{\mathrm{low}} Z\|_F^2 - \|\widehat{\Pi}_{\mathrm{low}}(T-I)Z\|_F^2 \\ &\ge \frac{1}{2}\|\widehat{\Pi}_{\mathrm{low}} Z\|_F^2 - \delta^2 \|Z\|_F^2. \end{aligned}$$

The two terms satisfy

$$\begin{aligned} \|\widehat{\Pi}_{\mathrm{low}} Z\|_F^2 &= \mathrm{Tr}\left(\widehat{\Pi}_{\mathrm{low}} \widehat{G}^{-1} C^2\right) \\ &\ge \frac{R}{c_0}\,\mathrm{Tr}\left(\widehat{\Pi}_{\mathrm{low}} C^2\right), \end{aligned}$$

and

$$\begin{aligned} \|Z\|_F^2 &= \mathrm{Tr}\left(C\widehat{G}^{-1}C\right) \\ &\le \frac{1}{1-\delta}\,\mathrm{Tr}\left(CG^{-1}C\right). \end{aligned}$$

This proves the first claim.

We now evaluate the two terms on the sketch event. Since $\widehat{G} \preceq (1+\delta)G$, the eigenvalue estimate for G and the fixed-scale decay of $h_{P,j}$ imply, for a sufficiently large absolute constant C_0,

$$\mathrm{rank}\left(\widehat{\Pi}_{\mathrm{low}}\right) \ge M - C_0 k.$$

Ky Fan's minimum principle and the eigenvalue estimate for C therefore give

$$\begin{aligned} \mathrm{Tr}\left(\widehat{\Pi}_{\mathrm{low}} C^2\right) &\ge \sum_{j > C_0 k} \nu_j(C)^2 \\ &\gtrsim k\,(a_k h_{P,k})^2. \end{aligned}$$

For the global source energy, set

$$U := SH^{1/2}, \qquad V := SA_*H^{1/2}, \qquad \Pi_U := U^\top (UU^\top)^{-1} U.$$

Then $G = UU^\top$, $C = VU^\top = UV^\top$, and

$$CG^{-1}C = V\Pi_U V^\top \preceq VV^\top.$$

Consequently, the weighted trace estimate on the joint sketch event gives

$$\mathrm{Tr}\left(CG^{-1}C\right) \le \mathrm{Tr}\left(SA_*^2HS^\top\right) \lesssim 1.$$

Combining the preceding estimates and using $Rh_{P,k} \asymp 1$,

$$\begin{aligned}\|B_*\widehat{\Psi}_L\|_G^2 &\gtrsim Rk\left(a_k h_{P,k}\right)^2 - C\frac{c_{\mathrm{rel}}^2}{R} \\ &\asymp ka_k^2 h_{P,k} - Cc_{\mathrm{rel}}^2 h_{P,k}.\end{aligned}$$

Finally, $ka_k^2 \asymp k^{1-2r} \gtrsim 1$. Choosing c_{rel} sufficiently small absorbs the second term and yields

$$\|B_*\widehat{\Psi}_L\|_G^2 \gtrsim k^{1-\beta_\alpha} + P^{-1}k^{1-\beta_\theta}.$$

□

F Additional Experimental Details

Primitive latent model. All six experiments use a diagonal stable teacher in ambient dimension d:

$$a_j = \rho j^{-r}, \qquad \sigma_{\xi,j} = c_\xi j^{-\alpha}, \qquad \sigma_{0,j} = c_0 j^{-\theta},$$

and

$$x_{i,0,j} \sim \mathcal{N}(0,\sigma_{0,j}), \qquad x_{i,p+1,j} = a_j x_{i,p,j} + \xi_{i,p+1,j}, \qquad \xi_{i,p+1,j} \sim \mathcal{N}(0,\sigma_{\xi,j}).$$

The initialization and innovation covariances are specified independently; in particular, the experiments do not initialize the process at stationarity. For every requested trajectory length P, define

$$m_{P,j} := \frac{1-a_j^{2P}}{P(1-a_j^2)}.$$

The coordinatewise averaged design covariance is computed exactly as

$$h_{P,j} := \frac{1}{P}\sum_{p=0}^{P-1}\mathbb{E}[x_{i,p,j}^2] = \sigma_{0,j}m_{P,j} + \frac{\sigma_{\xi,j}}{1-a_j^2}(1-m_{P,j}).$$

Writing $H_P = \mathrm{diag}(h_{P,1},\dots,h_{P,d})$, we recompute

$$G_P = SH_PS^\top, \qquad C_P = SA_*H_PS^\top, \qquad B_{*,P} = C_PG_P^{-1}$$

for each value of P. A negligible numerical ridge is used only when solving the finite-dimensional positive-definite system.

Approximation and population bias. For a Gaussian sketch $S_{\ell j} \sim \mathcal{N}(0,1/M)$, let

$$Q_P := SH_P^{1/2}, \qquad \Pi_{Q_P} := Q_P^\top(Q_PQ_P^\top)^{-1}Q_P.$$

The approximation experiments evaluate the observable projection error directly:

$$\mathrm{Approx}_{M,P} = \frac{1}{2}\left\|SA_*H_P^{1/2}(I-\Pi_{Q_P})\right\|_F^2.$$

For the population-bias experiments, write

$$G_P = U\,\mathrm{diag}(\mu_1,\dots,\mu_M)U^\top, \qquad \psi_L(s) := \prod_{t=1}^{L}(1-\gamma_t s).$$

The finite-dimensional bias is

$$\mathrm{Bias} = \frac{1}{2}\left\|B_{*,P}U\,\mathrm{diag}(\psi_L(\mu_j))U^\top\right\|_{G_P}^2.$$

The step sizes follow the WSD schedule in Section 2, with linear warmup, a constant-rate stable phase, and exponential decay to γ/L. Specifically,

$$L_{\mathrm{w}} = \max\{1,\lfloor 0.02L\rfloor\}, \qquad L_{\mathrm{s}} = L - \max\{1,\lfloor 0.02L\rfloor\}, \qquad R = \gamma L.$$

Thus warmup and decay each occupy 2% of the steps, with the remaining 96% in the stable phase for all reported step counts.

Trajectory sampling and empirical variance. Each empirical checkpoint samples complete trajectories from the primitive latent recursion and forms

$$\widehat{G} = \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} u_{i,p}u_{i,p}^{\top}, \qquad \widehat{C} = \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=0}^{P-1} y_{i,p}u_{i,p}^{\top}.$$

Before filtering, the base step is safeguarded exactly as in the main analysis:

$$\bar{\gamma} := \min\left\{\gamma, \frac{1}{2\|\widehat{G}\|_{\mathrm{op}}}\right\}.$$

The saved checkpoints have zero clipping rate, so $\bar{\gamma} = \gamma = 0.1$ throughout. If $\widehat{G} = \widehat{U}\,\mathrm{diag}(\widehat{\mu}_j)\widehat{U}^{\top}$, define

$$g_L(s) := \sum_{t=1}^{L}\bar{\gamma}_t\prod_{q=t+1}^{L}(1-\bar{\gamma}_q s), \qquad \widehat{E} := \widehat{C} - B_{*,P}\widehat{G}.$$

The empirical filters are evaluated as

$$\widehat{\mathcal{B}}_L(B_{*,P}) = B_{*,P}\widehat{U}\,\mathrm{diag}(\psi_L(\widehat{\mu}_j))\widehat{U}^{\top},$$

$$\widehat{\mathcal{V}}_L(\widehat{E}) = \widehat{E}\widehat{U}\,\mathrm{diag}(g_L(\widehat{\mu}_j))\widehat{U}^{\top}.$$

Thus every plotted variance point is computed from the same one-sided empirical GD recursion as the theorem. No covariance event or artificial perturbation is imposed.

For Regime II, the dotted reference curves retain the exact finite population spectrum and GD filter while omitting lagged score covariances. If Ω_P is the population residual covariance, the reference value is

$$\mathrm{Var}_L^{\mathrm{ff}} = \frac{\mathrm{Tr}(\Omega_P)}{2NP}\sum_{j=1}^{M}\big(\mu_j g_L(\mu_j)\big)^2.$$

The agreement with directly sampled trajectories therefore checks both the two-scale population spectrum and the finite GD filter.

Regime-I configuration and grids. Regime I uses

$$(\alpha, \theta, r, \rho, c_\xi, c_0, \gamma) = (1.5, 2.5, 0.2, 0.55, 0.05, 0.01, 0.1),$$

so $\theta \geq \alpha$, $\beta_\alpha = 1.9$, and $\beta_\theta = 2.9$. The approximation experiment uses $d = 768$, $P = 4096$, eight sketches, and

$$M \in \{12, 16, 24, 32, 48, 64, 96, 128\}.$$

The bias experiment uses $(M, d, P) = (512, 2048, 4096)$, three sketches, and

$$L \in \{4000, 6000, 8000, 12000, 16000, 24000\}.$$

The variance experiment uses $(M, d) = (20, 128)$, three sketches, and two trajectory repetitions per sketch. Its three sweeps are

$$\begin{aligned}
&N \in \{1024, 1536, 2048, 3072, 4096\}, && P = 4096, \quad L = 64000,\\
&P \in \{1024, 1536, 2048, 3072, 4096\}, && N = 4096, \quad L = 64000,\\
&L \in \{500, 1000, 2000, 3000, 4000, 6000, 8000,\\
&\qquad 16000, 32000, 64000\}, && N = P = 4096.
\end{aligned}$$

The Regime-I horizon sweeps probe finite-dimensional behavior as well as unsaturated scaling; they are not restricted to the strictly unsaturated range of Theorem 3.1. For the variance sweep, the plotted guide retains the covariance scale and the dimension cap, $\min\{M, (c_\xi R)^{1/\alpha}\}$. Its fitted slope over the full grid is 0.530, rather than the unsaturated exponent $1/\alpha = 0.667$.

Regime-II configuration and grids. Regime II uses

$$(\alpha, \theta, r, \rho, c_\xi, c_0, \gamma) = (1.8, 1.25, 0.35, 0.55, 0.03, 0.30, 0.1),$$

so $\alpha - 2r = 1.1 \le \theta < \alpha$, $\beta_\alpha = 2.5$, and $\beta_\theta = 1.95$. The approximation experiment uses $(M, d) = (512, 2048)$, eight sketches, and

$$P \in \{4, 6, 8, 12, 16, 24, 32, 48\}.$$

The bias experiment uses $(M, d, R) = (1024, 2048, 1200)$, five shared sketches, $L = 12000$, and

$$P \in \{4, 6, 8, 12, 16, 24\}.$$

The variance experiment uses $(M, d) = (40, 128)$, eight sketches, and two trajectory repetitions per sketch. At the fixed horizon $R = 2400$ ($L = 24000$), its data sweeps are

$$N \in \{16384, 24576, 32768, 49152, 65536, 98304, 131072\}, \quad P = 40,$$
$$P \in \{8, 12, 16, 24, 32, 40\}, \quad N = 65536.$$

The optimization sweep fixes $(N, P) = (65536, 40)$, uses target horizons

$$R \in \{1400, 2000, 2800, 3800, 5000, 6400, 8000\},$$

and the corresponding step counts

$$L \in \{14000, 20000, 28000, 38000, 50000, 64000, 80000\}.$$

The N- and P-sweeps use nested samples within each repetition, while the L-sweep applies all filters to shared empirical moments. Approximation and bias error bars are standard deviations across sketches; variance error bars are standard errors across sketch-level means.

Table F.1: Log–log slopes from the six saved experiment checkpoints. The guide column uses the asymptotic power laws in Regime I, except that its variance–horizon guide retains the effective-dimension cap, and the plotted finite two-scale or finite-filter curves in Regime II.

Regime	Quantity	Measured	Guide
I	Approx vs. M	-0.962	-0.900
I	Bias vs. R	-0.666	-0.600
I	Var vs. N	-0.972	-1.000
I	Var vs. P	-1.026	-1.000
I	Var vs. R	0.482	0.530
II	Approx vs. P	-0.949	-0.950
II	Bias vs. P	-0.224	-0.185
II	Var vs. N	-0.979	-1.000
II	Var vs. P	-1.359	-1.353
II	Var vs. R	0.524	0.484

Checkpoint agreement. For Regime II, the empirical approximation divided by its two-scale guide has mean 1.000 and range $[0.996, 1.003]$; the bias-to-guide ratio has mean 0.995 and range $[0.943, 1.019]$. The trajectory-sampled variance divided by the finite-filter guide has mean and range 1.029, $[0.959, 1.076]$ for the N-sweep; 0.979, $[0.962, 0.999]$ for the P-sweep; and 1.005, $[0.963, 1.034]$ for the R-sweep. These diagnostics support the theorem's distinction between the one-scale innovation-dominated regime and the two-scale initialization-dominated regime.