FROM SPECTRA TO JOINT SCHEDULES IN LLM PRE-TRAINING: $3 + 3(+2)$ SCALING-LAW REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Power-law learning curves are often treated as fixed properties of a model and its data, although learning-rate and batch-size schedules can change the observed loss. We study this dependence in noisy online SGD with linear random features. Conditional on the representation, an exact Volterra equation separates two response components: a forcing term that propagates unresolved target error and a memory kernel that propagates stochastic-error injections. We prove that either component follows a power law if and only if its cumulative weighted spectral mass has the corresponding low-spectrum scaling; individual eigenvalues and target coefficients need not obey coordinatewise power laws. Under a joint schedule, intrinsic time $T_t = \sum_{s<t} \eta_s$ controls optimization progress, while $r_t = B_t/\eta_t$ controls noise injection. Their interaction yields sharp conditions under which a schedule preserves, changes, or destroys the clean power law, together with a memory ceiling on noise reduction. The power-law random-feature model realizes this mechanism in $3+3(+2)$ propagation regimes with phase-dependent compute rates. Controlled nanoGPT experiments show that (1) learning-rate and batch-size schedules with matched B/η paths are nearly equivalent in intrinsic time, (2) a forcing-memory surrogate accurately predicts loss across schedules, and (3) its fitted exponents across real-world datasets identify the regime of LLMs in $3 + 3(+2)$ map.

1 INTRODUCTION

Power-law learning curves in scaling laws are widely used to forecast training progress and allocate model size, data, and compute in large language models (LLMs) pre-training (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). Such extrapolation is powerful as the fitted law can transfer beyond the scale and training configuration on which it was measured.

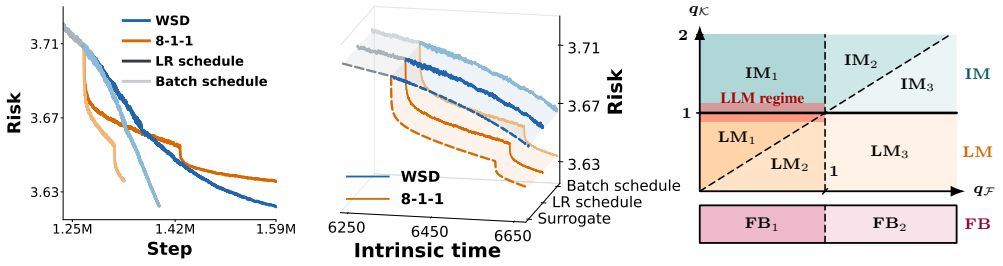
Yet it is often unclear what is expected to transfer: *the existence of a power law*, *its exponent*, or *the mechanism that generates it*. These notions need not coincide. An apparent power law may terminate when a finite range of slow modes is exhausted; and even with the representation and target fixed, changing the learning-rate or batch-size schedule (from LLM pre-training) can change both optimization progress and the accumulation of stochastic error. A scaling exponent is therefore not, in general, a fixed property of a model and its data.

Rather than treating **universality** as a single yes-no question, we specify two distinct problems studied in this paper. The first is one of *origin*: when do the underlying learning dynamics produce power-law response components? The second is one of *transfer*: once such components exist, when does a joint learning-rate/batch-size schedule preserve, change, or destroy their law in the observed loss?

It appears technically impossible to quantitatively study the above two problems in a trained language model. The representation learning, target alignment, stochasticity, and scheduling evolve together, so their separate roles are difficult to identify from the loss curve alone. We therefore study a *proxy* model, linear random features trained by noisy online stochastic gradient descent (SGD) (Rahimi & Recht, 2007; Liu et al., 2022; Mei & Montanari, 2022) to answer the above two questions.

This model is deliberately simple. The value of the proxy lies not in its literal similarity to an LLM, but in whether the response coordinates it reveals remain predictive outside the proxy. We therefore test two consequences in controlled 300M Nano-GPT pre-training (Karpathy, 2022): i) whether the

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.



(a) Loss vs. step (b) Loss vs. intrinsic time (c) $3 + 3(+2)$ map

Figure 1: **Proxy-derived response coordinates in 300M nanoGPT training.** (a) We control the same ratio path $r_t = B_t/\eta_t$ by changing batch-size and learning rate schedules equivalently, e.g., the WSD (Hu et al., 2024) and 8-1-1 (DeepSeek-AI, 2024), leading to different trajectories against optimizer step. (b) After reparameterization by intrinsic time, each trajectory coincides. A surrogate fitted only to the 8-1-1 trajectory also predicts the held-out WSD trajectory without refitting. (c) The fitted exponent satisfies $q_{\mathcal{K}} \approx 1$ and $q_{\mathcal{F}} < 1$, placing the LLM response near the $\mathrm{IM}_1/\mathrm{LM}_1$ boundary of our $3 + 3(+2)$ scaling law regimes. The source code can be found in the supplementary materials.

loss can be fully characterized by the theory-derived coordinates, i.e., the intrinsic time and the ratio between batch-size and learning rate; 2) whether a theory-derived forcing-memory surrogate fitted on one schedule predicts held-out schedules in LLM pre-training. Figure 1 verifies these tests.

1.1 Contributions and Findings

The prediction risk of our proxy model can be precisely characterized by Volterra recursion (Paquette et al., 2024). Its forcing term F propagates unresolved target error without stochastic feedback, whereas its memory kernel K describes how long the effect of one stochastic-error injection survives. The resulting mechanism has two successive stages:

$$\underbrace{\text{weighted low-spectrum mass}}_{\text{spectral origin}} \longrightarrow \underbrace{(F, K)}_{\text{forcing and memory}} \xrightarrow[T_t := \sum_{s<t} \eta_s,\ r_t := B_t/\eta_t]{\text{joint schedule}} \underbrace{R_{\sigma,t}}_{\text{observed loss}},$$

where (T_t, r_t) are the two natural schedule coordinates. Intrinsic time T_t measures optimization progress, while $r_t := B_t/\eta_t$ under the batch size B_t and the learning rate η_t controls stochastic-error injection per unit progress. The weighted low spectrum determines whether F and K follow power laws; the joint schedule then accumulates the memory response along the training trajectory. Consequently, componentwise power laws do not by themselves determine the exponent of the total loss. The observed curve is governed by the competition between clean target-error decay and the accumulated stochastic response. Our contributions are summarized as below.

- **Spectral origin of component power laws.** We prove separate if-and-only-if criteria for power laws in learnable forcing and one-injection memory. The decisive quantities are their cumulative weighted spectral masses near zero, rather than coordinatewise power laws for individual eigenvalues or target coefficients. Irregular spectra or targets can therefore produce canonical power-law responses, while power-law data are neither sufficient nor necessary.
- **Transformation under joint learning-rate-batch-size schedules.** We show that the noisy–clean gap is obtained by accumulating schedule-dependent stochastic injections through the memory kernel. This gives sharp boundaries between schedules that preserve, change, or destroy the clean power law, together with a memory ceiling beyond which reducing only late-stage noise cannot improve the decay rate. Below this ceiling, the gap identifies the asymptotic behavior of the ratio $r_t = B_t/\eta_t$, though not learning rate and batch size separately.
- **Concrete $3 + 3(+2)$ scaling law regimes.** We consider specific power-law eigenvalues and target coefficients, formulating a 3+3(+2)-regime map: 3 long-memory (LM), 3 integrable-memory (IM), and 2 finite-bulk propagation regimes. We derive an exact conditional formula for the finite-width noisy–clean gap in the finite-bulk regimes, asymptotic formulas for the full deterministic-equivalent risk across the phase map, and phase-dependent optimal rates under data and feature-compute budgets.

- **Proxy for LLM pre-training prediction.** Controlled nanoGPT (124M and 300M) experiments provide three tests of the theory: the loss is determined by the intrinsic time and the ratio path $r_t := B_t/\eta_t$, which provides a possible way to tune learning-rate or batch-size schedules under a fixed ratio r_t. Besides, a forcing-memory surrogate fitted on one schedule (e.g., 8-1-1) predicts held-out schedules without refitting (e.g., WSD). Across OpenWebText and FineWeb and across model scales, LLM response remains near $q_{\mathcal{K}} \approx 1$ and $q_{\mathcal{F}} < 1$.

1.2 Related work

Power-law existence and exponent universality. Recent work argues for shared exponents or derives a universal $1/3$ time law (Liu & Gore, 2026; Liu et al., 2026). Other studies show that scaling laws can change with the observation scale, terminate when a finite spectral range is exhausted, or vary across spectral phases (Xiao, 2024; Maloney et al., 2022; Bahri et al., 2024; Lin et al., 2024; Paquette et al., 2024). These results motivate separating the existence of a component power law from the invariance of the exponent observed in the total loss.

Spectral models and random-feature dynamics. Solvable spectral theories commonly prescribe power laws for eigenvalues or target coefficients and then derive the resulting learning curve (Maloney et al., 2022; Bordelon et al., 2024; Lin et al., 2024; Paquette et al., 2024). Our componentwise criteria instead characterize power laws through cumulative weighted spectral mass and therefore do not require coordinatewise power-law sequences. Random-feature models provide a setting in which spectral, target, finite-width, and stochastic effects can be separated (Mei & Montanari, 2022; Xiao et al., 2022; Bach, 2024; Ba et al., 2022; Moniri et al., 2024), while high-dimensional SGD limits provide the dynamical foundation for the exact recursion used here (Paquette et al., 2021; 2025; Atanasov et al., 2026). The closest PLRF result is the constant-schedule, clean-label $4 + 3$ compute-optimal classification of Paquette et al. (2024). We introduce label noise and joint schedules; our $3 + 3(+2)$ map classifies forcing–memory propagation rather than compute-optimal bottlenecks.

Schedules and transferable response laws. Schedule optimization has been studied from complementary starting points. Bordelon & Mori (2026) assume spectral and source power laws, whereas functional-scaling-law analyses begin from a specified response law (Li et al., 2025; 2026a; Wang et al., 2026). In particular, Li et al. (2025) fit one learning-rate schedule at fixed batch size and predict other schedules without refitting. Our theory addresses the preceding layer, when the underlying forcing and memory laws exist, and extends the response analysis to jointly varying learning rate and batch size. Our LLM experiments correspondingly test both changes in the ratio path B_t/η_t and different learning-rate–batch-size factorizations of the same path (Meterez et al., 2026b; Karkada et al., 2026; Hu & Lu, 2023).

From a proxy mechanism to LLM response. A tractable proxy becomes useful beyond its literal assumptions only when the coordinates it reveals make falsifiable predictions outside the proxy. Previous results on local quadratic models (Meterez et al., 2026b), kernel-based functional scaling laws (Li et al., 2025), and Gaussian universality (Karkada et al., 2026; Hu & Lu, 2023) demonstrate their potential in prediction of large complex models. Such mechanism-first viewpoint motivates us to test it directly in controlled plain-SGD LLM training, supported by the following two reasons.

First, whether a proxy model or an LLM is trained with plain SGD, the dynamics can always be cast as signal–noise learning related to forcing and memory terms. Second, paired learning-rate and batch-size factorizations with matched B_t/η_t paths follow different trajectories in optimizer step but nearly coincide at equal intrinsic time.

2 Problem setup and exact loss dynamics

The model follows the linear random-feature setting of Paquette et al. (2024), trained by online SGD, but differs in two ways motivated by LLM-pretraining practice (Brown et al., 2020; Hoffmann et al., 2022; Li et al., 2025; Wang et al., 2026): it introduces label noise as a tractable source of persistent stochastic error; it allows the learning rate and batch size to vary jointly over time.

2.1 Linear random features trained by online SGD

We consider Gaussian data $\boldsymbol{x} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{z} \in \mathbb{R}^d$, with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \succ \mathbf{0}$. It can be naturally extended to kernel feature maps under the hypercontractivity condition of Mei

et al. (2022). The data generation process is

$$y = f_\star(\boldsymbol{x}) + \varepsilon, \qquad f_\star(\boldsymbol{x}) := \langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle, \qquad \varepsilon \perp\!\!\!\perp \boldsymbol{x}, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{E}[\varepsilon^2] = \sigma^2 ,$$

where $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is the target and the label noise ε is with zero-mean and bounded variance σ^2. We use a linear random feature model of width m, $f_{\boldsymbol{a}}(\boldsymbol{x}) := \langle \boldsymbol{W}^\top \boldsymbol{x}, \boldsymbol{a} \rangle$, where $\boldsymbol{W} \in \mathbb{R}^{d \times m}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries and is fixed; only $\boldsymbol{a} \in \mathbb{R}^m$ is trained by SGD. Starting from $\boldsymbol{a}_0 = \boldsymbol{0}$, iteration t draws a fresh mini-batch of size B_t and updates with learning rate η_t:

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t - \frac{\eta_t}{B_t} \sum_{i=1}^{B_t} \boldsymbol{W}^\top \boldsymbol{x}_t^i \big(f_{\boldsymbol{a}_t}(\boldsymbol{x}_t^i) - y_t^i \big), \qquad T_t := \sum_{s<t} \eta_s, \qquad r_t := \frac{B_t}{\eta_t}. \tag{2.1}$$

Batches and label noises are independent across iterations. Here T_t is the optimization clock, while $1/r_t$ is the variance injected per unit intrinsic time. Conditional on $\boldsymbol{W}$, the excess risk is defined as

$$R_{\sigma,t} := \mathbb{E}\Big[\big(f_{\boldsymbol{a}_t}(\boldsymbol{x}) - f_\star(\boldsymbol{x}) \big)^2 \,\Big|\, \boldsymbol{W} \Big] = \mathbb{E}\Big[\|\boldsymbol{\Lambda}^{1/2} (\boldsymbol{W} \boldsymbol{a}_t - \boldsymbol{\theta}^\star)\|_2^2 \,\Big|\, \boldsymbol{W} \Big] .$$

The expectation averages all training mini-batches and label noises and an independent test covariate, conditioning on $\boldsymbol{W}$. All definitions and exact recursions in this section hold for finite d and m. The asymptotic results later take $m \to \infty$ and $d = d_m \to \infty$ jointly, with d/m bounded below by a constant strictly larger than one; see more details in Appendix A.

2.2 Forcing, Memory, and the Exact Volterra Equation

To identify the spectral quantities first, temporarily fix the schedule: $\eta_t \equiv \eta$, $B_t \equiv B$, and $T := \eta t$. The frozen representation induces $\widehat{\boldsymbol{H}} := \boldsymbol{\Lambda}^{1/2} \boldsymbol{W} \boldsymbol{W}^\top \boldsymbol{\Lambda}^{1/2}$ with its eigenpairs $(\widehat{\lambda}_j, \widehat{\boldsymbol{u}}_j)$. The initial energy aligned with the empirical mode j is $|\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \rangle|^2$. For Gaussian covariates, the corresponding second-moment mode has one-step survival factor

$$q_\eta(\lambda) := 1 - 2\eta\lambda + (1 + 1/B)\, \eta^2 \lambda^2. \tag{2.2}$$

Its power $q_\eta(\lambda)^t$ is the spectral filter that determines how much squared error in mode λ survives t steps: on the long-time scale $\lambda = O(T^{-1})$, $q_\eta(\lambda)^t \approx e^{-2\eta\lambda t} = e^{-2T\lambda}$. Summing the exact mode recursions gives the exact risk recursion

$$R_{\sigma,t} = F_{\boldsymbol{W}}(t) + \sum_{s=0}^{t-1} K_{\boldsymbol{W}}(t-1-s)(R_{\sigma,s} + \sigma^2) . \tag{2.3}$$

Here the forcing term $F_{\boldsymbol{W}}$ and the memory kernel $K_{\boldsymbol{W}}$ are defined as

$$F_{\boldsymbol{W}}(t) := \sum_j |\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \rangle|^2 q_\eta(\widehat{\lambda}_j)^t, \qquad K_{\boldsymbol{W}}(t) := \frac{\eta^2}{B} \sum_j \widehat{\lambda}_j^2 q_\eta(\widehat{\lambda}_j)^t ,$$

where $F_{\boldsymbol{W}}$ is the initial target error propagated without stochastic feedback. The memory kernel $K_{\boldsymbol{W}}$ is the impulse response of one SGD-variance injection. When $\sigma = 0$, Eq. (2.3) reduces to the clean-label recursion of Paquette et al. (2024). The zero-mode forcing is the finite-width approximation floor $R_{\mathrm{app}} := \sum_{\widehat{\lambda}_j = 0} |\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \rangle|^2$; write $F_{\boldsymbol{W},>0}$ for the learnable part of the forcing.

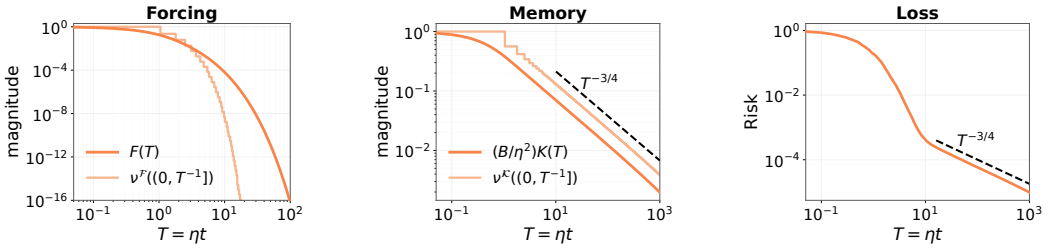
3 When Do Frozen Spectral Dynamics Produce Power Laws?

Every fixed positive mode in Eq. (2.2) decays exponentially. A power law can nevertheless emerge after infinitely many modes with different time scales are added. This section makes that statement precise for F and K; it does not yet claim a power law for the total loss.

3.1 From the Moving Spectral Cutoff to a Spectral Criterion

On the long-time scale $T = \eta t$, modes with $\lambda \gg T^{-1}$ have relaxed, while modes with $\lambda \ll T^{-1}$ are nearly untouched: $q_\eta(\lambda)^t \approx e^{-2T\lambda} \approx \mathbf{1}\{\lambda \lesssim T^{-1}\}$. Hence the remaining error is governed by cumulative spectral weight below the moving cutoff T^{-1}, not by the decay of a single mode (Paquette et al., 2024; Li et al., 2025). Accordingly, for the eigenpairs of $\widehat{\boldsymbol{H}}$, define the target-weighted and memory-weighted empirical spectral measures

$$\nu_{\boldsymbol{W}}^{\mathcal{F}} := \sum_j \Big| \Big\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \Big\rangle \Big|^2 \delta_{\widehat{\lambda}_j}, \qquad \nu_{\boldsymbol{W}}^{\mathcal{K}} := \sum_j \widehat{\lambda}_j^2 \delta_{\widehat{\lambda}_j}.$$



(a) Forcing (b) Memory (c) Loss

Figure 2: Finite-width rapid-target construction with $\lambda_j = j^{-0.8}$, $|\theta_j^\star|^2 = e^{-j}$, and $d = 524{,}288$. (a) Normalized forcing and cumulative forcing mass both decay faster than every inverse power of T. (b) Normalized one-injection memory and cumulative memory mass both track $T^{-3/4}$. (c) The minibatch-SGD loss curve averaged over 20 runs follows the $T^{-3/4}$ memory tail.

Here $\nu_{\boldsymbol{W}}^{\mathcal{F}}((0,x])$ is the target energy in learnable directions slower than x^{-1}, whereas $\nu_{\boldsymbol{W}}^{\mathcal{K}}((0,x])$ measures the strength of one variance injection stored in those directions. The zero atom $\nu_{\boldsymbol{W}}^{\mathcal{F}}(\{0\}) = R_{\mathrm{app}}$ is the approximation floor and is excluded from $F_{\boldsymbol{W},>0}$.

These measures give the exact spectral representations and their long-time Laplace approximations:

$$F_{\boldsymbol{W},>0}(t) = \int_{(0,\infty)} q_\eta(\lambda)^t \, \nu_{\boldsymbol{W}}^{\mathcal{F}}(\mathrm{d}\lambda) \approx \int_{(0,\infty)} e^{-2T\lambda} \, \nu_{\boldsymbol{W}}^{\mathcal{F}}(\mathrm{d}\lambda) = 2T \int_0^\infty e^{-2Tx} \nu_{\boldsymbol{W}}^{\mathcal{F}}((0,x]) \, \mathrm{d}x,$$

$$\frac{B}{\eta^2} K_{\boldsymbol{W}}(t) = \int_{(0,\infty)} q_\eta(\lambda)^t \, \nu_{\boldsymbol{W}}^{\mathcal{K}}(\mathrm{d}\lambda) \approx \int_{(0,\infty)} e^{-2T\lambda} \, \nu_{\boldsymbol{W}}^{\mathcal{K}}(\mathrm{d}\lambda) = 2T \int_0^\infty e^{-2Tx} \nu_{\boldsymbol{W}}^{\mathcal{K}}((0,x]) \, \mathrm{d}x.$$

Accordingly, the following theorem characterizes exactly when either component follows a temporal power law. The fully quantified statement is given in Theorem A.4.

Theorem 3.1 (Componentwise spectral criterion (informal))**.** Fix a stable constant schedule, condition on $\boldsymbol{W}$ at each width, and take a joint width–time limit with $m, t \to \infty$ and $T = \eta t \to \infty$. Under appropriate uniform spectral-window conditions, with $x = T^{-1} \downarrow 0$, for $q_{\mathcal{F}}, q_{\mathcal{K}} > 0$,

$$\nu_{\boldsymbol{W}}^{\mathcal{F}}((0,x]) \propto x^{q_{\mathcal{F}}} \iff F_{\boldsymbol{W},>0}(t) \propto T^{-q_{\mathcal{F}}}, \qquad \nu_{\boldsymbol{W}}^{\mathcal{K}}((0,x]) \propto x^{q_{\mathcal{K}}} \iff \frac{B}{\eta^2} K_{\boldsymbol{W}}(t) \propto T^{-q_{\mathcal{K}}}.$$

Theorem 3.1 shows that the decisive quantities are the cumulative weighted spectral masses near zero, rather than pointwise power-law formulas for the eigenvalues or target coefficients. It can be empirically validated by Figure 2. In Figures 2(a) and 2(b), each temporal response is compared with its corresponding cumulative spectral mass. The forcing response and forcing mass both decay faster than every inverse power of T, whereas the memory response and memory mass both scale as $T^{-3/4}$, illustrating the two componentwise equivalences in Theorem 3.1. Separately, Figure 2(c) shows that the minibatch-SGD loss initially follows the exponentially decaying forcing response and then crosses over to the $T^{-3/4}$ power-law memory tail.

3.2 When power laws appear—and when they do not

Based on Theorem 3.1, we provide several examples to see whether power laws appear or not. For instance, power-law data are neither sufficient nor necessary. See more details in Appendix D.2.

- **Only power-law data are insufficient.** Zipf-like rank–frequency laws make power-law structure familiar in language data (He et al., 2025; Mikhaylovskiy, 2025). Such data-side scaling does not determine target alignment in the spectral dynamics studied here. For $\alpha > 1/4$, take $\lambda_j = j^{-2\alpha}$ and $|\theta_j^\star|^2 = e^{-j}$. The target places exponentially little energy in slow directions. For the corresponding deterministic construction, $-\log F_{\boldsymbol{W},>0}(t) \asymp T^{1/(2\alpha+1)}$. The forcing therefore decays faster than every inverse power of T, although the eigenspectrum is an exact power law. The memory decay is unchanged because it does not depend on the target. Figure 2 visualizes this deterministic construction.
- **Power-law data are not necessary either.** For example, the constructions $\lambda_j = e^{-j^\gamma}$, $|\theta_j^\star| \asymp j^{-\beta}$, $0 < \gamma < 1$, and $\lambda_j = e^{-j}$, $\lambda_j |\theta_j^\star|^2 \propto e^{-\tau j}$, both produce power-law temporal orders on their corresponding scaling windows.

- **Irregular targets can give the canonical exponent.** For $\lambda_j = j^{-2\alpha}$, with $\alpha > 1/4$ and $2\alpha + 2\beta > 1$, write $|\theta_j^\star|^2 = j^{-2\beta} c_j$, where $c_j \geq 0$. If $n^{-1}\sum_{j\leq n} c_j \to 1$, then $\sum_{\lambda_j \leq x} \lambda_j |\theta_j^\star|^2 \propto x^{(2\alpha+2\beta-1)/(2\alpha)}$. Thus oscillatory, mass-compensated sparse, and frozen random-amplitude targets can retain the canonical forcing exponent.
- **A finite spectrum leads to exponential decay.** It may display a long intermediate scaling window before entering this asymptotic regime.

Non-identifiability: From forcing and memory to the observed loss. Theorem 3.1 is componentwise: its if-and-only-if conclusions apply separately to $F_{\boldsymbol{W},>0}$ and $K_{\boldsymbol{W}}$, not to the observed loss. $F_{\boldsymbol{W},>0}$ gives the learnable forcing, whereas $K_{\boldsymbol{W}}$ describes the survival of one stochastic injection. The observed loss is therefore not obtained by directly comparing these two raw components: the memory contribution must first be accumulated over past injections and propagated through feedback.

4 PLRF PROPAGATION REGIMES: $3 + 3(+2)$ SCALING LAWS

We now instantiate the general spectral criterion of Section 3 in the canonical power-law random-feature (PLRF) model of Paquette et al. (2024). The resulting $3 + 3(+2)$ phase map classifies how learnable forcing and one-injection memory propagate, rather than assigning eight universal exponents to the observed loss. The schedule must still accumulate memory over past noise injections; Section 5 carries out this step and determines the resulting learning curve.

Consider the following power-law setting in PLRF

$$\lambda_j = j^{-2\alpha}, \qquad \theta_j^\star = j^{-\beta}, \qquad 2\alpha + 2\beta > 1 \text{ (ensuring } \sum_j \lambda_j |\theta_j^\star|^2 < \infty),$$

the moving cutoff $\lambda_{j_T} \asymp T^{-1}$ corresponds to $j_T \asymp T^{1/(2\alpha)}$. Summing the unresolved tail gives

$$F(T) \approx \sum_{j \gtrsim j_T} j^{-2(\alpha+\beta)} \propto T^{-q_{\mathcal{F}}}, \quad q_{\mathcal{F}} := \frac{2(\alpha+\beta)-1}{2\alpha}, \qquad \frac{B}{\eta^2} K(T) \approx \sum_{j \gtrsim j_T} j^{-4\alpha} \propto T^{-q_{\mathcal{K}}}, \quad q_{\mathcal{K}} := 2 - \frac{1}{2\alpha}.$$

This cutoff calculation gives the PLRF response exponents; Corollary A.10 gives the corresponding rigorous population-filter asymptotics.

Why use $(q_{\mathcal{F}}, q_{\mathcal{K}})$ instead of (α, β): For PLRF, $q_{\mathcal{F}}$ and $q_{\mathcal{K}}$ are algebraic functions of the microscopic source and capacity parameters (α, β), i.e., a change of coordinates. The purpose of the response coordinates is not to create new regimes by relabeling the PLRF phase diagram. Rather, $q_{\mathcal{F}}$ and $q_{\mathcal{K}}$ are respectively the decay exponents of forcing and one-injection memory, and are the only microscopic information entering the subsequent schedule-transfer law. They remain well defined for irregular spectra and targets for which no coordinatewise (α, β) power laws exist, and models with the same response exponents have the same leading schedule behavior. We therefore use (α, β) to describe the PLRF origin of a regime and $(q_{\mathcal{F}}, q_{\mathcal{K}})$ to describe its dynamical propagation. The finite-bulk regimes are kept separate because their memory remains explicitly coupled to width and cannot be represented by a width-independent $q_{\mathcal{K}} > 0$, shown in Figure 3.

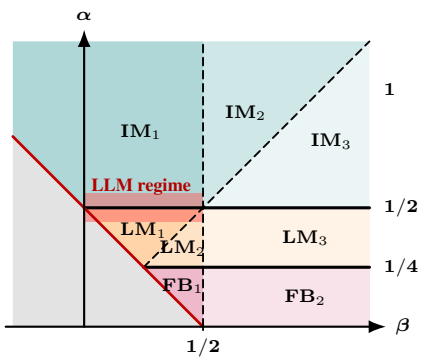


Figure 3: **PLRF regimes in (α, β).** Colors match the response-coordinate map in Figure 1; the red band marks the fitted LLM regime near $\alpha = 1/2$.

Long memory (LM) has $0 < q_{\mathcal{K}} < 1$, so its survival kernel is not integrable; **integrable memory (IM)** has $q_{\mathcal{K}} > 1$. Within the LM and IM families, the subscripts record the position of $q_{\mathcal{F}}$ relative to $q_{\mathcal{K}}$ and 1.

For $0 < \alpha < 1/4$, the squared spectral mass $\sum_{j\leq m} \lambda_j^2 \asymp m^{1-4\alpha}$ grows with width, so no width-independent decaying exponent $q_{\mathcal{K}} > 0$ exists. This is the **finite-bulk (FB)** regime. The right panel of Figure 1 organizes the propagation regimes in the response coordinates $(q_{\mathcal{F}}, q_{\mathcal{K}})$ that govern schedule behavior. Figure 3 pulls the same LM/IM partition back to the PLRF parameters (α, β), with finite

bulk shown separately because no width-independent $q_{\mathcal{K}}$ exists there, making direct comparison with the $4+3$ taxonomy possible.

Relative to the compute-optimal $4+3$ taxonomy of Paquette et al. (2024), the propagation viewpoint makes two structural changes: Phase Ib splits across the finite-bulk and forcing–memory boundaries, while that taxonomy's IVa–IVb crossover $\alpha = 1 - 1/\sqrt{2}$ lies inside LM_3, rather than on a propagation boundary. We use the propagation taxonomy throughout.

The fitted $q_{\mathcal{K}} \approx 1$ corresponds to $\alpha \approx 1/2$, where the $4+3$ theory predicts the **near-square-root** compute-optimal width scaling reported by Chinchilla (Paquette et al., 2024; Hoffmann et al., 2022); see Appendix F.

Phasewise forcing. The phase map becomes concrete through the deterministic-equivalent forcing. In the proved open source range $\beta < 1 + 2\alpha$, and on the window $1 \ll T \le C_0 m^{2\alpha}$,

$$\mathcal{F}(T,m) \asymp \begin{cases} m^{-2\alpha q_{\mathcal{F}}} + T^{-q_{\mathcal{F}}}, & \mathrm{FB}_1, \mathrm{LM}_1, \mathrm{LM}_2, \mathrm{IM}_1, \\ m^{-2\alpha} + T^{-q_{\mathcal{F}}}, & \mathrm{FB}_2, \mathrm{LM}_3, \\ m^{-2\alpha} + T^{-q_{\mathcal{F}}} + m^{-1} T^{-(q_{\mathcal{K}}-1)}, & \mathrm{IM}_2, \mathrm{IM}_3. \end{cases} \tag{4.1}$$

Here $\mathcal{F}(T,m)$ denotes the deterministic-equivalent approximation to $F_{\boldsymbol{W}}(t)$ at intrinsic time $T = \eta t$. Every row contains a finite-width floor and a target-aligned source transient; only IM_2 and IM_3 contain the additional feature-distortion transient. The largest term determines the forcing on the stated width–time window.

Eq. (4.1) describes the forcing component, not the observed total risk. This is why the $3+3(+2)$ map differs from the $4+3$ compute-optimal taxonomy of Paquette et al. (2024): it classifies how forcing and memory propagate before resource optimization. Section 5 supplies the missing step by accumulating one-injection memory along the schedule and comparing the resulting noise response with clean learning.

5 How joint schedules transform the observed scaling law

The exponent $q_{\mathcal{K}}$ characterizes how the contribution of one stochastic-error injection decays with its age. A training schedule creates a stream of such injections. At intrinsic time u, the ratio $r(u) = B(u)/\eta(u)$ sets how much noise is injected, while the memory profile $k(T-u)$ sets how much survives until time T. The observed loss is determined by their accumulation and its competition with clean learning.

5.1 Schedule response and preserve–change–destroy classification

Recall $T_t = \sum_{s<t} \eta_s$, and interpolate $r(u) = r_s = B_s/\eta_s$ for $T_s \le u < T_{s+1}$. For a constant schedule, define the reference survival profile $k(\eta t) := (B/\eta^2) K_{\boldsymbol{W}}(t)$. Under a varying schedule, the constant-schedule age-only kernel $K_{\boldsymbol{W}}(t-s)$ becomes the two-time kernel $K_{t,s}$, the exact weight with which stochastic error injected at step s survives until step t. Subtracting the clean recursion from the noisy one gives the first relation below. For a fixed infinite-spectrum dynamics with $\sigma^2 > 0$, we assume the continuum approximation and uniformly subcritical feedback in the remaining two:

$$R_{\sigma,t} - R_{0,t} = \sum_{s<t} K_{t,s} \left[\sigma^2 + R_{\sigma,s} - R_{0,s}\right], \qquad \sum_{s<t} K_{t,s} \asymp \int_0^{T_t} \frac{k(T_t - u)}{r(u)}\, \mathrm{d}u, \qquad \sup_t \sum_{s<t} K_{t,s} < 1\,.$$

Along the intrinsic-time grid, write $T = T_t$, $R_\sigma(T) := R_{\sigma,t}$, and $R_0(T) := R_{0,t}$. The following theorem converts power-law k and r into the noise exponent $q_{\mathcal{N}}$, whose comparison with the clean exponent q_0 determines whether the clean loss law is preserved, changed, or destroyed.

Theorem 5.1 (Schedule transformation of loss)**.** Suppose $k(T) \sim c_{\mathcal{K}} T^{-q_{\mathcal{K}}}$, with $c_{\mathcal{K}} > 0$ and $q_{\mathcal{K}} \neq 1$, and let $r(T) \sim c_r T^{\vartheta}$ be eventually monotone, with $c_r > 0$. On every decaying branch,

$$R_\sigma(T) - R_0(T) \asymp T^{-q_{\mathcal{N}}(\vartheta)}, \qquad q_{\mathcal{N}}(\vartheta) := \begin{cases} \vartheta + q_{\mathcal{K}} - 1, & \mathrm{LM}, \quad 1 - q_{\mathcal{K}} < \vartheta < 1, \\ q_{\mathcal{K}}, & \mathrm{LM}, \quad \vartheta > 1, \\ \min\{\vartheta, q_{\mathcal{K}}\}, & \mathrm{IM}, \quad \vartheta > 0. \end{cases}$$

At $\vartheta = 1$ in LM, the gap is $\asymp T^{-q_{\mathcal{K}}} \log T$. If, in addition, $R_0(T) - R_{\mathrm{app}} \sim c_0 T^{-q_0}$, then

$$R_\sigma(T) - R_{\mathrm{app}} \asymp T^{-\min\{q_0, q_{\mathcal{N}}(\vartheta)\}}, \quad q_0 > 0\,,$$

up to the displayed LM logarithmic correction.

The theorem yields a phase classification by comparing the noise exponent $q_{\mathcal{N}}(\vartheta)$ with the clean exponent q_0. Figure 4 summarizes this comparison. The case $q_{\mathcal{K}} = 1$ is excluded because it is the marginal LM/IM boundary: cumulative memory grows as $\int_0^T k(v)\,\mathrm{d}v \asymp \log T$, so the gap acquires logarithmic corrections rather than following the pure-power formulas above.

Consequently, the leading clean-loss decay is *preserved*, *critical*, or *changed* according as $q_{\mathcal{N}}(\vartheta)$ is above, equal to, or below q_0. At $\vartheta = 1 - q_{\mathcal{K}}$ in LM and $\vartheta = 0$ in IM, the gap has no positive-power decay and therefore *destroys* convergence to R_{app}; below these edges, uniform row stability is impossible. Since $q_{\mathcal{N}}(\vartheta) \le q_{\mathcal{K}}$, strict preservation is possible only when $q_0 < q_{\mathcal{K}}$. The complete boundary asymptotics and tunable branches are given in Appendix D.

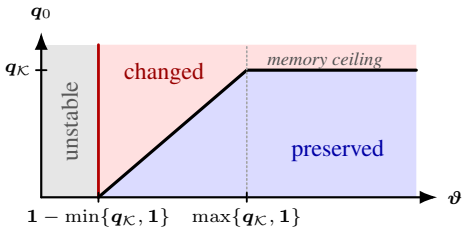


Figure 4: **LM/IM schedule response.** The red line marks the destroy boundary, where the noisy–clean gap has no positive-power decay; the gray region to its left is unstable. The black curve separates changed from preserved scaling and plateaus at the memory ceiling.

The LLM experiment in Appendix F.1 directly tests the schedule family in Theorem 5.1: it varies ϑ in $r(T) \propto T^{\vartheta}$ at matched terminal intrinsic time, provides evidence for the predicted memory ceiling through the diminishing response at large ϑ.

What the loss can and cannot identify. Below the LM memory ceiling, a noisy–clean gap asymptotic $R_\sigma(T) - R_0(T) \sim cT^{-q}$, $0 < q < q_{\mathcal{K}}$, identifies the ratio exponent $\vartheta = 1 - q_{\mathcal{K}} + q$. It identifies $r = B/\eta$, not learning rate and batch size separately; details are given in Appendix D.

5.2 PLRF FINITE-BULK AND FULL-RISK ASYMPTOTICS

Theorem 5.1 assumes a width-independent survival profile. For $0 < \alpha < 1/4$, an unresolved spectral band instead carries squared mass $m^{1-4\alpha}$. On the finite-bulk window,

$$R_{\sigma,t} - R_{0,t} \asymp \sigma^2 m^{1-4\alpha} \sum_{s<t} \frac{\eta_s^2}{B_s} = \sigma^2 m^{1-4\alpha} \int_0^{T_t} \frac{\mathrm{d}u}{r(u)}. \tag{5.1}$$

This is why FB is not a third branch of the LM/IM map: memory and width remain coupled.

Let $\mathcal{F}(T, m)$ be the forcing profile in Eq. (4.1), and write $\Delta T_s := T_{s+1} - T_s$. The PLRF deterministic-equivalent risk admits the following forcing–memory representation:

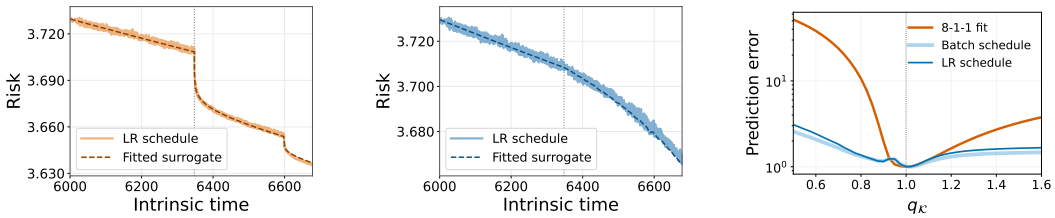
$$\mathcal{R}_{\sigma,t} \asymp \begin{cases} \mathcal{F}(T_t, m) + \displaystyle\sum_{s<t} \frac{\Delta T_s}{r_s} (1 + T_t - T_{s+1})^{-q_{\mathcal{K}}} [\mathcal{F}(T_s, m) + \sigma^2], & \alpha > \frac{1}{4}, \\ \mathcal{F}(T_t, m) + \sigma^2 m^{1-4\alpha} \displaystyle\sum_{s<t} \frac{\Delta T_s}{r_s}, & 0 < \alpha < \frac{1}{4}. \end{cases} \tag{5.2}$$

The first term is the unresolved signal, while the second accumulates surviving stochastic injections. In the LM/IM response, $\mathcal{F}(T_s, m)$ records mini-batch variance even with clean labels, whereas σ^2 is the additional label-noise variance. Precise source windows and empirical-to-DE transfer conditions are given in Appendix C.

Forcing-memory surrogate for LLM. We use the $\alpha > 1/4$ branch because only in this regime does the width-independent profile $k(v) \asymp (1+v)^{-q_{\mathcal{K}}}$ exist, with $q_{\mathcal{K}} = 2 - 1/(2\alpha) > 0$. The $\alpha > 1/4$ branch of Eq. (5.2) suggests a 7 parameter surrogate for LLM loss:

$$\widehat{L}(T) = L_\infty + A_{\mathcal{F}}(1+T)^{-q_{\mathcal{F}}} + \int_0^T \frac{A_0 + A_1(1+u)^{-q_{\mathcal{F}}}}{r(u)} \big(1 + c_{\mathcal{K}}(T-u)\big)^{-q_{\mathcal{K}}}\,\mathrm{d}u\,. \tag{5.3}$$

Its seven parameters $(L_\infty, A_{\mathcal{F}}, A_0, A_1, c_{\mathcal{K}}, q_{\mathcal{F}}, q_{\mathcal{K}})$ can be fitted to an observed loss curve, inspired by Li et al. (2025). Section 5.4 and Appendix F fit these parameters on an 8-1-1 schedule and evaluate the resulting surrogate on a WSD schedule without refitting.



(a) 8-1-1 fit (b) WSD zero-refit transfer (c) Prediction error

Figure 5: **LLM surrogate fit, zero-refit transfer, and effective $q_{\mathcal{K}}$.** (a) For a 300M plain-SGD LLM, we fix $q_{\mathcal{K}} = 1$ in Eq. (5.3) and fit the remaining 6 parameters to the 8-1-1 learning-rate schedule. (b) With all fitted parameters frozen, it predicts the WSD learning-rate schedule without refitting. (c) Fixing $q_{\mathcal{K}}$, fitting the remaining parameters only on 8-1-1, and evaluating the WSD schedules without refitting yields prediction errors, each divided by its best value, that are minimized near $q_{\mathcal{K}} \approx 1$.

5.3 Optimal ratios and sharp resource rates

The preceding forcing–memory representation also turns schedule design into an optimization over the ratio path $r(u)$. In the LM/IM branches, for fixed T and data budget D, this means minimizing $\int_0^T (1+T-u)^{-q_{\mathcal{K}}}[\mathcal{F}(u,m)+\sigma^2]/r(u)\,\mathrm{d}u$, subject to $\int_0^T r(u)\,\mathrm{d}u = D$. Cauchy–Schwarz gives

$$r^\star(u) \propto \sqrt{(1+T-u)^{-q_{\mathcal{K}}}[\mathcal{F}(u,m)+\sigma^2]}, \qquad \int_0^T r^\star(u)\,\mathrm{d}u = D.$$

Hence more samples are allocated where an injection is both large and likely to survive until time T. This recovers the joint optimum of Bordelon & Mori (2026) and its fixed-batch power-decay and WSD factorizations (Bordelon & Mori, 2026; Li et al., 2026a).

A representative resource balance. In most LM/IM branches, the leading terms reduce to $\mathcal{R}_\sigma(T) \asymp T^{-p/(2\alpha)} + D^{-1}T^{1/(2\alpha)}$, with $p = 2\alpha + 2\beta - 1$. Balancing signal and controlled noise gives $T^\star \asymp D^{2\alpha/(1+p)}$, $\mathcal{R}_\sigma^\star(D) \asymp D^{-p/(1+p)}$. Width floors, feature distortion, finite-bulk amplification, and the constraint $\mathfrak{f} \asymp mD$ produce the remaining phase-dependent compute rates.

Across the PLRF regimes, optimizing over admissible schedules yields the optimal risk orders under data and feature-compute constraints in Corollary C.23 and Table 4. The upper bounds are attainable and match the corresponding lower bounds.

5.4 LLM experiments

Figure 1 previews the LLM results. We fit the 7 parameters of the surrogate in Eq. (5.3) only to the 8-1-1 trajectory and use the frozen fit to predict WSD without refitting. Figures 5(a) and 5(b) show the fit to 8-1-1 and the zero-refit prediction of WSD, respectively. The prediction errors in Figure 5(c) are minimized near $q_{\mathcal{K}} \approx 1$, placing the effective LLM response near the LM/IM boundary previewed in Figure 1(c). Independent fits on OpenWebText (Gokaslan & Cohen, 2019), FineWeb (Penedo et al., 2024), and peS2o V2 (Soldaini & Lo, 2023) yield $q_{\mathcal{K}} = 1.017$, 0.952, and 0.995, respectively. Their agreement near one supports a dataset-robust effective response coordinate. Full experimental details are provided in Appendix F.

6 Conclusion and discussion

Our results characterize a power-law learning curve as a dynamical response jointly shaped by the spectrum, target alignment, noise, and schedule. Target-weighted and squared-spectrum masses determine forcing and memory, respectively, while B_t/η_t controls noise injection in intrinsic time. We characterize when schedules preserve, change, or destroy clean power-law decay. In the canonical PLRF model, these mechanisms yield the $3+3(+2)$ map, an exact conditional finite-bulk gap law, and phase-dependent deterministic-equivalent risk and resource-optimal rates. Gaussian online-SGD experiments support the exact conditional and DE predictions. LLM experiments show our proxy surrogate yields cross-schedule prediction without refitting, supporting the response-level mechanism beyond frozen features. The factorization-collapse experiment further distinguishes the optimizer coordinates: plain SGD trajectories nearly collapse under B_t/η_t, whereas Muon (Jordan et al., 2024; Liu et al., 2025) trajectories nearly collapse under the coordinate B_t/η_t^2. The fitted $q_{\mathcal{K}} \approx 1$ corresponds to an effective $\alpha \approx 1/2$, where the $4+3$ theory predicts the near-square-root compute-optimal width scaling reported by Chinchilla (Paquette et al., 2024; Hoffmann et al., 2022).

Reproducibility Statement

The assumptions, limit regimes, and complete proofs for all theoretical results are provided in the appendix, including the finite-width, deterministic-equivalent, and long-time qualifications used in the main text. Appendix F specifies the model and optimizer configurations, schedule constructions, datasets, validation protocols, and scope of the LLM experiments. The supplementary material contains the experiment code, frozen configurations, processed validation traces, and plotting scripts needed to verify the reported numerical results and regenerate all experimental figures; it also documents the public data sources and the preparation and launch procedures for fresh training runs. Large training corpora and model checkpoints are not redistributed.

AI Use Statement

We used generative AI tools to assist with mathematical proof development, experimental design and implementation, and manuscript drafting and editing. All AI-assisted proofs, code, experimental results, and text were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

References

Alexander Atanasov, Blake Bordelon, Jacob A Zavatone-Veth, Courtney Paquette, and Cengiz Pehlevan. Two-point deterministic equivalence for stochastic gradient dynamics in linear models. *Advances in Theoretical and Mathematical Physics*, 30(1):3–39, 2026. doi: 10.4310/ATMP.260412235902. URL https://doi.org/10.4310/ATMP.260412235902.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

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

Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the origin of neural scaling laws: From random graphs to natural language. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.10684. Spotlight.

Nicholas H. Bingham, Charles M. Goldie, and Jozef L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1989.

Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model, 2026. URL https://arxiv.org/abs/2602.04774. Preprint.

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. PMLR, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

DeepSeek-AI. DeepSeek LLM: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3323. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bd18189308a4c45c7d71ca83acf3deaa-Abstract-Conference.html.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Yanjin He, Qingkai Zeng, and Meng Jiang. Pre-trained models perform the best when token distributions follow Zipf's law. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28009–28021. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1421. URL https://aclanthology.org/2025.emnlp-main.1421/.

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, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract.html.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f. Spotlight.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

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.

Dhruva Karkada, Joseph Turnbull, Yuxi Liu, and James B. Simon. Predicting kernel regression learning curves from only raw data statistics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c01bbfc198331bc0cbfb22b313877309-Abstract-Conference.html.

Andrej Karpathy. nanoGPT. https://github.com/karpathy/nanoGPT, 2022.

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3389. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/92abec9d3f278c648dfe99c8b8f35954-Abstract-Conference.html.

Binghui Li, Zilin Wang, Fengling Chen, Shiyang Zhao, Ruiheng Zheng, and Lei Wu. Optimal learning rate schedules under functional scaling laws: Power decay and warmup–stable–decay (extended abstract). In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4722–4723. PMLR, 2026a. URL https://proceedings.mlr.press/v336/li26b.html.

Jiaxiang Li, Zhiqi Bu, and Shiyun Xu. Towards joint scaling laws with optimal batch size schedules, 2026b. URL https://arxiv.org/abs/2607.27731.

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.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2022.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Yizhou Liu and Jeff Gore. Neural scaling universality: If exponents are fixed, time to understand coefficients. *arXiv preprint arXiv:2606.25008*, 2026. URL https://arxiv.org/abs/2606.25008.

Yizhou Liu, Ziming Liu, Cengiz Pehlevan, and Jeff Gore. Universal one-third time scaling in learning peaked distributions. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.03685.

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

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Alexandru Meterez, Depen Morwani, Jingfeng Wu, Costin-Andrei Oncescu, Cengiz Pehlevan, and Sham M. Kakade. Seesaw: Accelerating training by balancing batch size and learning rate scheduling. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/187f9ba4cd5a59477305ac712282a1cd-Paper-Conference.pdf.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model, 2026b. URL https://arxiv.org/abs/2607.21716. Preprint.

Nikolay Mikhaylovskiy. Zipf's and heaps' laws for tokens and LLM-generated texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15469–15481. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.837. URL https://aclanthology.org/2025.findings-emnlp.837/.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: Deterministic equivalents, test error, and GCV estimator, 2024. URL https://arxiv.org/abs/2403.08938.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36106–36159. PMLR, 2024. URL https://proceedings.mlr.press/v235/moniri24a.html.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3548–3626, 2021. URL https://proceedings.mlr.press/v134/paquette21a.html.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214:1–90, 2025. doi: 10.1007/s10107-024-02171-3. URL https://doi.org/10.1007/s10107-024-02171-3.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1dccfc3ee01871d05e33457c61037d59-Abstract-Conference.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, https://github.com/allenai/peS2o.

Jinbo Wang, Binghui Li, Zhanpeng Zhou, Mingze Wang, Yuxuan Sun, Jiaqi Zhang, Xunliang Cai, and Lei Wu. Fast catch-up, late switching: Optimal batch size scheduling via functional scaling laws. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0c7aaeb2745951f01224b3f95c75b528-Abstract-Conference.html.

Lechao Xiao. Rethinking conventional wisdom in machine learning: From generalization to scaling. *arXiv preprint arXiv:2409.15156*, 2024.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4558–4570, 2022.

APPENDIX CONTENTS

Notation guide **18**

A Spectral representations and formal power-law criteria **19**

A.1 From exact dynamics to deterministic-equivalent risk 19

A.2 Spectral criteria for power-law learning dynamics 22

B Proofs of the spectral power-law criteria **26**

B.1 Positive spectral representation of the deterministic equivalent 27

B.2 Uniform Tauberian transfer for moving spectra 27

B.3 Spectral criteria for forcing and memory 29

B.4 Accumulating memory over training 29

B.5 From forcing and memory asymptotics to observed loss under stable feedback . . . 31

C Joint learning-rate and batch-size schedules **32**

C.1 Exact risk dynamics under joint schedules 32

C.2 Time-inhomogeneous PLRF deterministic equivalent 33

C.3 Stable schedules and spectral-mode decay 34

C.4 PLRF regimes and admissible schedule families 35

C.5 Finite-bulk empirical spectrum, kernel, and gap 36

C.6 Long-memory and integrable-memory deterministic-equivalent risk asymptotics . . 38

C.7 Finite-bulk deterministic-equivalent risk asymptotics 44

C.8 Conditional transfer from DE risk asymptotics to realized SGD 46

C.9 Optimal ratio and schedule realizations 47

D Sharp transfer for regularly varying joint schedules **54**

D.1 From the empirical spectral tail to two-time memory 55

D.2 Examples and counterexamples for the transfer assumptions 57

D.3 Long-memory schedule transfer with feedback 64

D.4 Integrable-memory schedule transfer 68

D.5 From the exact noisy–clean gap back to B/η 70

D.6 Full preserve–change–destroy classification 72

D.7 Polynomial schedules in iteration time 75

E Constant learning-rate and batch-size schedules **76**

E.1 Scaling inputs and optimization setup 77

E.2 Discrete infinite-width fixed-noise optimum 79

E.3 Source-window rates and trace-class plateau 79

E.4 Proof of Theorem E.4 82

E.5 Clean-to-noisy crossover scales 84

F End-to-end language-model experiments **87**

F.1 Testing ratio-path effects at matched intrinsic time 87

F.2 Testing factorization invariance at matched ratio paths 88

F.3 A forcing-memory surrogate predicts unseen schedules 89

ORGANIZATION AND SCOPE OF THE APPENDIX RESULTS

The appendices are organized around the questions needed to verify the main text. Appendix A records the spectral representations and the complete formal criteria for power-law learning curves, while Appendix B proves the general criteria by translating cumulative slow-mode weight into long-time decay and then accounting for repeated noise feedback. Width-dependent finite-bulk results are kept with the exact finite-width and deterministic-equivalent risk results in Appendices C.5 and C.7. Appendices C and D turn to schedules in which learning rate and batch size vary together. They derive the resulting two-time risk equation, show how input power laws—including slowly varying corrections—pass to the noisy–clean gap, and prove when that gap determines the schedule ratio. The formal joint-schedule statements and their propagation, finite-bulk, and optimization proofs are now collected together in Appendix C. Appendix E places each constant-schedule result beside its calculation and optimality proof. Within Appendix D.2, a stretched-exponential example shows that temporal power laws need not come from a pointwise power-law spectrum. Numerical checks at finite width appear beside the results they are designed to illustrate, while Appendix F collects end-to-end language-model external-validity checks separately from the theorem-facing evidence.

Limit regimes: what grows and in what order. The results below use several asymptotic regimes because they answer different questions rather than provide interchangeable approximations to the same limit. Exact finite-width dynamics describe a realized feature map over a finite training horizon, while the fixed-width terminal limit describes what remains after all positive modes have relaxed. A fixed-time deterministic equivalent replaces the random empirical spectrum at a prescribed horizon. A fixed-limit long-time law then studies an infinite-spectrum response as $T \to \infty$, whereas a joint width–time limit asks whether that response remains valid along finite-width models with $T = T_m \to \infty$. Resource limits additionally allow width, horizon, and schedule to vary with the available data or feature compute. Table 1 summarizes the limit hierarchy used by the statements and proofs that follow.

Table 1: Limit regimes of the principal results.

Limit regime	**Results in this paper**
Exact finite width: $d, m, \boldsymbol{W}$ fixed, $t < \infty$	Exact finite-bulk kernel and noisy–clean gap (Lemma C.8 and Corollary C.9)
Fixed-width terminal: $d, m, \boldsymbol{W}$ fixed, $T \to \infty$	Stable terminal-risk transfer and the noise-induced plateau (Theorem A.9 and Propositions E.3 and E.5)
Fixed-time DE: $d, m \to \infty$, $d/m \to c$, T fixed	Positive spectral representation and DE risk recursion for varying schedules (Lemmas A.1 and C.5 and Proposition C.1)
Fixed infinite-spectrum system: $T \to \infty$	Schedule response, full classification, and gap-to-ratio identification (Theorems 5.1, A.6, D.13, D.16 and D.19 and Corollary D.23)
Joint width–time: $m, T_m \to \infty$, $1 \ll T_m \lesssim m^{2\alpha}$	Spectral iff criteria, uniform transfer, and LM/IM/FB risk laws (Theorems 3.1, B.1, C.14 and D.4, Corollaries A.10 and C.15, and Propositions C.18 and C.21)
Resource limit: $D \to \infty$ or $\mathfrak{f} \to \infty$; m, T, r vary	Optimal ratio control and phasewise data/compute rates (Theorem C.22, Corollary C.23, and Lemma C.24)

In the canonical PLRF model, the smallest positive spectral scale is $m^{-2\alpha}$. At fixed width, once $T \gg m^{2\alpha}$, all positive modes have been resolved and the remaining transient decays exponentially:

$$m \text{ fixed}, \qquad T \gg m^{2\alpha} \quad \Longrightarrow \quad \text{exponential transient}.$$

A temporal power law instead requires a joint window in which the filter scale T^{-1} remains above the finite-width cutoff:

$$m \to \infty, \qquad T = T_m \to \infty, \qquad 1 \ll T_m \ll m^{2\alpha}.$$

Thus the joint width–time limit defines the phenomenon studied here rather than serving only as a technical device. The order of limits also matters. Writing $R(T, m)$ schematically for the risk at

Table 2: Analytical levels and representative formal results.

Analytical level	**Formal results**
Exact conditional finite-width dynamics	Exact finite-bulk kernel and noisy–clean gap (Lemma C.8 and Corollary C.9).
Deterministic equivalent	Positive DE spectral representation, varying-schedule DE recursion, and conditional transfer to realized SGD (Lemma A.1 and Propositions C.1 and C.21).
Power-law asymptotics	Spectral iff criteria, uniform Tauberian and schedule transfer, and LM/IM/FB risk laws (Theorems A.4, B.1, C.14 and D.4 and Proposition C.18).

width m and intrinsic time T, the two orders can differ for some spectral families:

$$\lim_{m\to\infty}\lim_{T\to\infty} R(T,m) \neq \lim_{T\to\infty}\lim_{m\to\infty} R(T,m).$$

The left-hand side first trains each finite-width model to its terminal behavior. The right-hand side first produces an infinite spectrum, so progressively slower modes remain unresolved as T grows.

Analytical levels: what is exact and what is approximated. The limit hierarchy above specifies which variables grow and in what order. A separate distinction concerns what is proved at each level. Conditional on the frozen features $\boldsymbol{W}$, the finite-width risk dynamics are exact and require no spectral approximation. A deterministic equivalent then replaces the random empirical spectrum by deterministic spectral measures as width and dimension grow. This approximation is initially a fixed-horizon statement: validity for every fixed T does not by itself imply validity along $T = T_m \to \infty$. Power-law asymptotics make this additional long-time passage, either directly for empirical spectra or through the DE. They therefore require uniform control near the moving spectral scale T^{-1}, expressed through the corresponding spectral-control, no-escape, and source-window conditions. Table 2 records the formal results at each level.

NOTATION GUIDE

Hats on $\widehat{\boldsymbol{H}}, \widehat{\lambda}_j, \widehat{\boldsymbol{u}}_j$ mark empirical spectral objects after drawing $\boldsymbol{W}$; hats on scalar functions in Appendix D denote Laplace transforms. Plain F, K, S, R denote **exact dynamical quantities**, while $\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{R}$ denote their **deterministic-equivalent** counterparts. Where present, subscript $\boldsymbol{W}$ makes conditioning on the frozen features explicit; on risks, subscript 0 marks clean labels. Spectral measures with subscript m below are DE objects.

Table 3: Core notation; proof-local symbols are defined at first use.

Symbol	Meaning	Symbol	Meaning
Model and spectrum			
d	input dimension	m	feature width
α	spectral-decay exponent	β	target-regularity exponent
p	target-energy tail exponent $2\alpha + 2\beta - 1$	$\boldsymbol{x}$	covariate
$\boldsymbol{\Lambda}$	population covariance	λ_j	population eigenvalue
$\boldsymbol{\theta}^\star$	teacher coefficients	$f_\star$	teacher predictor
ε	label noise	σ^2	label-noise variance
$\boldsymbol{W}$	frozen feature matrix	$\boldsymbol{a}_t$	readout at update t
$\widehat{\boldsymbol{H}}$	$\boldsymbol{\Lambda}^{1/2}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Lambda}^{1/2}$	$\widehat{\lambda}_j$	empirical eigenvalue
$\widehat{\boldsymbol{u}}_j$	empirical eigenvector	$\boldsymbol{M}_m(z)$	diagonal matrix-valued DE resolvent
$\mathfrak{m}(z)$	scalar resolvent fixed point		
Dynamics and spectral objects			
η_t	learning rate	B_t	batch size
T_t	$\sum_{s<t}\eta_s$	T	terminal intrinsic time
r_s	B_s/η_s	$r(u)$	intrinsic-time ratio path
ϑ	regular-variation exponent of r	L_r	slow factor of r
$q_\eta(\lambda)$	one-step modal retention	$k(v)$	unit-injection survival at age v
F_t	exact varying-schedule forcing	$K_{t,s}$	exact two-time kernel
$F_{\boldsymbol{W}}$	exact conditional forcing	$K_{\boldsymbol{W}}$	exact conditional kernel
$\nu^{\mathcal{F}}_{\boldsymbol{W}}$	target-weighted empirical spectral measure	$\nu^{\mathcal{K}}_{\boldsymbol{W}}$	squared-eigenvalue empirical memory measure
$\boldsymbol{\mu}_m$	positive diagonal matrix-valued DE spectral measure	$\nu^{\mathcal{K}}_m$	λ^2-weighted DE memory measure
$\mu^{\mathcal{F}}_m$	target-weighted DE forcing measure	$\mu^{\mathcal{K}}_m$	DE trace spectral measure
$\mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m)$	joint-schedule DE forcing	$\mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t,s,m)$	joint-schedule DE kernel
$F_{\boldsymbol{W},0}$	exact zero-mode forcing	$F_{\boldsymbol{W},>0}$	exact learnable forcing
$\mathcal{F}_0$	DE null-space forcing component	$\mathcal{F}_{pp}$	DE pure-point forcing component
$\mathcal{F}_{ac}$	DE absolutely-continuous forcing component	$\mathcal{K}_{pp}$	DE pure-point memory component
Risks, scaling laws, and budgets			
$R_{\sigma,t}$	exact conditional noisy risk	$R_{0,t}$	exact conditional clean risk
$R_\sigma(T)$	exact fixed-limit noisy risk	$R_0(T)$	exact fixed-limit clean risk
$\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m)$	joint-schedule DE noisy risk	$\mathcal{R}_0$	clean DE risk; arguments suppressed
R_{app}	exact zero-mode floor	Z_t	exact noisy–clean gap
$\widetilde{R}_\sigma(w)$	exact risk generating series	$\mathcal{R}^\star_\sigma$	admissible-class DE optimum
$\mathcal{F}(T,m)$	PLRF large-clock phase-order forcing profile	$q_{\mathcal{F}}$	general learnable-forcing exponent; target-aligned source coordinate in the PLRF map
$L_{\mathcal{F}}$	forcing slow factor	$q_{\mathcal{K}}$	general one-injection memory exponent; PLRF coordinate on the LM and IM branches
$L_{\mathcal{K}}$	memory slow factor	q_0	complete floor-centered clean-loss exponent

Continued on the next page

Notation guide (continued)

Symbol	Meaning	Symbol	Meaning
L_0	centered clean-loss slow factor	$q_{\mathcal{N}}(\vartheta)$	schedule-induced noisy–clean-gap exponent
D_t	processed samples by update t	D	data budget
$\mathfrak{f}$	compute proxy, $\mathfrak{f} \asymp mD$	$\sim$	ratio tends to one
$\asymp$	same order	$\lesssim$	upper-order bound
$\ll$	negligible ratio	$\kappa_{\boldsymbol{W}}$	exact total kernel mass
ΔT_s	intrinsic-time increment, η_s		
Abbreviations and regimes			
PLRF	power-law random-feature model	FSL	functional scaling law
SGD	stochastic gradient descent	FLOP	floating-point operation
DE	deterministic equivalent	WSD	warmup–stable–decay
FB	finite-bulk regime	LM	long-memory regime
IM	integrable-memory regime		

A Spectral representations and formal power-law criteria

This appendix records the exact and deterministic-equivalent spectral representations used in the paper and states the formal criteria for the learnable forcing, one-injection memory, accumulated noise, and their transfer to the observed loss. The corresponding spectral and renewal proofs are collected in Appendix B, while the width-dependent finite-bulk results are proved in Appendices C.5 and C.7. Throughout, $a \asymp b$ hides dimension-independent positive constants.

A.1 From exact dynamics to deterministic-equivalent risk

We use the model, data distribution, online SGD update, conditional prediction risk, and PLRF parameters defined in Sections 2 and 4. Thus $\lambda_j = j^{-2\alpha}$, $\theta_j^\star = j^{-\beta}$, and $2\alpha + 2\beta > 1$. We take $d \geq c_0 m$, $c_0 > 1$, with $d/m \to c \in (1, \infty)$ when $\alpha < 1/2$ and $d/m \to c \in (1, \infty]$ when $\alpha > 1/2$.

For the constant-schedule spectral formulas below, set $\eta_t \equiv \eta$, $B_t \equiv B = \Theta(1)$, and retain $\boldsymbol{a}_0 = \mathbf{0}$. All sampling, independence, and conditioning conventions are those of Section 2. For this constant schedule, the stable learning-rate scale is

$$\eta \asymp \begin{cases} 1, & \alpha > \frac{1}{2}, \\ m^{2\alpha-1}, & 0 < \alpha < \frac{1}{2}. \end{cases} \tag{A.1}$$

The hidden constant is chosen strictly below the threshold in Eq. (A.14).

We first record the exact recursion stated in Section 2, then construct the positive deterministic-equivalent measures and assemble the corresponding risk. The formal power-law criteria follow in the next subsection.

A.1.1 Exact conditional forcing–memory dynamics

The operator $\widehat{\boldsymbol{H}}$, filter q_η, forcing $F_{\boldsymbol{W}}$, memory kernel $K_{\boldsymbol{W}}$, and approximation floor R_{app} are defined in Section 2; the two empirical spectral measures and their integral representations are given in Section 3. We verify here the exact recursion stated in Eq. (2.3). Let $\boldsymbol{e}_t := \boldsymbol{\Lambda}^{1/2}(\boldsymbol{W}\boldsymbol{a}_t - \boldsymbol{\theta}^\star)$ and $\rho_j(t) := \langle \widehat{\boldsymbol{u}}_j, \boldsymbol{e}_t \rangle$. Let

$$\mathscr{F}_t := \sigma\big(\boldsymbol{W}, \{(\boldsymbol{x}_s^i, y_s^i) : 0 \leq s < t,\ 1 \leq i \leq B\}\big), \qquad \mathbb{E}_t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathscr{F}_t].$$

Then $(\boldsymbol{a}_t, \boldsymbol{e}_t, \rho_j(t))$ are $\mathscr{F}_t$-measurable and the next mini-batch is independent of $\mathscr{F}_t$. Gaussian moments give

$$\mathbb{E}_t[\rho_j^2(t+1)] = q_\eta(\widehat{\lambda}_j)\rho_j^2(t) + \frac{\eta^2}{B}\widehat{\lambda}_j^2(\|\boldsymbol{e}_t\|_2^2 + \sigma^2).$$

Iterating this scalar recursion and summing over j proves Eq. (2.3).

Terminal risk and feedback stability. The exact Volterra equation also determines the fixed-width terminal level. Define the ordinary generating series, interpreted as formal power series,

$$\widetilde{R}_\sigma(w) := \sum_{t\geq 0} R_{\sigma,t} w^t, \qquad \widetilde{F}_{\boldsymbol{W}}(w) := \sum_{t\geq 0} F_{\boldsymbol{W}}(t) w^t, \qquad \widetilde{K}_{\boldsymbol{W}}(w) := \sum_{t\geq 0} K_{\boldsymbol{W}}(t) w^t.$$

Coefficient summation in Eq. (2.3) gives

$$\widetilde{R}_\sigma(w) = \frac{\widetilde{F}_{\boldsymbol{W}}(w) + \sigma^2 w \widetilde{K}_{\boldsymbol{W}}(w)/(1-w)}{1 - w\widetilde{K}_{\boldsymbol{W}}(w)}.$$

The reciprocal denominator $(1 - w\widetilde{K}_{\boldsymbol{W}}(w))^{-1}$ sums repeated reinjections of previous prediction error and is the transfer function of the SGD feedback.

For fixed $(d, m, \boldsymbol{W})$, the terminal risk can be computed exactly. Assume $\eta(1+1/B)\lambda_{\max}(\widehat{\boldsymbol{H}}) < 2$ and define the exact kernel mass

$$\kappa_{\boldsymbol{W}} := \sum_{t\geq 0} K_{\boldsymbol{W}}(t) = \frac{\eta}{B} \sum_{\widehat{\lambda}_j > 0} \frac{\widehat{\lambda}_j}{2 - (1+1/B)\eta\widehat{\lambda}_j}.$$

Since the kernel is nonnegative, $\kappa_{\boldsymbol{W}} < 1$ gives $|w\widetilde{K}_{\boldsymbol{W}}(w)| < 1$ for $|w| \leq 1$. The feedback is therefore stable, with input–output gain at most $(1-\kappa_{\boldsymbol{W}})^{-1}$. The positive-spectrum forcing then vanishes as $t \to \infty$, and the exact conditional terminal risk is

$$\lim_{t\to\infty} R_{\sigma,t} = \frac{R_{\mathrm{app}} + \sigma^2 \kappa_{\boldsymbol{W}}}{1 - \kappa_{\boldsymbol{W}}}. \tag{A.2}$$

Thus the terminal level combines frozen-representation error with persistent training-label noise and its feedback amplification. The centered-transfer result is stated in Theorem A.9 and proved in Appendix B.5.

A.1.2 Positive spectral measures for the deterministic equivalent

The random-matrix step preserves the same forcing–memory picture: it replaces the realized spectral projection by deterministic positive measures. Target weighting will produce the forcing measure, while the trace measure, together with the squared-eigenvalue variance weight, will produce memory.

To construct these measures, write λ_j for the diagonal entries of $\boldsymbol{\Lambda}$; in PLRF, $\lambda_j = j^{-2\alpha}$. The diagonal matrix-valued DE resolvent is

$$\boldsymbol{M}_m(z) := \mathrm{diag}\left(\frac{1}{\lambda_j \mathfrak{m}(z) - z} : 1 \leq j \leq d\right),$$

where $\mathfrak{m}(z)$ is the analytic solution of

$$\mathfrak{m}(z) = \left(1 + \frac{1}{m}\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j \mathfrak{m}(z) - z}\right)^{-1}.$$

Choose the Cauchy–Stieltjes branch of Paquette et al. (2024, Proposition E.1), characterized by $\mathrm{Im}\,\mathfrak{m}(z) < 0$ on the upper half-plane and $\mathfrak{m}(z) \to 1$ as $|z| \to \infty$.

For the matrix-valued measure $\boldsymbol{\mu}_m$ whose positivity and uniqueness are established next, the two scalar projections relevant to learning are

$$\mu_m^{\mathcal{F}}(A) := \left\langle \boldsymbol{\mu}_m(A)\boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star, \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star \right\rangle, \qquad \mu_m^{\mathcal{K}}(A) := \mathrm{tr}\,\boldsymbol{\mu}_m(A). \tag{A.3}$$

The squared-spectrum weighting used by every DE memory kernel is

$$\nu_m^{\mathcal{K}}(\mathrm{d}\lambda) := \lambda^2 \mu_m^{\mathcal{K}}(\mathrm{d}\lambda). \tag{A.4}$$

Its target-weighted zero atom is the DE approximation floor $\mathcal{F}_0(m) := \mu_m^{\mathcal{F}}(\{0\})$.

The next lemma turns the DE resolvent into genuine positive spectral measures, so the contour formulas used later are ordinary spectral averages rather than only formal complex integrals. It also records the moments and functional calculus needed below.

Lemma A.1 (Positive spectral measures of the PLRF deterministic equivalent)**.** For every finite m, d, there is a unique finite positive-semidefinite diagonal matrix-valued Borel measure $\boldsymbol{\mu}_m$, compactly supported on $[0, \infty)$, such that

$$\boldsymbol{M}_m(z) = \int_{[0,\infty)} \frac{\boldsymbol{\mu}_m(\mathrm{d}\lambda)}{\lambda - z}, \qquad z \in \mathbb{C} \setminus [0, \infty). \tag{A.5}$$

It satisfies

$$\boldsymbol{\mu}_m([0,\infty)) = \boldsymbol{I}_d, \qquad \int \lambda \, \boldsymbol{\mu}_m(\mathrm{d}\lambda) = \boldsymbol{\Lambda}, \qquad \int \lambda^2 \, \boldsymbol{\mu}_m(\mathrm{d}\lambda) = \boldsymbol{\Lambda}^2 + \frac{\operatorname{tr} \boldsymbol{\Lambda}}{m} \boldsymbol{\Lambda}. \tag{A.6}$$

The scalar measures in Eq. (A.3) are finite and nonnegative, with

$$\mu_m^{\mathcal{F}}([0,\infty)) = (\boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda} \boldsymbol{\theta}^\star, \qquad \mu_m^{\mathcal{K}}([0,\infty)) = d,$$
$$\int \lambda \, \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) = (\boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}^2 \boldsymbol{\theta}^\star, \quad \int \lambda \, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda) = \operatorname{tr} \boldsymbol{\Lambda}, \quad \int \lambda^2 \, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda) = \operatorname{tr}(\boldsymbol{\Lambda}^2) + \frac{(\operatorname{tr} \boldsymbol{\Lambda})^2}{m}. \tag{A.7}$$

If h is analytic on a neighborhood of $\operatorname{supp} \mu_m^{\mathcal{K}}$, then

$$-\frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}} h(z) \boldsymbol{M}_m(z) \, \mathrm{d}z = \int h(\lambda) \boldsymbol{\mu}_m(\mathrm{d}\lambda). \tag{A.8}$$

The proof of Lemma A.1 is in Appendix B.1.

A.1.3 Assembling the deterministic-equivalent risk

We now return to the risk. The DE retains the exact system's input–output structure: its spectral transforms define forcing and memory, and stable positive feedback keeps the noisy–clean gap within a fixed factor of cumulative memory.

Let $\mathscr{C}$ enclose the deterministic-equivalent spectral support and no other singularities. The DE forcing and memory are

$$\mathcal{F}(t, m) := -\frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}} \left\langle \boldsymbol{M}_m(z) \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \right\rangle q_\eta(z)^t \, \mathrm{d}z, \tag{A.9}$$

$$\mathcal{K}(t, m) := \frac{\eta^2}{B} \operatorname{tr} \left[-\frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}} \boldsymbol{M}_m(z) z^2 q_\eta(z)^t \, \mathrm{d}z \right]. \tag{A.10}$$

By Eq. (A.8), these contour expressions are positive spectral transforms.

$$\mathcal{F}_0(m) = -\operatorname*{Res}_{z=0} \left\langle \boldsymbol{M}_m(z) \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \right\rangle, \qquad \mathcal{F}_{>0}(t, m) := \mathcal{F}(t, m) - \mathcal{F}_0(m), \tag{A.11}$$

so that $\mathcal{F}(t, m) = \mathcal{F}_0(m) + \mathcal{F}_{>0}(t, m)$. Only the zero-versus-positive split is universal. For general $(\lambda_i, \theta_i^\star)$, we keep the positive-spectrum forcing intact; the later pure-point/absolutely-continuous split is PLRF-specific. The temporal spectral and accumulation criteria below are first stated for exact conditional quantities; the transfer theorem then covers both exact and DE losses.

To assemble the components into risk, define the causal discrete convolution for sequences $a, b : \mathbb{N}_0 \to \mathbb{R}$ by

$$(a * b)(0) := 0, \qquad (a * b)(t) := \sum_{s=0}^{t-1} a(t - 1 - s) b(s), \quad t \geq 1.$$

The clean DE risk is the bounded response to the forcing under this kernel:

$$\mathcal{R}_0 = \mathcal{F} + \mathcal{K} * \mathcal{R}_0. \tag{A.12}$$

This is the zero-noise constant-schedule resolvent-DE surrogate for the SGD dynamics in Section 2.

Its noisy counterpart adds cumulative label-noise injection:

$$\mathcal{R}_\sigma = \mathcal{F} + \mathcal{K} * \mathcal{R}_\sigma + \sigma^2 \mathcal{S}, \qquad \mathcal{S}_t(m) := \sum_{s=0}^{t-1} \mathcal{K}(s, m), \quad \mathcal{S}_0(m) := 0. \tag{A.13}$$

The comparison requires two uniform stability conditions: every positive mode contracts, and the total Volterra feedback remains below one:

$$\eta\left(1+\frac{1}{B}\right)\sup\{\lambda>0:\lambda \text{ lies in the DE spectral support of } \boldsymbol{M}_m\}\le 2-\delta,$$
$$\|\mathcal{K}\|_{\ell^1}:=\sum_{t=0}^{\infty}\mathcal{K}(t,m)\le\kappa<1. \tag{A.14}$$

Here $\delta>0$ and $\kappa<1$ are fixed uniformly in m.

Because the kernel is nonnegative, the direct injection gives the lower bound, while repeated stable feedback amplifies it by at most $(1-\kappa)^{-1}$. Hence

$$\mathcal{R}_0(t,m)+\sigma^2\mathcal{S}_t(m)\le\mathcal{R}_\sigma(t,m)\le\mathcal{R}_0(t,m)+\frac{\sigma^2}{1-\kappa}\mathcal{S}_t(m). \tag{A.15}$$

This is the reduction used below: once the clean-risk and cumulative-memory scales are known, the noisy risk follows up to constants fixed by stability.

A.2 Spectral criteria for power-law learning dynamics

This subsection follows the temporal components to the observed loss. Weighted empirical tails determine forcing and one-injection memory, cumulative memory determines the effect of persistent noise, and stable Volterra feedback transfers these components to the observed loss.

A positive function L is slowly varying if $L(cT)/L(T)\to 1$ for every fixed $c>0$.

Forcing and memory: spectral tails determine temporal decay. Forcing has a power law exactly when the target-weighted low-spectrum mass does, while one-injection memory has a power law exactly when the squared-spectrum mass does. A mode of size λ is learned around λ^{-1}, so at time T both component decays are read near the moving cutoff T^{-1}.

We study $K_{\boldsymbol{W}}$ because persistent noise depends on its sum, $S_{\boldsymbol{W},t}:=\sum_{u<t}K_{\boldsymbol{W}}(u)$, analyzed next.

Every fixed finite spectrum eventually decays exponentially, so the equivalences use the following common joint-limit condition. It excludes step-boundary artifacts and moving spectral packets without presupposing a target power law. Proofs are in Appendix B.3.

Assumption A.2 (Constant-schedule stability)**.** A constant schedule $\eta_t\equiv\eta$, $B_t\equiv B$ satisfies constant-schedule stability if there are constants $\delta>0$ and $\kappa<1$, uniform over the width, such that:

(a) *Pointwise contraction.*
$$\eta\left(1+\frac{1}{B}\right)\lambda_{\max}(\widehat{\boldsymbol{H}})\le 2-\delta.$$

(b) *Row stability.*
$$\sup_{t\ge1}\sum_{s<t}K_{\boldsymbol{W}}(t-1-s)=\sum_{h\ge0}K_{\boldsymbol{W}}(h)\le\kappa.$$

For the DE recursion, pointwise contraction is imposed uniformly over $\lambda\in\operatorname{supp}\mu_m^{\mathcal{K}}$, and row stability replaces $K_{\boldsymbol{W}}(h)$ by $\mathcal{K}(h,m)$.

Assumption A.3 (Uniform spectral window)**.** Retain the constant schedule, assume part *(a)* of Assumption A.2, and consider a joint sequence with $m\to\infty$, $t\to\infty$, and $T=\eta t\to\infty$. For the component under consideration, let $A_m(x)$ denote its cumulative weighted spectral mass below x. Fix $q>0$, a slowly varying function L, and a normalization constant $c_A>0$, and set

$$a_m:=c_AT^{-q}L(T).$$

There exist $C<\infty$ and $\epsilon\in(0,q)$, independent of m,t, such that eventually, for every $y>0$,

$$\frac{A_m(y/T)}{a_m}\le C\max\{y^{q-\epsilon},y^{q+\epsilon}\},\qquad y>0. \tag{A.16}$$

Theorem A.4 (Spectral criteria for power-law forcing and memory)**.** Retain the constant schedule $\eta_t \equiv \eta$, $B_t \equiv B$. For each width, let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, with $\lambda_j \geq 0$ and $(\boldsymbol{\theta}^\star)^\top \mathbf{\Lambda} \boldsymbol{\theta}^\star < \infty$, so the covariance and finite-energy target are otherwise arbitrary. Condition on $\boldsymbol{W}$, let $(\widehat{\lambda}_j, \widehat{\boldsymbol{u}}_j)$ be the empirical eigenpairs in Section 2, and retain the empirical measures $\nu^{\mathcal{F}}_{\boldsymbol{W}}, \nu^{\mathcal{K}}_{\boldsymbol{W}}$ above.

Forcing. Under zero initialization, the initial residual is $-\mathbf{\Lambda}^{1/2}\boldsymbol{\theta}^\star$. Let $q_{\mathcal{F}} > 0$ and let $L_{\mathcal{F}}$ be slowly varying. Assume Assumption A.3 with

$$A_m(x) = \nu^{\mathcal{F}}_{\boldsymbol{W}}((0, x]), \qquad q = q_{\mathcal{F}}, \qquad L = L_{\mathcal{F}}, \qquad c_A = 1.$$

Then

$$\nu^{\mathcal{F}}_{\boldsymbol{W}}((0, T^{-1}]) \sim T^{-q_{\mathcal{F}}} L_{\mathcal{F}}(T)$$

if and only if

$$F_{\boldsymbol{W},>0}(t) \sim \Gamma(q_{\mathcal{F}} + 1)(2T)^{-q_{\mathcal{F}}} L_{\mathcal{F}}(T) \qquad (T \to \infty).$$

Memory. Let $q_{\mathcal{K}} > 0$ and let $L_{\mathcal{K}}$ be slowly varying. Assume Assumption A.3 with

$$A_m(x) = \nu^{\mathcal{K}}_{\boldsymbol{W}}((0, x]), \qquad q = q_{\mathcal{K}}, \qquad L = L_{\mathcal{K}}, \qquad c_A = \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)}.$$

Then

$$\nu^{\mathcal{K}}_{\boldsymbol{W}}((0, T^{-1}]) \sim \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)} T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T)$$

if and only if

$$K_{\boldsymbol{W}}(t) \sim \frac{\eta^2}{B} T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T). \tag{A.17}$$

Each asymptotic equivalence above is understood locally uniformly under constant-factor rescalings of T.

Fixed-margin stability is standard (Paquette et al., 2021; 2025); fixed-spectrum Potter bounds follow from Bingham et al. (1989, Theorem 1.5.6).

Remark A.5 (Joint-limit uniformity)**.** Empirical spectral control is available (Misiakiewicz & Saeed, 2024; Defilippis et al., 2024); the simultaneous width–time uniformity used above is assumed.

Cumulative memory: growth or convergence. Summing one-injection memory creates the dividing line $q_{\mathcal{K}} = 1$. Below it, cumulative memory has asymptotic scale $(\eta/B)T^{1-q_{\mathcal{K}}} L_{\mathcal{K}}(T)$; above it, memory converges with a remaining tail of the same order. At equality, a fixed infinite spectrum has a borderline integral. The theorem also gives a direct spectral criterion for the terminal tail without assuming a power law for $K_{\boldsymbol{W}}$. The proof is in Appendix B.4.

Under the same constant schedule and conditional on $\boldsymbol{W}$, retain $S_{\boldsymbol{W},t}$ above and define its total mass by $S_{\boldsymbol{W},\infty} := \sum_{u=0}^{\infty} K_{\boldsymbol{W}}(u)$.

Theorem A.6 (When cumulative memory grows or converges)**.** Under the same constant schedule, condition on $\boldsymbol{W}$, retain q_η and $K_{\boldsymbol{W}}$. Assume part *(a)* of Assumption A.2.

Cumulative-memory asymptotics. Let $q_{\mathcal{K}} > 0$, let $L_{\mathcal{K}}$ be slowly varying, and consider a joint limit in which $m \to \infty$, $t \to \infty$, and $T = \eta t \to \infty$. Set

$$A^{\mathcal{K}}_m(x) := \nu^{\mathcal{K}}_{\boldsymbol{W}}((0, x]), \qquad a^{\mathcal{K}}_m := \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)} T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T), \qquad \widetilde{A}^{\mathcal{K}}_m(y) := \frac{A^{\mathcal{K}}_m(y/T)}{a^{\mathcal{K}}_m}.$$

Suppose the memory cutoff asymptotic

$$\widetilde{A}^{\mathcal{K}}_m(y) \longrightarrow y^{q_{\mathcal{K}}} \tag{A.18}$$

holds locally uniformly for $y \in (0, \infty)$. For the branch under consideration, also suppose that there exist $C < \infty$ and $\epsilon > 0$ such that, eventually, for every $y > 0$,

$$\widetilde{A}^{\mathcal{K}}_m(y) \leq C \max\{y^{q_{\mathcal{K}} - \epsilon}, y^{q_{\mathcal{K}} + \epsilon}\}, \qquad y > 0. \tag{A.19}$$

with constants uniform along the joint limit and

$$0 < \epsilon < \min\{q_{\mathcal{K}}, 1 - q_{\mathcal{K}}\} \quad \text{if } 0 < q_{\mathcal{K}} < 1, \qquad 0 < \epsilon < q_{\mathcal{K}} - 1 \quad \text{if } q_{\mathcal{K}} > 1.$$

Then, for $0 < q_{\mathcal{K}} < 1$,

$$S_{\boldsymbol{W},t} \sim \frac{\eta}{B} \frac{1}{1 - q_{\mathcal{K}}} T^{1-q_{\mathcal{K}}} L_{\mathcal{K}}(T), \tag{A.20}$$

whereas for $q_{\mathcal{K}} > 1$,

$$S_{\boldsymbol{W},\infty} - S_{\boldsymbol{W},t} \sim \frac{\eta}{B} \frac{1}{q_{\mathcal{K}} - 1} T^{1-q_{\mathcal{K}}} L_{\mathcal{K}}(T). \tag{A.21}$$

Under Assumption A.3, Eq. (A.18) is equivalent to Eq. (A.17). All these hypotheses and conclusions are locally uniform under constant-factor rescalings of T.

Direct terminal-tail criterion. Without assuming a power law for $K_{\boldsymbol{W}}$, consider the same joint intermediate limit as in Theorem A.4, and let $q_{\mathcal{S}} > 0$ and $L_{\mathcal{S}}$ be slowly varying. Define

$$A_m^{\mathrm{tail}}(x) := \frac{\eta}{B} \sum_{0 < \widehat{\lambda}_j \le x} \frac{\widehat{\lambda}_j}{2 - (1 + 1/B)\eta\widehat{\lambda}_j}.$$

Suppose that, for some $\epsilon \in (0, q_{\mathcal{S}})$ and $C < \infty$, it eventually satisfies, for every $y > 0$,

$$\frac{A_m^{\mathrm{tail}}(y/T)}{T^{-q_{\mathcal{S}}} L_{\mathcal{S}}(T)} \le C \max\{y^{q_{\mathcal{S}} - \epsilon}, y^{q_{\mathcal{S}} + \epsilon}\}, \qquad y > 0.$$

Under this envelope, the following two locally uniform families are equivalent. For $c > 0$, put $t_c := \lfloor ct \rfloor$ and $T_c := \eta t_c$, so $T_c \sim cT$ locally uniformly in c. Then, for every compact $J \subset (0, \infty)$,

$$\sup_{c \in J} \left| \frac{A_m^{\mathrm{tail}}(1/T_c)}{T_c^{-q_{\mathcal{S}}} L_{\mathcal{S}}(T_c)} - 1 \right| \longrightarrow 0$$

if and only if

$$\sup_{c \in J} \left| \frac{S_{\boldsymbol{W},\infty} - S_{\boldsymbol{W},t_c}}{\Gamma(q_{\mathcal{S}} + 1)(2T_c)^{-q_{\mathcal{S}}} L_{\mathcal{S}}(T_c)} - 1 \right| \longrightarrow 0.$$

Remark A.7 (Exact finite-width accumulation)**.** The finite-width identities underlying the theorem are

$$S_{\boldsymbol{W},t} = \frac{\eta}{B} \sum_{\widehat{\lambda}_j > 0} \frac{\widehat{\lambda}_j [1 - q_\eta(\widehat{\lambda}_j)^t]}{2 - (1 + 1/B)\eta\widehat{\lambda}_j}, \qquad S_{\boldsymbol{W},\infty} - S_{\boldsymbol{W},t} = \frac{\eta}{B} \sum_{\widehat{\lambda}_j > 0} \frac{\widehat{\lambda}_j q_\eta(\widehat{\lambda}_j)^t}{2 - (1 + 1/B)\eta\widehat{\lambda}_j}. \tag{A.22}$$

For a fixed infinite-spectrum kernel that satisfies $K(t) \sim (\eta^2/B) T^{-1} L_{\mathcal{K}}(T)$ at all sufficiently large times, ordinary one-sequence summation gives $(\eta/B) \int_1^T L_{\mathcal{K}}(u)\, \mathrm{d}u/u$ when this integral diverges, or the remaining tail $(\eta/B) \int_T^\infty L_{\mathcal{K}}(u)\, \mathrm{d}u/u$ when it converges. For the DE kernel, retain the squared-spectrum measure in Eq. (A.4) and set $A_{m,\mathrm{DE}}^{\mathcal{K}}(x) := \nu_m^{\mathcal{K}}((0, x])$. The same accumulation and remaining-tail conclusions hold when this cumulative measure satisfies the corresponding cutoff and endpoint-envelope hypotheses and the DE stability margin holds. The direct terminal-tail criterion instead uses

$$A_{m,\mathrm{DE}}^{\mathrm{tail}}(x) := \frac{\eta}{B} \int_{(0,x]} \frac{\lambda}{2 - (1 + 1/B)\eta\lambda}\, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda).$$

The same DE proofs apply under the stated cutoff, endpoint-envelope, and stability hypotheses. The next result transfers these forcing and memory asymptotics to loss.

From components to the observed loss. Under either transfer condition, stable feedback introduces no scale beyond the displayed forcing, cumulative-memory, and renewal terms. The clean loss follows the floor and forcing, while the noisy loss also follows cumulative memory; renewal feedback adds one explicit kernel-sized term. At a terminal plateau, the same components govern centered convergence when the terminal tail remains uniformly controlled under renewal and the two positive post-feedback responses remain quantitatively noncancelling. The theorem covers both the exact conditional recursion and its deterministic equivalent.

The exact conditional loss and its deterministic equivalent obey the same positive Volterra system, so we state their common transfer result once. Use the following neutral notation for the exact quantities in Eq. (2.3) and the DE quantities in Eqs. (A.9), (A.10) and (A.13):

$$\begin{aligned}(\mathsf{F},\mathsf{K};\mathsf{R}_\sigma,\mathsf{R}_0) &= (F_{\boldsymbol{W}},K_{\boldsymbol{W}};R_{\sigma,\cdot},R_{0,\cdot}), && \text{exact conditional},\\ &= (\mathcal{F},\mathcal{K};\mathcal{R}_\sigma,\mathcal{R}_0), && \text{deterministic equivalent}.\end{aligned}$$

Suppress the explicit $\boldsymbol{W}$-conditioning and use m as the common width index. In either case, write

$$\mathsf{F}(t,m)=\mathsf{F}_0(m)+\mathsf{F}_{>0}(t,m),\qquad \mathsf{S}_t(m):=\sum_{u=0}^{t-1}\mathsf{K}(u,m),\qquad \kappa_m:=\sum_{t\ge 0}\mathsf{K}(t,m).$$

Whenever the fixed-width terminal limit exists and $\kappa_m<1$, denote it by $P_m:=\lim_{t\to\infty}\mathsf{R}_\sigma(t,m)$. For the centered response, also set

$$\overline{\mathsf{K}}_m(t):=\kappa_m-\mathsf{S}_t(m),\qquad \mathsf{G}_m g:=g+\sum_{n\ge 1}\mathsf{K}^{(*n)}*g.$$

Assumption A.8 (Long-tail conditions for transferring component laws to risk)**.** Assume part *(b)* of Assumption A.2. Uniformly along the common intermediate window, both $\mathsf{F}_{>0}$ and K are long-tailed: each is asymptotically unchanged by any fixed time shift.

Theorem A.9 (How stable feedback transfers components to observed loss)**.** Under the same constant schedule, fix either of the two preceding rows and assume Assumption A.8. The transfer has two forms:

(i) *Forcing preservation.* The leading loss scales are

$$\mathsf{R}_0(t,m)\asymp\mathsf{F}_0(m)+\mathsf{F}_{>0}(t,m),\qquad \mathsf{R}_\sigma(t,m)\asymp\mathsf{F}_0(m)+\mathsf{F}_{>0}(t,m)+\sigma^2\mathsf{S}_t(m).\tag{A.23}$$

These comparisons hold provided

$$\left[\sum_{n\ge 1}\mathsf{K}^{(*n)}(t,m)\right]\sum_{s=0}^{t}\mathsf{F}_{>0}(s,m)=o(\mathsf{F}_{>0}(t,m))$$

uniformly in the window.

(ii) *Renewal feedback.* The same comparisons hold after adding

$$\left[\sum_{s\ge 0}\mathsf{F}_{>0}(s,m)\right]\mathsf{K}(t,m)$$

to both right-hand sides. This conclusion holds when $\sum_{t\ge 0}\mathsf{F}_{>0}(t,m)<\infty$ and the standard uniform subexponential relations

$$(\mathsf{K}*\mathsf{K})(t,m)\asymp\mathsf{K}(t,m),\qquad (\mathsf{K}*\mathsf{F}_{>0})(t,m)\asymp\|\mathsf{K}\|_{\ell^1}\mathsf{F}_{>0}(t,m)+\left[\sum_{s\ge 0}\mathsf{F}_{>0}(s,m)\right]\mathsf{K}(t,m)$$

hold with summable domination over the convolution powers.

If $\mathsf{F}_{>0}(t,m)\to 0$, the fixed-width terminal plateau exists and satisfies

$$P_m=\frac{\mathsf{F}_0(m)+\sigma^2\kappa_m}{1-\kappa_m}.\tag{A.24}$$

For the centered response, the positive defective resolvent gives the exact identity

$$\mathsf{R}_\sigma(t,m)-P_m=(\mathsf{G}_m\mathsf{F}_{>0})(t)-(P_m+\sigma^2)(\mathsf{G}_m\overline{\mathsf{K}}_m)(t).\tag{A.25}$$

To reduce this identity to the original forcing and kernel tail, assume, uniformly across the joint window, the terminal-tail renewal domination

$$\sum_{n\ge 1}(\mathsf{K}^{(*n)}*\overline{\mathsf{K}}_m)(t)\le C_{\mathrm{tail}}\overline{\mathsf{K}}_m(t)\tag{A.26}$$

and the quantitative noncancellation condition

$$|\mathsf{R}_\sigma(t,m) - P_m| \geq c_{\mathrm{nc}} \left[(\mathsf{G}_m \mathsf{F}_{>0})(t) + (P_m + \sigma^2)(\mathsf{G}_m \overline{\mathsf{K}}_m)(t)\right], \qquad c_{\mathrm{nc}} > 0. \tag{A.27}$$

Here $C_{\mathrm{tail}} < \infty$ and $c_{\mathrm{nc}} > 0$ are independent of m, t throughout that window. Then, in case *(i)*,

$$|\mathsf{R}_\sigma(t,m) - P_m| \asymp \mathsf{F}_{>0}(t,m) + \left(P_m + \sigma^2\right)\left[\kappa_m - \mathsf{S}_t(m)\right]. \tag{A.28}$$

In case *(ii)*, the right-hand side additionally contains $\left[\sum_{s\geq 0} \mathsf{F}_{>0}(s,m)\right] \mathsf{K}(t,m)$.

The transfer theorem is generic; we now evaluate its spectral inputs in PLRF.

PLRF exponents. For PLRF, the abstract spectral criteria become two explicit coordinates. With intrinsic time $T = \eta t$, finite target energy gives $q_{\mathcal{F}} = (2\alpha + 2\beta - 1)/(2\alpha)$; on the LM and IM branches $\alpha > 1/4$, one-injection memory has $q_{\mathcal{K}} = 2 - 1/(2\alpha)$. Target alignment therefore affects forcing, whereas memory depends only on spectral decay. The corollary records the population-filter constants and admissible parameter range before we explain the cutoff calculation.

Corollary A.10 (PLRF forcing and memory exponents)**.** Consider the PLRF specialization $\lambda_i = i^{-2\alpha}$ and $\theta_i^\star = i^{-\beta}$. If $2\alpha + 2\beta > 1$, then the forcing exponent is $q_{\mathcal{F}} = (2\alpha + 2\beta - 1)/(2\alpha)$. As $T \to \infty$, the population target filter satisfies

$$\sum_i \lambda_i(\theta_i^\star)^2 e^{-2T\lambda_i} \sim \frac{\Gamma\left(1 + \frac{2\alpha+2\beta-1}{2\alpha}\right)}{2\alpha + 2\beta - 1} (2T)^{-\frac{2\alpha+2\beta-1}{2\alpha}}.$$

If additionally $\alpha > 1/4$, then the memory exponent is $q_{\mathcal{K}} = 2 - 1/(2\alpha)$. As $T \to \infty$, the population memory filter satisfies

$$\sum_i \lambda_i^2 e^{-2T\lambda_i} \sim \frac{\Gamma\left(3 - \frac{1}{2\alpha}\right)}{4\alpha - 1} (2T)^{-2+1/(2\alpha)}.$$

These exponents come from the moving spectral cutoff. At intrinsic time $T = \eta t$, the critical modes satisfy $q_\eta(\lambda)^t = e^{-2T\lambda}(1 + o(1))$. Consequently, in the common intermediate window,

$$\mathcal{F}_{pp}(t,m) \asymp \sum_i \lambda_i(\theta_i^\star)^2 e^{-2T\lambda_i}, \qquad \mathcal{K}_{pp}(t,m) \asymp \frac{\eta^2}{B} \sum_i \lambda_i^2 e^{-2T\lambda_i}.$$

These are respectively unlearned target error and one surviving variance injection. For joint schedules, use elapsed intrinsic time and the injection weight in Eq. (C.29).

Since only modes with $\lambda_i T \lesssim 1$ remain, the filters reduce to

$$\sum_{0<\lambda_i \lesssim T^{-1}} \lambda_i(\theta_i^\star)^2, \qquad \sum_{0<\lambda_i \lesssim T^{-1}} \lambda_i^2.$$

Direct summation gives the displayed exponents; see Appendix B.3.

Two boundaries delimit these formulas. As $2\alpha + 2\beta \downarrow 1$, $q_{\mathcal{F}} \downarrow 0$. At $\alpha = 1/4$, the LM/IM memory formula gives way to FB.

For $\alpha < 1/4$, Lemmas C.7 and C.8 give the empirical width-scale band and its joint-schedule kernel. The resulting exact conditional noisy–clean gap is closed in Corollary C.9, without a target-forcing comparison. Separately, Lemma C.16 and Proposition C.17 give the DE kernel and gap; explicit DE total-risk asymptotics retain the forcing range in Lemma C.11. For $\alpha > 1/4$, Lemma C.12 proves LM/IM DE kernel transfer, with that source restriction only for the explicit PLRF forcing profile.

B Proofs of the spectral power-law criteria

This section supplies the proof path from frozen spectral structure to the observed learning curve. Appendix B.1 first proves that the deterministic-equivalent resolvent defines positive spectral measures. Appendix B.2 then shows why weighted spectral mass near zero is equivalent to long-time decay. Appendix B.3 applies this principle to the learnable target error and the survival of one noise injection. Appendix B.4 sums these injections over training, and Appendix B.5 shows how stable recurrence carries the component rates to the observed loss. Finite-bulk results, whose spectral scale depends explicitly on width, are kept with the corresponding exact finite-width and deterministic-equivalent risk results in Appendices C.5 and C.7.

B.1 Positive spectral representation of the deterministic equivalent

Proof of Lemma A.1. The proof has two main steps: a Herglotz representation gives positivity and support, while the expansion at infinity gives the required moments. For $1 \le j \le d$, set $r_{j,m}(z) := 1/(\lambda_j \mathfrak{m}(z) - z)$. The selected branch places this function in the positive Herglotz class. By Paquette et al. (2024, Proposition E.1), $\operatorname{Im} \mathfrak{m}(z) < 0$ on the upper half-plane, hence $\operatorname{Im}(\lambda_j \mathfrak{m}(z) - z) < 0$ and $\operatorname{Im} r_{j,m}(z) > 0$. Together with $r_{j,m}(z) = -z^{-1} + O(|z|^{-2})$, the Herglotz–Stieltjes representation gives a unique probability measure $\mu_{j,m}$ with

$$r_{j,m}(z) = \int_{\mathbb{R}} \frac{\mu_{j,m}(\mathrm{d}\lambda)}{\lambda - z}.$$

We next verify that this measure is supported on the nonnegative real axis and has compact support. For $z = -s < 0$, the fixed-point equation is equivalent to

$$1 = u + \frac{1}{m} \sum_{j=1}^{d} \frac{\lambda_j u}{\lambda_j u + s}.$$

The right-hand side increases strictly from zero past one on $(0, 1)$, so it has a unique solution $u = \mathfrak{m}(-s) \in (0, 1)$. Its positive u-derivative and the implicit-function theorem continue the branch across the negative axis, excluding mass on $(-\infty, 0)$; the far-field estimate of Paquette et al. (2024, Proposition E.2) gives compact support.

Collecting the scalar measures now gives $\boldsymbol{\mu}_m(A) := \operatorname{diag}\big(\mu_{1,m}(A), \dots, \mu_{d,m}(A)\big)$. This proves Eq. (A.5) and positivity. The expansion at infinity, $\mathfrak{m}(z) = 1 + \operatorname{tr}(\boldsymbol{\Lambda})/(mz) + O(|z|^{-2})$, identifies the stated moment identities; in particular,

$$r_{j,m}(z) = -\frac{1}{z} - \frac{\lambda_j}{z^2} - \frac{\lambda_j^2 + \lambda_j \operatorname{tr}(\boldsymbol{\Lambda})/m}{z^3} + O(|z|^{-4}).$$

Comparing coefficients with the Stieltjes expansion gives

$$\mu_{j,m}([0,\infty)) = 1, \qquad \int \lambda \, \mu_{j,m}(\mathrm{d}\lambda) = \lambda_j, \qquad \int \lambda^2 \, \mu_{j,m}(\mathrm{d}\lambda) = \lambda_j^2 + \frac{\operatorname{tr} \boldsymbol{\Lambda}}{m} \lambda_j.$$

This proves Eq. (A.6) and Eq. (A.7). It remains to identify the zero atom and the contour calculus. The near-zero estimate of Paquette et al. (2024, Appendix E) makes zero an isolated at-most-simple pole, so

$$-\operatorname*{Res}_{z=0} \left\langle \boldsymbol{M}_m(z) \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \right\rangle = \mu_m^{\mathcal{F}}(\{0\}).$$

Finally, substituting the measure representation into the contour integral and applying Cauchy's formula componentwise gives Eq. (A.8). □

B.2 Uniform Tauberian transfer for moving spectra

Theorem B.1 (Uniform Tauberian equivalence for m-dependent spectra)**.** Let $A_m : [0,\infty) \to [0,\infty)$ be bounded, nondecreasing, and right-continuous, with $A_m(0) = 0$, and consider any joint sequence with $m \to \infty$ and $T = T_m \to \infty$. Fix $\tau > 0$, let L be slowly varying at infinity, and set $a_m := T^{-\tau} L(T)$. Assume the following two conditions.

(a) *Uniform spectral control.* For some $\epsilon \in (0, \tau)$ and $C < \infty$, eventually

$$\frac{A_m(y/T)}{a_m} \le C \max\{y^{\tau-\epsilon}, y^{\tau+\epsilon}\}, \qquad y > 0. \tag{B.1}$$

(b) *Compatible training filter.* Let $h_m : (0,\infty) \to [0,\infty)$ satisfy, for every $A < \infty$,

$$\sup_{0<y\le A} |T h_m(y/T) - y| \longrightarrow 0, \tag{B.2}$$

and at every point where A_m increases,

$$h_m(\lambda) \ge c_0 \lambda \tag{B.3}$$

for a constant $c_0 > 0$ independent of m.

Then the following statements are equivalent:

(i) For every compact $K \subset (0, \infty)$,

$$\sup_{c \in K} \left| \frac{A_m(c/T)}{a_m} - c^\tau \right| \longrightarrow 0. \tag{B.4}$$

(ii) For every compact $K \subset (0, \infty)$,

$$\sup_{c \in K} \left| \frac{\int_0^\infty e^{-2cTh_m(\lambda)} \, \mathrm{d}A_m(\lambda)}{a_m} - \Gamma(\tau + 1)(2c)^{-\tau} \right| \longrightarrow 0. \tag{B.5}$$

Proof. Set $B_m(y) := A_m(y/T)/a_m$. Then

$$B_m(y) \le C \max\{y^{\tau-\epsilon}, y^{\tau+\epsilon}\}, \qquad y > 0. \tag{B.6}$$

Assume first Eq. (B.4). Then $B_m(y) \to y^\tau$ locally uniformly on $(0, \infty)$. Uniformly for c in a fixed compact subset of $(0, \infty)$, Eq. (B.2) gives $e^{-2cTh_m(y/T)} \to e^{-2cy}$ on compact y-intervals. The $(0, \delta]$ contribution is uniformly $O(\delta^{\tau-\epsilon})$ by Eq. (B.6); moreover, Eq. (B.3) and integration by parts give, uniformly for $c \ge c_- > 0$,

$$\begin{aligned} \int_{(A,\infty)} e^{-2cTh_m(y/T)} \, \mathrm{d}B_m(y) &\le \int_{(A,\infty)} e^{-2c_- c_0 y} \, \mathrm{d}B_m(y) \\ &\le 2c_- c_0 \int_A^\infty e^{-2c_- c_0 y} B_m(y) \, \mathrm{d}y, \end{aligned}$$

which vanishes uniformly in m as $A \to \infty$. Passing to the limit on $[\delta, A]$, then sending $\delta \downarrow 0$ and $A \to \infty$, gives locally uniformly in $c > 0$

$$\frac{\int_0^\infty e^{-2cTh_m(\lambda)} \, \mathrm{d}A_m(\lambda)}{a_m} \longrightarrow \int_0^\infty e^{-2cy} \tau y^{\tau-1} \, \mathrm{d}y = \Gamma(\tau + 1)(2c)^{-\tau},$$

which proves Eq. (B.5).

Conversely, assume Eq. (B.5). The envelope and Helly's theorem give, from every subsequence, one on which B_m converges at continuity points to a nondecreasing $B_\star$. The same endpoint estimates and Eq. (B.2) give, for every $c > 0$,

$$\int_0^\infty e^{-2cy} \, \mathrm{d}B_\star(y) = \Gamma(\tau + 1)(2c)^{-\tau}.$$

Uniqueness of Laplace–Stieltjes transforms gives $B_\star(y) = y^\tau$, so $B_m(c) \to c^\tau$ for $c > 0$. Monotonicity and continuity of c^τ make convergence uniform on compact subsets, proving Eq. (B.4). □

Remark B.2. The function $A_m(x)$ is the cumulative weighted spectral mass below x. Its monotonicity and right-continuity mean that it represents a finite positive measure, so contributions from different spectral modes cannot cancel. The joint limit examines the moving scale $\lambda \asymp T^{-1}$, and $a_m = T^{-\tau} L(T)$ is the expected amount of mass at that scale.

The *Uniform spectral control* assumption ensures that mass outside this moving window cannot remain hidden from the cutoff asymptotic while still affecting the training response. For example, a group of spectral mass may be located at $\lambda_m = y_m/T$, where $y_m \to \infty$. It eventually lies above every fixed cutoff c/T. The envelope prevents such moving mass from becoming large enough to change the filtered response.

The *Compatible training filter* assumption treats h_m as the effective decay rate of a spectral mode. Its local condition makes the training filter behave like e^{-2cy} for $\lambda = y/T$, while its lower bound suppresses faster modes away from the cutoff. Together, these assumptions ensure that the long-time response is determined by spectral mass near T^{-1}.

B.3 Spectral criteria for forcing and memory

This subsection proves the forcing and memory criteria stated in Theorem A.4. After specifying their common constant-schedule limit, we derive both by applying Theorem B.1 to the corresponding weighted spectral measure.

The proof has two parallel applications of Theorem B.1: one to learnable forcing and the other to one-injection memory. Both use the same constant-schedule limit and effective training filter. Apply Assumption A.3 with the component's weighted spectral mass and normalization. Part *(a)* of Assumption A.2 verifies the *Compatible training filter* condition below, while the spectral-window envelope is the *Uniform spectral control* condition in Eq. (B.1). The two applications differ only in their weighted spectral measure and normalization.

Proof of Theorem A.4. **Learnable forcing.**

For the conditioned empirical spectrum, set $A_m(x) := \nu^{\mathcal{F}}_{\boldsymbol{W}}((0,x])$. For every positive empirical eigenvalue, set $h_\eta(\lambda) := -\log q_\eta(\lambda)/(2\eta)$. With $x = \eta\lambda$, stability gives

$$1 - q_\eta(\lambda) = x\left[2 - \left(1 + \frac{1}{B}\right)x\right] \geq \delta x.$$

Since $-\log q \geq 1 - q$, $h_\eta(\lambda) \geq \delta\lambda/2$ on the positive support. Since $T = \eta t$, uniformly for $0 < y \leq A$,

$$Th_\eta(y/T) = -\frac{t}{2}\log\left[1 - \frac{2y}{t} + \left(1 + \frac{1}{B}\right)\frac{y^2}{t^2}\right] = y + O\left(\frac{A^2}{t}\right).$$

Hence h_η satisfies Eqs. (B.2) and (B.3). With $T = \eta t$,

$$F_{\boldsymbol{W},>0}(t) = \sum_{\widehat{\lambda}_j > 0} \left|\left\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star \right\rangle\right|^2 e^{-2Th_\eta(\widehat{\lambda}_j)} = \int_0^\infty e^{-2Th_\eta(\lambda)}\,\mathrm{d}A_m(\lambda).$$

Apply Theorem B.1 with $a_m = T^{-q_{\mathcal{F}}}L_{\mathcal{F}}(T)$ to obtain the stated cutoff and forcing asymptotics, uniformly under constant-factor changes of T.

One-injection memory.

For the conditioned empirical spectrum, set $A_m(x) := \nu^{\mathcal{K}}_{\boldsymbol{W}}((0,x])$. With the same map h_η used in the forcing proof, the exact conditional kernel becomes

$$\frac{B}{\eta^2}K_{\boldsymbol{W}}(t) = \int_0^\infty e^{-2Th_\eta(\lambda)}\,\mathrm{d}A_m(\lambda).$$

The filter verification in the forcing proof is unchanged. Apply Theorem B.1 with $a_m = 2^{q_{\mathcal{K}}}T^{-q_{\mathcal{K}}}L_{\mathcal{K}}(T)/\Gamma(q_{\mathcal{K}}+1)$ to obtain the claimed equivalence and constant-factor uniformity. □

B.4 Accumulating memory over training

Proof of Theorem A.6. For the same constant schedule and every $\lambda > 0$, $1 - q_\eta(\lambda) = \eta\lambda[2 - (1 + 1/B)\eta\lambda]$. The quadratic identity

$$q_\eta(\lambda) = \frac{1}{B+1} + \left(1 + \frac{1}{B}\right)\left(\eta\lambda - \frac{B}{B+1}\right)^2$$

and the stability margin show, on every positive empirical eigenvalue, that $0 < q_\eta(\lambda) < 1$ and

$$2 - \left(1 + \frac{1}{B}\right)\eta\lambda \geq \delta, \qquad q_\eta(\lambda)^t \leq e^{-t[1-q_\eta(\lambda)]} \leq e^{-\delta T\lambda}.$$

Substitution into the empirical kernel sum and exact geometric summation give the two identities in Eq. (A.22).

We first prove the cumulative-memory asymptotics directly from those identities. Retain $A_m^{\mathcal{K}}, a_m^{\mathcal{K}}$, and $\widetilde{A}_m^{\mathcal{K}}$ from the theorem. On the rescaled spectral support define the proof-local filters

$$\phi_{m,t}^{\mathrm{acc}}(y) := \frac{1-q_\eta(y/T)^t}{y[2-(1+1/B)y/t]}, \qquad \phi_{m,t}^{\mathrm{tail}}(y) := \frac{q_\eta(y/T)^t}{y[2-(1+1/B)y/t]}.$$

Because $T=\eta t$, the exact identities become

$$\frac{S_{\boldsymbol{W},t}}{(\eta/B)Ta_m^{\mathcal{K}}} = \int_{(0,\infty)} \phi_{m,t}^{\mathrm{acc}}(y)\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}}(y), \qquad \frac{S_{\boldsymbol{W},\infty}-S_{\boldsymbol{W},t}}{(\eta/B)Ta_m^{\mathcal{K}}} = \int_{(0,\infty)} \phi_{m,t}^{\mathrm{tail}}(y)\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}}(y). \quad \text{(B.7)}$$

On every compact interval $0<a\le y\le A<\infty$,

$$q_\eta(y/T)^t = \left[1-\frac{2y}{t}+\left(1+\frac{1}{B}\right)\frac{y^2}{t^2}\right]^t \longrightarrow e^{-2y}$$

uniformly. Hence

$$\phi_{m,t}^{\mathrm{acc}}(y) \longrightarrow \frac{1-e^{-2y}}{2y}, \qquad \phi_{m,t}^{\mathrm{tail}}(y) \longrightarrow \frac{e^{-2y}}{2y}$$

uniformly there. The cutoff asymptotic Eq. (A.18) makes the rescaled Stieltjes measures converge on the same compact interval to $\mathrm{d}(y^{q_{\mathcal{K}}})$.

It remains to make this compact limit uniform at zero and infinity. The preceding stability bounds and $1-q^t\le t(1-q)$ for $0\le q<1$ give

$$0\le \phi_{m,t}^{\mathrm{acc}}(y) \le \min\left\{1,\frac{1}{\delta y}\right\}, \qquad 0\le \phi_{m,t}^{\mathrm{tail}}(y)\le \frac{e^{-\delta y}}{\delta y}. \quad \text{(B.8)}$$

For $0<q_{\mathcal{K}}<1$, the endpoint envelope therefore gives

$$\int_{(0,a]} \phi_{m,t}^{\mathrm{acc}}\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}} \lesssim a^{q_{\mathcal{K}}-\epsilon}.$$

A dyadic decomposition at infinity gives

$$\int_{(A,\infty)} \phi_{m,t}^{\mathrm{acc}}\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}} \lesssim \sum_{j\ge 0}\frac{\widetilde{A}_m^{\mathcal{K}}(2^{j+1}A)}{2^jA} \lesssim A^{q_{\mathcal{K}}+\epsilon-1}\sum_{j\ge 0}2^{j(q_{\mathcal{K}}+\epsilon-1)},$$

which vanishes as $A\to\infty$ because $q_{\mathcal{K}}+\epsilon<1$.

For $q_{\mathcal{K}}>1$, Stieltjes integration by parts at zero instead gives

$$\int_{(0,a]} \phi_{m,t}^{\mathrm{tail}}\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}} \lesssim \int_{(0,a]} y^{-1}\,\mathrm{d}\widetilde{A}_m^{\mathcal{K}}(y) \lesssim a^{q_{\mathcal{K}}-\epsilon-1},$$

which vanishes because $q_{\mathcal{K}}-\epsilon>1$. At infinity, a dyadic decomposition combines the second bound in Eq. (B.8) with the polynomial endpoint envelope; its resulting exponentially damped series vanishes uniformly as $A\to\infty$.

We may now pass to the compact limit and then send $a\downarrow 0$ and $A\to\infty$. The limiting integrals are

$$\int_0^\infty \frac{1-e^{-2y}}{2y}\,\mathrm{d}(y^{q_{\mathcal{K}}}) = \frac{\Gamma(q_{\mathcal{K}}+1)}{2^{q_{\mathcal{K}}}(1-q_{\mathcal{K}})}, \qquad 0<q_{\mathcal{K}}<1,$$
$$\int_0^\infty \frac{e^{-2y}}{2y}\,\mathrm{d}(y^{q_{\mathcal{K}}}) = \frac{\Gamma(q_{\mathcal{K}}+1)}{2^{q_{\mathcal{K}}}(q_{\mathcal{K}}-1)}, \qquad q_{\mathcal{K}}>1.$$

Substituting the definition of $a_m^{\mathcal{K}}$ into Eq. (B.7) proves Eqs. (A.20) and (A.21) with their displayed η/B constants. The uniform convergence theorem for slowly varying functions makes this rescaling explicit: for c in a compact subset of $(0,\infty)$, put

$$t_c := \lfloor ct\rfloor, \qquad T_c := \eta t_c, \qquad a_{m,c}^{\mathcal{K}} := \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}}+1)}T_c^{-q_{\mathcal{K}}}L_{\mathcal{K}}(T_c).$$

Then $T_c/T \to c$, and $a^{\mathcal{K}}_{m,c}/a^{\mathcal{K}}_m \to c^{-q_{\mathcal{K}}}$ uniformly over such c. Repeating the preceding rescaling with $y = T_c\lambda$ gives the same compact limits and endpoint bounds uniformly in c. This proves the stated constant-factor uniformity.

For the direct terminal-tail criterion, use the separate cumulative measure A^{tail}_m defined in the theorem. Its transform under the forcing proof's h_η is exactly

$$S_{\boldsymbol{W},\infty} - S_{\boldsymbol{W},t} = \int_0^\infty e^{-2Th_\eta(\lambda)}\,\mathrm{d}A^{\mathrm{tail}}_m(\lambda).$$

The denominator in A^{tail}_m is at least δ, and the filter verification in the forcing proof is unchanged. Thus Theorem B.1, with $a_m = T^{-q_{\mathcal{S}}}L_{\mathcal{S}}(T)$, gives both locally uniform families. Apply its cutoff family at T/T_c and its transform family at T_c/T. Since $T_c/T \to c$ and $L_{\mathcal{S}}(T_c)/L_{\mathcal{S}}(T) \to 1$ locally uniformly in c, this is exactly the displayed direct equivalence.

Finally, replacing the empirical squared-spectrum measure by $\nu^{\mathcal{K}}_m(\mathrm{d}\lambda)$ gives the DE cumulative-memory argument, while replacing A^{tail}_m by $A^{\mathrm{tail}}_{m,\mathrm{DE}}$ gives the DE direct-tail argument. Every step uses the corresponding cutoff, endpoint-envelope, and stability hypotheses stated after the theorem. □

B.5 From forcing and memory asymptotics to observed loss under stable feedback

Proof of Theorem A.9. Stability makes the nonnegative repeated-feedback series summable:

$$\sum_{n\ge1}\|\mathsf{K}^{(*n)}\|_{\ell^1} = \sum_{n\ge1}\kappa_m^n = \frac{\kappa_m}{1-\kappa_m} \le \frac{\kappa}{1-\kappa}.$$

The clean solution is $\mathsf{F} + \sum_{n\ge1}\mathsf{K}^{(*n)} * \mathsf{F}$. Its constant-term response lies in $[\mathsf{F}_0, \mathsf{F}_0/(1-\kappa)]$.

In case (i), split at a fixed fraction of t: the long-tail assumptions control bounded lags, summability of the feedback kernels controls the middle part, and cumulative forcing controls the opposite endpoint. Thus

$$\mathsf{F}_{>0}(t,m) + \sum_{n\ge1}(\mathsf{K}^{(*n)} * \mathsf{F}_{>0})(t,m) \asymp \mathsf{F}_{>0}(t,m).$$

In case (ii), the two uniform subexponential relations applied successively to the convolution powers give

$$(\mathsf{K}^{(*n)} * \mathsf{F}_{>0})(t,m) \asymp \kappa_m^n\mathsf{F}_{>0}(t,m) + n\kappa_m^{n-1}\left[\sum_{s\ge0}\mathsf{F}_{>0}(s,m)\right]\mathsf{K}(t,m).$$

The assumed summable domination permits summation over n, yielding

$$\mathsf{F}_{>0}(t,m) + \sum_{n\ge1}(\mathsf{K}^{(*n)} * \mathsf{F}_{>0})(t,m) \asymp \mathsf{F}_{>0}(t,m) + \left[\sum_{s\ge0}\mathsf{F}_{>0}(s,m)\right]\mathsf{K}(t,m).$$

This proves the clean comparison in Eq. (A.23); positivity of the common Volterra system adds $\sigma^2\mathsf{S}_t$, proving its noisy comparison.

If $\mathsf{F}_{>0}(t,m) \to 0$, taking the limit in the Volterra equation and using $\lim_{t\to\infty}\mathsf{S}_t(m) = \kappa_m$ gives Eq. (A.24). Put $d_t := P_m - \mathsf{R}_\sigma(t,m)$. Subtracting the finite-time equation from its terminal fixed point gives

$$d_t = -\mathsf{F}_{>0}(t,m) + (\mathsf{K} * d)(t,m) + \big(P_m + \sigma^2\big)\,\overline{\mathsf{K}}_m(t).$$

The stability bound makes the positive resolvent G_m well defined. Iterating the last recursion gives the exact identity

$$P_m - \mathsf{R}_\sigma(t,m) = -(\mathsf{G}_m\mathsf{F}_{>0})(t) + (P_m+\sigma^2)(\mathsf{G}_m\overline{\mathsf{K}}_m)(t),$$

which is precisely Eq. (A.25) after multiplication by -1. Moreover, $P_m + \sigma^2 = (\mathsf{F}_0(m) + \sigma^2)/(1-\kappa_m) \ge 0$. Together with positivity of G_m, this proves directly that the two terms on the right-hand side of Eq. (A.25) are nonnegative.

The forcing estimates already established above give, uniformly in the common window, $(\mathsf{G}_m\mathsf{F}_{>0})(t) \asymp \mathsf{F}_{>0}(t,m)$ in case *(i)*, whereas in case *(ii)*,

$$(\mathsf{G}_m\mathsf{F}_{>0})(t) \asymp \mathsf{F}_{>0}(t,m) + \left[\sum_{s\geq 0} \mathsf{F}_{>0}(s,m)\right] \mathsf{K}(t,m).$$

For the terminal response, positivity and Eq. (A.26) give the explicit two-sided comparison

$$\overline{\mathsf{K}}_m(t) \leq (\mathsf{G}_m\overline{\mathsf{K}}_m)(t) \leq (1 + C_{\mathrm{tail}})\overline{\mathsf{K}}_m(t).$$

Consequently, $(P_m+\sigma^2)(\mathsf{G}_m\overline{\mathsf{K}}_m)(t) \asymp (P_m+\sigma^2)[\kappa_m - \mathsf{S}_t(m)]$. Finally, the exact response identity and Eq. (A.27) imply that $|\mathsf{R}_\sigma(t,m) - P_m|$ is comparable to the sum of the two nonnegative terms in Eq. (A.25). Substitution of the preceding positive-response estimates proves Eq. (A.28) in case *(i)* and its stated additional renewal term in case *(ii)*. □

C JOINT LEARNING-RATE AND BATCH-SIZE SCHEDULES

This appendix develops the schedule-dependent dynamics used in the main text in five steps. First, Appendices C.1 and C.2 derive the exact conditional risk recursion for joint learning-rate and batch-size schedules and its time-inhomogeneous PLRF deterministic equivalent. Next, Appendices C.3 and C.4 state the conditions under which intrinsic time controls how errors in each spectral mode decay, then define the PLRF regimes and admissible schedule families. We then establish the empirical finite-bulk spectrum, kernel, and exact noisy–clean gap, followed by the deterministic-equivalent risk asymptotics in the long-memory, integrable-memory, and finite-bulk regimes; see Appendices C.5 to C.7. Appendix C.8 states the additional comparison conditions needed to transfer the DE total-risk orders to realized finite-width SGD. Finally, Appendix C.9 formulates the optimal ratio-control problem, proves the matching data and feature-compute rates, and constructs integer schedules that attain them. Each proof is placed with the result it establishes.

C.1 EXACT RISK DYNAMICS UNDER JOINT SCHEDULES

This subsection derives the exact conditional risk dynamics for arbitrary deterministic learning-rate and batch-size schedules. The SGD update in Eq. (C.1) gives the modal recursion in Eq. (C.2), which sums to the closed risk equation in Eq. (C.3).

Return to the deterministic schedule of Section 5, with integer $B_t \geq 1$ and $\eta_t \geq 0$. Intrinsic time controls signal learning and B_t/η_t controls noise per intrinsic-time increment, although exact dynamics retain the individual steps:

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t - \frac{\eta_t}{B_t}\sum_{i=1}^{B_t} \boldsymbol{W}^\top \boldsymbol{x}_t^i \left(f_{\boldsymbol{a}_t}(\boldsymbol{x}_t^i) - y_t^i\right). \tag{C.1}$$

Schedules are fixed before sampling. With (T_t, r_s) from Eq. (2.1), set $\Delta T_s := T_{s+1} - T_s = \eta_s$ with $1/r_s := 0$ when $\eta_s = 0$, so step s has injection weight $\eta_s^2/B_s = \Delta T_s/r_s$.

To track the joint schedule exactly, let

$$q_s(z) := 1 - 2\eta_s z + \left(1 + \frac{1}{B_s}\right)\eta_s^2 z^2, \qquad Q_{s,t}(z) := \prod_{u=s}^{t-1} q_u(z), \qquad Q_{t,t}(z) := 1.$$

With $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{u}}_j = \widehat{\lambda}_j\widehat{\boldsymbol{u}}_j$, $\boldsymbol{e}_t := \boldsymbol{\Lambda}^{1/2}(\boldsymbol{W}\boldsymbol{a}_t - \boldsymbol{\theta}^\star)$, and $\rho_j(t) := \langle \widehat{\boldsymbol{u}}_j, \boldsymbol{e}_t\rangle$, we can now derive the risk equation directly.

Conditional on the frozen random features, the joint update gives

$$\boldsymbol{e}_{t+1} = \boldsymbol{e}_t - \frac{\eta_t}{B_t}\sum_{i=1}^{B_t} \widehat{\boldsymbol{H}}\boldsymbol{z}_t^i \left(\langle \boldsymbol{z}_t^i, \boldsymbol{e}_t\rangle - \varepsilon_t^i\right).$$

For one fresh Gaussian covariate,

$$\mathbb{E}[z_j\langle \boldsymbol{z}, \boldsymbol{e}_t\rangle] = \rho_j(t), \qquad \mathbb{E}[z_j^2\langle \boldsymbol{z}, \boldsymbol{e}_t\rangle^2] = \|\boldsymbol{e}_t\|_2^2 + 2\rho_j^2(t).$$

The B_t diagonal terms contribute $B_t\|\boldsymbol{e}_t\|_2^2+2B_t\rho_j^2(t)$, the $B_t(B_t-1)$ ordered cross terms contribute $B_t(B_t-1)\rho_j^2(t)$, and the independent label noises contribute $B_t\sigma^2$. Their sum gives the exact modal recursion

$$\mathbb{E}_t[\rho_j^2(t+1)] = q_t(\widehat{\lambda}_j)\rho_j^2(t) + \frac{\eta_t^2}{B_t}\widehat{\lambda}_j^2\left(\|\boldsymbol{e}_t\|_2^2+\sigma^2\right), \tag{C.2}$$

and finite iteration gives

$$\mathbb{E}[\rho_j^2(t)] = Q_{0,t}(\widehat{\lambda}_j)\rho_j^2(0) + \sum_{s<t}\frac{\eta_s^2}{B_s}\widehat{\lambda}_j^2 Q_{s+1,t}(\widehat{\lambda}_j)\left(\mathbb{E}\|\boldsymbol{e}_s\|_2^2+\sigma^2\right).$$

Define the propagated initial error and the weight of one variance injection by

$$F_t := \sum_j \rho_j^2(0)Q_{0,t}(\widehat{\lambda}_j), \qquad K_{t,s} := \frac{\eta_s^2}{B_s}\sum_j \widehat{\lambda}_j^2 Q_{s+1,t}(\widehat{\lambda}_j).$$

Summing the finite iterate over j yields the exact conditional risk equation below; this finite-dimensional step needs no stability condition or asymptotic spectral input:

$$R_{\sigma,t} = F_t + \sum_{s=0}^{t-1} K_{t,s}\left(R_{\sigma,s}+\sigma^2\right), \qquad R_{\sigma,t} := \mathbb{E}\left[\|\boldsymbol{e}_t\|_2^2 \mid \boldsymbol{W}\right]. \tag{C.3}$$

The fixed-batch recursion and its finite iterate require no second calculation: set $B_s \equiv B$ in the preceding formulas, so that $q_s = q_{\eta_s}$ and $\eta_s^2/B_s = \eta_s^2/B$, and then sum over j. Cauchy's functional calculus with the exact resolvent $(\widehat{\boldsymbol{H}} - z\boldsymbol{I}_d)^{-1}$ gives the corresponding empirical spectral formulas.

C.2 Time-inhomogeneous PLRF deterministic equivalent

The positive representation in Lemma A.1 defines a deterministic time-inhomogeneous surrogate whose exact mode-decay factors $Q_{s,t}$ depend on the full factorization $(\boldsymbol{\eta}, \boldsymbol{B})$, not only on T_t.

For deterministic schedules $\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots)$ and $\boldsymbol{B} = (B_0, B_1, \ldots)$, with $\eta_s \geq 0$ and $B_s \in \mathbb{N}$, retain q_s and $Q_{s,t}$ from the exact joint recursion and define

$$\mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m) := \int_{[0,\infty)} Q_{0,t}(\lambda)\,\mu_m^{\mathcal{F}}(\mathrm{d}\lambda), \tag{C.4}$$

$$\mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t,s,m) := \frac{\eta_s^2}{B_s}\int_{[0,\infty)} \lambda^2 Q_{s+1,t}(\lambda)\,\mu_m^{\mathcal{K}}(\mathrm{d}\lambda), \qquad 0 \leq s < t. \tag{C.5}$$

The corresponding time-inhomogeneous PLRF resolvent-DE risk is specified by

$$\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m) = \mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m) + \sum_{s=0}^{t-1}\mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t,s,m)\left[\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(s,m)+\sigma^2\right]. \tag{C.6}$$

Proposition C.1 (Joint-schedule PLRF resolvent-DE recursion)**.** For any deterministic $(\boldsymbol{\eta}, \boldsymbol{B})$ with $\eta_s \geq 0$ and $B_s \in \mathbb{N}$, the DE recursion in Eq. (C.6) has a unique nonnegative solution at every finite horizon.

Let $\delta > 0$. If

$$\sup_{\substack{s\geq 0\\ \lambda\in\operatorname{supp}\mu_m^{\mathcal{K}}}} \eta_s\left(1+\frac{1}{B_s}\right)\lambda \leq 2-\delta, \tag{C.7}$$

then

$$\sup_{t\geq 1}\sum_{s<t}\mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t,s,m) \leq \frac{\operatorname{tr}\boldsymbol{\Lambda}}{\delta}\sup_s\frac{\eta_s}{B_s}. \tag{C.8}$$

Consequently, if the right-hand side is at most $\kappa < 1$, then

$$\sup_{t\geq 0}\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m) \leq \frac{(\boldsymbol{\theta}^\star)^\top\boldsymbol{\Lambda}\boldsymbol{\theta}^\star + \kappa\sigma^2}{1-\kappa}. \tag{C.9}$$

Proof. By Eq. (A.8), Eqs. (C.4) and (C.5) are the contour formulas obtained by inserting $Q_{0,t}$ and $Q_{s+1,t}$, without approximating growing-degree polynomials on the contour.

For every $\lambda \geq 0$, $q_s(\lambda) = (1-\eta_s\lambda)^2 + \eta_s^2\lambda^2/B_s \geq 0$. Thus all mode-decay factors, forcing values, and kernel entries are nonnegative; forward substitution gives the unique nonnegative finite-horizon solution.

Under Eq. (C.7),

$$1 - q_s(\lambda) = \eta_s\lambda\left[2 - \left(1 + \frac{1}{B_s}\right)\eta_s\lambda\right] \geq \delta\eta_s\lambda.$$

It also follows that $0 \leq q_s(\lambda) \leq 1$. Hence

$$\begin{aligned}\sum_{s<t}\frac{\eta_s^2}{B_s}\lambda^2 Q_{s+1,t}(\lambda) &\leq \frac{1}{\delta}\left(\sup_u \frac{\eta_u}{B_u}\right)\lambda\sum_{s<t}[1-q_s(\lambda)]Q_{s+1,t}(\lambda)\\ &= \frac{1}{\delta}\left(\sup_u \frac{\eta_u}{B_u}\right)\lambda[1-Q_{0,t}(\lambda)]\\ &\leq \frac{1}{\delta}\left(\sup_u \frac{\eta_u}{B_u}\right)\lambda.\end{aligned}$$

Integration against $\mu_m^{\mathcal{K}}$, using its first moment, proves Eq. (C.8). Stability also gives $\mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m) \leq \mu_m^{\mathcal{F}}([0,\infty)) = (\boldsymbol{\theta}^\star)^\top\boldsymbol{\Lambda}\boldsymbol{\theta}^\star$. Finite-horizon suprema in Eq. (C.6) prove Eq. (C.9). □

Setting $B_s \equiv B$ and then $\eta_s \equiv \eta$ in Eqs. (C.4) to (C.6) gives the fixed-batch specialization and then the constant-schedule convolution recursion in Eq. (A.13).

C.3 Stable schedules and spectral-mode decay

The exact finite-width and deterministic-equivalent risk results below require control of each update and of accumulated feedback. We first state these conditions and prove common bounds describing how errors in each spectral mode decay with elapsed intrinsic time.

Assumption C.2 (Joint-schedule stability)**.** A deterministic schedule satisfies joint-schedule stability if there are constants $\delta > 0$ and $\kappa < 1$, uniform over the width and horizon, such that:

(a) *Pointwise contraction.*

$$\sup_s \eta_s\left(1 + \frac{1}{B_s}\right)\lambda_{\max}(\widehat{\boldsymbol{H}}) \leq 2 - \delta.$$

(b) *Row stability.*

$$\sup_{t\geq 1}\sum_{s<t} K_{t,s} \leq \kappa.$$

For the DE recursion, pointwise contraction is imposed uniformly over $\lambda \in \operatorname{supp}\mu_m^{\mathcal{K}}$, and row stability replaces $K_{t,s}$ by $\mathcal{K}_{t,s}$.

Here a kernel row fixes the current time t: $\sum_{s<t} K_{t,s}$ adds the weights of all past risks entering the risk at t. Keeping every such total uniformly below one makes repeated feedback a contraction on bounded risk paths, so its amplification remains controlled.

Lemma C.3 (Empirical row-stability certificate)**.** Under part *(a)* of Assumption C.2,

$$\sup_{t\geq 1}\sum_{s<t} K_{t,s} \leq \frac{\operatorname{tr}(\widehat{\boldsymbol{H}})}{\delta}\sup_s \frac{\eta_s}{B_s}.$$

Consequently, row stability holds whenever

$$\frac{\operatorname{tr}(\widehat{\boldsymbol{H}})}{\delta}\sup_s \frac{\eta_s}{B_s} < 1. \tag{C.10}$$

Proof. For each empirical eigenvalue, the identity

$$\frac{(\eta_s^2/B_s)\widehat{\lambda}_j^2}{1-q_s(\widehat{\lambda}_j)} = \frac{\eta_s\widehat{\lambda}_j/B_s}{2-(1+1/B_s)\eta_s\widehat{\lambda}_j}$$

and pointwise stability bound the right-hand side by $\eta_s\widehat{\lambda}_j/(B_s\delta)$. Summing in s and using the telescoping identity for $[1-q_s(\widehat{\lambda}_j)]Q_{s+1,t}(\widehat{\lambda}_j)$, then summing in j, proves the stated bound. □

Assumption C.4 (Regular spectral-mode decay)**.** A deterministic schedule satisfies the required one-step regularity if either:

(a) $\sup_{s,z\in\mathrm{spec}(\widehat{\boldsymbol{H}})} \eta_s z < 1$;

(b) for some $B_{\max} < \infty$, $\sup_s B_s \le B_{\max}$.

For DE statements, condition *(a)* takes the supremum over $\operatorname{supp}\mu_m^{\mathcal{K}}$ instead of $\mathrm{spec}(\widehat{\boldsymbol{H}})$.

Lemma C.5 (Intrinsic-time decay of individual spectral modes)**.** Assume part *(a)* of Assumption C.2 and Assumption C.4. Then, for constants $0 < c < C < \infty$,

$$e^{-Cz(T_t-T_s)} \le Q_{s,t}(z) \le e^{-cz(T_t-T_s)}.$$

If $z(T_t - T_s) = O(1)$ and $\max_{s\le u<t}\eta_u/(T_t-T_s) \to 0$, then

$$Q_{s,t}(z) = \exp\{-2z(T_t-T_s)+o(1)\}.$$

Proof. The joint one-step factorization is $q_s(z) = (1-\eta_s z)^2 + \eta_s^2 z^2/B_s$, with $1-q_s(z) = \eta_s z[2-(1+1/B_s)\eta_s z]$. Pointwise stability gives $0 \le q_s(z) \le 1$. Under condition *(a)*, choose $\rho < 1$ such that $\eta_s z \le \rho$ uniformly; then $q_s(z) \ge (1-\rho)^2$. A uniform batch bound also keeps the minimum of this quadratic uniformly positive. Thus $-\log q_s(z)$ is comparable to $1-q_s(z)$, and summation proves the two exponential bounds. Taylor expansion gives

$$\log Q_{s,t}(z) = -2z(T_t-T_s) + O\left(z^2\sum_{u=s}^{t-1}\eta_u^2\right),$$

so the relative-mesh condition proves the critical-scale limit. □

Remark C.6 (Fixed-batch specialization)**.** For $B_t \equiv B$, the uniform batch bound applies and $q_s = q_{\eta_s} \ge 1/(B+1)$. Thus both spectral-mode decay conclusions hold without a separate joint log-regularity assumption.

Thus elapsed intrinsic time determines spectral-mode decay in the fine-mesh limit; at finite steps, the exact decay still depends on the individual steps.

C.4 PLRF REGIMES AND ADMISSIBLE SCHEDULE FAMILIES

We now specialize the preceding stability and spectral-mode decay requirements to the PLRF parameter regimes used below.

The eight open PLRF propagation subregimes below comprise three families: finite bulk (FB), long memory (LM), and integrable memory (IM). This taxonomy is PLRF-specific.

Definition D1 (Open PLRF propagation subregimes)**.** Write $p := 2\alpha+2\beta-1$ and assume $p > 0$. Set $q_{\mathcal{F}} := p/(2\alpha)$; when $\alpha > 1/4$, also write $q_{\mathcal{K}} := 2-1/(2\alpha)$. The eight open subregimes are

$$\begin{array}{c|lll} 0<\alpha<\frac{1}{4} & \mathrm{FB}_1 : q_{\mathcal{F}}<1, & \mathrm{FB}_2 : q_{\mathcal{F}}>1, & \\ \frac{1}{4}<\alpha<\frac{1}{2} & \mathrm{LM}_1 : q_{\mathcal{F}}<q_{\mathcal{K}}, & \mathrm{LM}_2 : q_{\mathcal{K}}<q_{\mathcal{F}}<1, & \mathrm{LM}_3 : q_{\mathcal{F}}>1, \\ \alpha>\frac{1}{2} & \mathrm{IM}_1 : q_{\mathcal{F}}<1, & \mathrm{IM}_2 : 1<q_{\mathcal{F}}<q_{\mathcal{K}}, & \mathrm{IM}_3 : q_{\mathcal{F}}>q_{\mathcal{K}}. \end{array}$$

Within subregime IM_1, the stability-limited hard branch is

$$\alpha > \frac{1}{2}, \qquad \frac{1}{2}-\alpha < \beta < 0, \qquad 0 < p < 2\alpha - 1.$$

The lower β bound is precisely finite target energy; feature distortion occurs in $\mathrm{IM}_{2,3}$. We exclude the propagation boundaries $\alpha \in \{1/4, 1/2\}$, $\beta \in \{1/2, \alpha\}$, and $2\alpha + 2\beta = 1$. The excluded optimizer crossover $\beta = 0$ splits the two IM_1 branches, while $\alpha = 1$ splits interior and source-saturated compute branches within $\mathrm{IM}_{2,3}$; neither is a propagation boundary.

Definition D2 (DE-admissible integer schedule family)**.** Fix a model, open subregime, $0 < C_0 < \infty$, and small $c_\star > 0$. A deterministic, sample-independent family $(\eta_s > 0, B_s \in \mathbb{N})$ is DE-admissible along width–horizon sequences satisfying $T_t \to \infty$ and $T_t \le C_0 m^{2\alpha}$ if it satisfies the DE versions of part *(a)* of Assumption C.2 and Assumption C.4 uniformly, its intrinsic mesh is uniformly bounded, and $\sup_s r_s^{-1} \le c_\star m^{2\alpha-1}$ or $c_\star$ for $0 < \alpha < 1/2$, $\alpha \ne 1/4$, or $\alpha > 1/2$, respectively. All class constants are budget-independent, with no finite upper ratio cap; see Lemma C.13 and Proposition C.17.

C.5 Finite-bulk empirical spectrum, kernel, and gap

When $\alpha < 1/4$, the squared spectral mass diverges in the infinite-spectrum limit, so a width-independent memory profile no longer applies. At finite width, the empirical spectrum is cut off at scale $m^{-2\alpha}$. The next result locates this spectral band; the following kernel and gap results show how it controls exact conditional SGD.

Lemma C.7 (Finite-bulk empirical spectral event)**.** Suppose $\lambda_j = j^{-2\alpha}$, $0 < \alpha < 1/4$, and $d/m \to c \in (1, \infty)$. There are constants $a_-, a_+, b_0, C_{\mathrm{tr}} > 0$ and events $\mathcal{E}_m$, depending only on $\boldsymbol{W}$, with $\mathbb{P}_{\boldsymbol{W}}(\mathcal{E}_m) \to 1$, such that on $\mathcal{E}_m$,

$$\mathrm{tr}(\widehat{\boldsymbol{H}}^2) \le C_{\mathrm{tr}} m^{1-4\alpha}, \qquad \#\left\{ j : a_- m^{-2\alpha} \le \widehat{\lambda}_j \le a_+ m^{-2\alpha} \right\} \ge b_0 m. \tag{C.11}$$

Proof. The m nonzero eigenvalues of $\widehat{\boldsymbol{H}}$ equal those of $\boldsymbol{W}^\top \boldsymbol{\Lambda} \boldsymbol{W}$. Choose $0 < \varepsilon < \min\{1, c-1\}$, set $J = \lfloor \varepsilon m \rfloor$, and restrict $(\boldsymbol{\Lambda}, \boldsymbol{W})$ to rows $J < i \le d$. Then $\boldsymbol{\Lambda}_{\mathrm{tail}} \asymp m^{-2\alpha} \boldsymbol{I}$, while Gaussian singular-value bounds give $\boldsymbol{W}_{\mathrm{tail}}^\top \boldsymbol{W}_{\mathrm{tail}} \asymp \boldsymbol{I}_m$ with probability tending to one. Hence the tail Gram matrix has all its eigenvalues in a fixed multiple of $m^{-2\alpha}$. The positive semidefinite head has rank at most J, so rank interlacing leaves at least $m - J \asymp m$ eigenvalues in the band in Eq. (C.11).

For the trace upper bound, separating equal and unequal Gaussian-column indices gives

$$\mathbb{E}\,\mathrm{tr}(\widehat{\boldsymbol{H}}^2) = \left(1 + \frac{1}{m}\right) \mathrm{tr}(\boldsymbol{\Lambda}^2) + \frac{(\mathrm{tr}\,\boldsymbol{\Lambda})^2}{m} \asymp m^{1-4\alpha}.$$

Gaussian Poincaré applied to this trace, whose gradient with respect to $\boldsymbol{W}$ is $4\boldsymbol{\Lambda}\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{\Lambda} \boldsymbol{W})$, yields

$$\mathrm{Var}\left[\mathrm{tr}(\widehat{\boldsymbol{H}}^2)\right] \le \frac{C}{m} \mathbb{E}\left[\|\boldsymbol{W}\|_{\mathrm{op}}^4 \,\mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{\Lambda}^2 \boldsymbol{W})\right] \le C \frac{\mathrm{tr}(\boldsymbol{\Lambda}^2)}{m} = o\big(m^{2-8\alpha}\big)\,.$$

Here we used Gaussian operator-norm moments, $d \asymp m$, and $\mathrm{tr}(\boldsymbol{\Lambda}^2) \asymp m^{1-4\alpha}$. Concentration and the spectral-band event give the claimed common event. □

Lemma C.8 (Finite-bulk kernel for subregimes FB_1 and FB_2)**.** Suppose $\lambda_j = j^{-2\alpha}$, $0 < \alpha < 1/4$, and $d/m \to c > 1$. Fix $0 < C_0 < \infty$. Up to the horizon t, assume common margins in part *(a)* of Assumption C.2 and in Assumption C.4, with the suprema in those conditions restricted to $s < t$. On the events $\mathcal{E}_m$ in Lemma C.7, the following holds simultaneously for every deterministic, sample-independent schedule satisfying these margins and every $0 \le s < t$ with $T_t \le C_0 m^{2\alpha}$:

$$K_{t,s} \asymp \frac{\Delta T_s}{r_s} m^{1-4\alpha}, \qquad \sup_{u \le t} \sum_{s<u} K_{u,s} \lesssim m^{1-4\alpha} T_t \sup_{s<t} \frac{1}{r_s}.$$

The first expression is zero at a zero step. Under $B_s \equiv B$ and $\sup_{s<t} \eta_s \lesssim m^{2\alpha-1}$, these statements reduce to

$$K_{t,s} \asymp \frac{\eta_s^2}{B} m^{1-4\alpha}, \qquad \sup_{u \le t} \sum_{s<u} K_{u,s} \lesssim m^{1-4\alpha} \left(\sup_{s<t} \frac{\eta_s}{B}\right) T_t = o(1), \tag{C.12}$$

whenever $T_t = o(m^{2\alpha})$. The result is target-independent and applies to both finite-bulk subregimes FB_1 and FB_2.

Proof. On $\mathcal{E}_m$, Eq. (C.11) supplies the trace upper bound and $\Theta(m)$ eigenvalues with $\widehat{\lambda} \asymp m^{-2\alpha}$. The upper bound uses $0 \le Q_{s+1,t} \le 1$. Apply the spectral-mode decay lemma after extending the finite-horizon schedule by zero steps. On this bulk band, its lower bound and $T_t - T_{s+1} \le C_0 m^{2\alpha}$ give $Q_{s+1,t}(\widehat{\lambda}) \gtrsim 1$, so these $\Theta(m)$ modes give the matching lower bound after multiplication by $\eta_s^2/B_s = \Delta T_s/r_s$. Finally,

$$\sum_{s<u} \frac{\Delta T_s}{r_s} \le T_u \sup_{s<t} \frac{1}{r_s},$$

which proves the joint row-mass estimate. Under the fixed-batch substitution, $\Delta T_s/r_s = \eta_s^2/B$ and

$$\frac{1}{B}\sum_{s<u} \eta_s^2 \le \left(\sup_{s<t} \frac{\eta_s}{B}\right) T_u.$$

This proves Eq. (C.12); the peak bound and strict window make the latter $o(1)$. □

Corollary C.9 (Exact finite-bulk noisy–clean gap)**.** Retain the model, source window, and common schedule margins of Lemma C.8. Fix $\sigma^2 > 0$, a target with finite population energy, and $\kappa \in (0,1)$. There is a sufficiently small $c_\star > 0$ such that, on the same $\boldsymbol{W}$-only events, simultaneously for every eligible schedule and finite horizon satisfying

$$T_t \le C_0 m^{2\alpha}, \qquad \sup_{s<t} \frac{1}{r_s} \le c_\star m^{2\alpha-1},$$

the exact conditional risks obey

$$R_{\sigma,t} - R_{0,t} \asymp \sigma^2 m^{1-4\alpha} \sum_{s<t} \frac{\eta_s^2}{B_s} = \sigma^2 m^{1-4\alpha} \int_0^{T_t} \frac{\mathrm{d}u}{r(u)}.$$

The comparison constants are independent of width, horizon, schedule, target, and σ^2; they may depend on the fixed α, c, C_0, κ and the stability and propagation margins. On the strict window $0 < T_t = o(m^{2\alpha})$,

$$R_{\sigma,t} - R_{0,t} = \sigma^2 \sum_{s<t} K_{t,s} \left[1 + O\left(\frac{T_t}{m^{2\alpha}}\right)\right]$$

uniformly over the same schedule class.

Proof. Fix one of the events in Lemma C.8 and write $Z_v := R_{\sigma,v} - R_{0,v}$. Its row bound gives, for $\kappa_t := \max_{v \le t} \sum_{s<v} K_{v,s}$,

$$\kappa_t \lesssim m^{1-4\alpha} T_t \sup_{s<t} \frac{1}{r_s} \lesssim c_\star \frac{T_t}{m^{2\alpha}} \le C c_\star C_0.$$

Choose $c_\star$ so that $C c_\star C_0 \le \kappa < 1$. Subtracting the clean and noisy exact recursions in Eq. (C.3) yields

$$Z_v = \sum_{s<v} K_{v,s}(\sigma^2 + Z_s).$$

Positivity and forward induction give $0 \le Z_v \le \sigma^2 \kappa_t/(1-\kappa_t)$ for every $v \le t$. Substitution at the terminal row therefore gives the pointwise sandwich

$$\sigma^2 \sum_{s<t} K_{t,s} \le Z_t \le \frac{\sigma^2}{1-\kappa_t} \sum_{s<t} K_{t,s}.$$

The two-sided kernel bound in Lemma C.8, together with $\eta_s^2/B_s = (T_{s+1} - T_s)/r_s$, proves the asserted order and integral identity. On the strict window, $\kappa_t = O(T_t/m^{2\alpha})$, so the same sandwich gives the relative refinement. All implications hold on the common $\boldsymbol{W}$-only event, proving the stated simultaneous high-probability conclusion. □

The result holds for both finite-bulk target subregimes because the clean forcing cancels from the gap. It therefore requires only finite target energy, not a canonical source profile.

A concrete schedule family. Fix $B_0 \in \mathbb{N}$, $\vartheta \geq 0$, and $c_\eta > 0$ with $c_\eta/B_0 \leq c_\star$, and set

$$\eta_s \equiv c_\eta m^{2\alpha-1}, \qquad B_s = \left\lceil B_0(1+T_s)^\vartheta \right\rceil.$$

On $\mathcal{E}_m$, $\lambda_{\max}(\widehat{\boldsymbol{H}}) \leq \sqrt{C_{\mathrm{tr}}} m^{1/2-2\alpha}$, so the pointwise and logarithmic margins hold for all sufficiently large m. The chosen c_η/B_0 also gives the required ratio peak. Since $B_s \asymp B_0(1+T_s)^\vartheta$ and $\eta_s \to 0$, cellwise comparison gives

$$\sum_{s<t} \frac{\eta_s^2}{B_s} \asymp \frac{c_\eta m^{2\alpha-1}}{B_0} \int_0^{T_t} (1+u)^{-\vartheta}\, \mathrm{d}u.$$

Consequently, uniformly as $T_t \to \infty$ with $T_t \leq C_0 m^{2\alpha}$,

$$R_{\sigma,t} - R_{0,t} \asymp \frac{\sigma^2 c_\eta}{B_0} m^{-2\alpha} \begin{cases} T_t^{1-\vartheta}, & 0 \leq \vartheta < 1, \\ \log T_t, & \vartheta = 1, \\ 1, & \vartheta > 1. \end{cases}$$

Boundary case. The case $\alpha = 1/4$ is excluded because $\mathrm{tr}(\boldsymbol{\Lambda}^2) \asymp \log m$, which introduces logarithmic corrections to the trace and kernel estimates.

C.6 Long-memory and integrable-memory deterministic-equivalent risk asymptotics

DE forcing and memory filters in the LM and IM regimes. We now establish the spectral estimates for the PLRF deterministic equivalent, beginning with forcing and then memory. Recall $q_{\mathcal{K}} = 2 - 1/(2\alpha)$, and put

$$h_\alpha(u) := (1+u)^{-q_{\mathcal{K}}}, \qquad A_{\alpha,m} := \begin{cases} m^{1-2\alpha}, & 1/4 < \alpha < 1/2, \\ 1, & \alpha > 1/2. \end{cases}$$

For $\alpha > 1/4$, DE spectral measures and spectral-mode decay bounds give the LM/IM memory weight h_α, with schedule dependence through intrinsic time and $\eta_s^2/B_s = \Delta T_s / r_s$. The class-uniform DE risk bound and explicit PLRF forcing asymptotic are Theorem C.14 and Corollary C.15.

Lemma C.10 (Direct low-source PLRF DE forcing filter)**.** Fix $0 < \alpha < 1/2$, $\beta < 1/2$, and $p = 2\alpha + 2\beta - 1 > 0$. Suppose $d/m \in [c_-, c_+] \Subset (1, \infty)$. For every fixed $0 < c_0 \leq c \leq c_1 < \infty$ and $0 < C_0 < \infty$,

$$\int e^{-cu\lambda}\, \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \asymp m^{-p} + (1+u)^{-p/(2\alpha)}, \qquad 0 \leq u \leq C_0 m^{2\alpha}, \tag{C.13}$$

uniformly in m, u, c.

Proof. Write $A_m(x) := \mu_m^{\mathcal{F}}([0,x])$, $u_x := \mathfrak{m}(-x)$, and $s_m := \mathrm{tr}(\boldsymbol{\Lambda})/m$. For $x > 0$, the fixed-point equation at $-x$ and the Stieltjes representation in Lemma A.1 give

$$1 = u_x + \frac{1}{m} \sum_{j=1}^{d} \frac{\lambda_j u_x}{\lambda_j u_x + x}, \qquad \frac{1}{\lambda_j u_x + x} = \int \frac{\mu_{j,m}(\mathrm{d}\lambda)}{\lambda + x}.$$

Since $s_m \lesssim m^{-2\alpha}$, the first identity implies $u_x \geq c > 0$ whenever $x \geq m^{-2\alpha}$. The second identity then gives

$$\mu_{j,m}([0,x]) \leq \frac{2x}{\lambda_j u_x + x} \lesssim \min\left\{1, \frac{x}{\lambda_j}\right\}.$$

For $m^{-2\alpha} \leq x \leq 1$, put $J = x^{-1/(2\alpha)} \leq m < d$. Because $\lambda_j(\theta_j^\star)^2 = j^{-2\alpha-2\beta}$, splitting at J yields

$$A_m(x) \lesssim x \sum_{j \leq J} j^{-2\beta} + \sum_{j > J} j^{-2\alpha-2\beta} \lesssim x^{p/(2\alpha)}.$$

For $x < m^{-2\alpha}$, monotonicity gives $A_m(x) \lesssim m^{-p}$; for $x > 1$, $A_m(x) \le \mu_m^{\mathcal{F}}([0,\infty)) \asymp 1 \le x^{p/(2\alpha)}$. Thus, for every $x > 0$,

$$A_m(x) \lesssim m^{-p} + x^{p/(2\alpha)}. \tag{C.14}$$

We next obtain a matching lower mass on each population scale. The moment identities in Eq. (A.6) imply

$$\mathbb{E}X_{j,m} = \lambda_j, \qquad \mathbb{E}X_{j,m}^2 = \lambda_j^2 + s_m\lambda_j,$$

for $X_{j,m} \sim \mu_{j,m}$. Since $s_m/\lambda_j \le s_m/\lambda_d \lesssim 1$, Paley–Zygmund followed by Markov gives constants $0 < b_- < b_+ < \infty$ and $b_0 > 0$, independent of j, m, such that

$$\mu_{j,m}([b_-\lambda_j, b_+\lambda_j]) \ge b_0, \qquad 1 \le j \le d. \tag{C.15}$$

Let $1 \le u \le C_0 m^{2\alpha}$ and $J_u := \min\{d, \lfloor u^{1/(2\alpha)} \rfloor\}$. Summing Eq. (C.15) over $\lceil J_u/2 \rceil \le j \le J_u$ gives, uniformly for $c \in [c_0, c_1]$,

$$\int e^{-cu\lambda}\, \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \gtrsim J_u^{-p}.$$

If $J_u < d$, then $J_u^{-p} \asymp u^{-p/(2\alpha)}$ and $m^{-p} \lesssim u^{-p/(2\alpha)}$; if $J_u = d$, then $J_u^{-p} \asymp m^{-p}$ and $u^{-p/(2\alpha)} \lesssim m^{-p}$. This proves the required lower bound for $u \ge 1$. For $0 \le u \le 1$, the upper bound follows from the total target mass, while the $j = 1$ instance of Eq. (C.15) gives a uniform positive lower bound.

For the remaining upper bound with $u \ge 1$, decompose the positive axis into $[0, u^{-1}]$ and $(2^k u^{-1}, 2^{k+1}u^{-1}]$, $k \ge 0$. By Eq. (C.14),

$$\int e^{-cu\lambda}\, \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \le A_m(u^{-1}) + \sum_{k\ge 0} e^{-c_0 2^k} A_m(2^{k+1}u^{-1}) \lesssim m^{-p} + u^{-p/(2\alpha)}.$$

Combining the two bounds proves Eq. (C.13), including the cutoff endpoint; its constants may depend on C_0 and the fixed aspect-ratio interval. □

For $p = 2\alpha + 2\beta - 1$, define the appendix forcing template

$$\Phi_{\alpha,\beta}(u,m) := \begin{cases} m^{-p} + (1+u)^{-p/(2\alpha)}, & \mathrm{FB}_1, \mathrm{LM}_1, \mathrm{LM}_2, \mathrm{IM}_1, \\ m^{-2\alpha} + (1+u)^{-p/(2\alpha)}, & \mathrm{FB}_2, \mathrm{LM}_3, \\ m^{-2\alpha} + (1+u)^{-p/(2\alpha)} + m^{-1}(1+u)^{-1+1/(2\alpha)}, & \mathrm{IM}_2, \mathrm{IM}_3. \end{cases} \tag{C.16}$$

The main text records the large-clock order of this bounded template as $\mathcal{F}(T,m)$; the uniform statements use the $(1+u)$-regularized representative above.

Lemma C.11 (Uniform PLRF DE forcing filters in the proved source range)**.** Fix $\alpha > 0$, $\alpha \notin \{1/4, 1/2\}$, an admissible PLRF dimension sequence, and $0 < C_0 < \infty$. Suppose $p = 2\alpha+2\beta-1 > 0$, $\beta < 1 + 2\alpha$, and (α, β) lies off the PLRF critical lines. For every fixed $0 < c_- \le c \le c_+ < \infty$,

$$\int e^{-cu\lambda}\, \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \asymp \Phi_{\alpha,\beta}(u,m), \qquad 0 \le u \le C_0 m^{2\alpha}, \tag{C.17}$$

The constants may depend on the fixed branch, C_0, c_-, c_+, and the admissible dimension limit, but not on m or u.

Lemma C.12 (Uniform DE memory filters in the LM and IM regimes)**.** Fix $\alpha > 1/4$, $\alpha \ne 1/2$, an admissible PLRF dimension sequence, and $0 < C_0 < \infty$. For every fixed $0 < c_- \le c \le c_+ < \infty$,

$$\int \lambda^2 e^{-cu\lambda}\, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda) \asymp h_\alpha(u), \qquad 0 \le u \le C_0 m^{2\alpha}, \tag{C.18}$$

uniformly in m, u, c. Moreover,

$$\int_0^{C_0 m^{2\alpha}} h_\alpha(u)\, \mathrm{d}u \asymp A_{\alpha,m}. \tag{C.19}$$

The constants may depend on the fixed branch, C_0, c_-, c_+, and the admissible dimension limit, but not on m or u.

Proof of Lemmas C.11 and C.12. Step 1: a constant-schedule DE input. Retain the weighted memory measure $\nu_m^{\mathcal{K}}$ from Eq. (A.4). This memory measure and the forcing measure $\mu_m^{\mathcal{F}}$ are the resolvent-DE analogues of the full empirical measures $\nu_{\boldsymbol{W}}^{\mathcal{K}}$ and $\nu_{\boldsymbol{W}}^{\mathcal{F}}$ from Section 3. The fixed-point stability and far-field estimates of Paquette et al. (2024, Propositions E.1–E.2), together with their large-z expansion (Paquette et al., 2024, Proposition E.6), give a common finite enclosure

$$\Lambda_\star := \sup_m \sup\big(\operatorname{supp} \mu_m^{\mathcal{K}}\big) < \infty \tag{C.20}$$

along every admissible dimension sequence. For the constant reference schedule, set $B = 1$ and

$$\eta_m := \eta_\star \begin{cases} m^{2\alpha-1}, & 0 < \alpha < 1/2, \\ 1, & \alpha > 1/2, \end{cases}$$

where $\eta_\star > 0$ is small enough that, uniformly in m,

$$\sup_m \sup_{\lambda \in \operatorname{supp} \mu_m^{\mathcal{K}}} \eta_m \lambda < 1,$$

and the reference schedule lies a fixed distance below the pointwise and kernel-norm convergence thresholds used in Paquette et al. (2024, Appendices F–H). This choice works on both sides of $\alpha = 1/4$. Indeed, pointwise stability follows from Eq. (C.20), while at batch size $B = 1$

$$\sum_{r \ge 0} \mathscr{K}(r) = \frac{\eta_m}{2} \int \frac{\lambda}{1 - \eta_m \lambda} \, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda) \lesssim \eta_m \operatorname{tr} \mathbf{\Lambda} \lesssim \eta_\star.$$

The last bound uses $\operatorname{tr} \mathbf{\Lambda} \asymp m^{1-2\alpha}$ for $\alpha < 1/2$ and summability for $\alpha > 1/2$; choosing $\eta_\star$ sufficiently small gives a common kernel-norm margin. For this proof, abbreviate $q_m(\lambda) := 1 - 2\eta_m \lambda + 2\eta_m^2 \lambda^2$.

The precise parameter map to Paquette et al. (2024) is

$$d_{4+3} = m, \qquad v_{4+3} = d, \qquad B_{4+3} = 1, \qquad \gamma_{4+3} = \eta_m.$$

Equivalently, before setting $B_{4+3} = 1$, the averaged-batch convention used here satisfies $\gamma_{4+3} B_{4+3} = \eta_m$ and $\gamma_{4+3}^2 B_{4+3} = \eta_m^2 / B_{4+3}$; the second identity matches the kernel prefactor. For the memory lemma, where $\alpha > 1/4$, Eq. (A.8) identifies the complete constant-schedule DE kernel in their notation as

$$\mathscr{K}(r) = \eta_m^2 \int q_m(\lambda)^r \nu_m^{\mathcal{K}}(\mathrm{d}\lambda),$$

not merely its pure-point surrogate. Their complete-kernel comparison (Paquette et al., 2024, Proposition G.1) gives $\mathscr{K}(r) \asymp \mathscr{K}_{pp}(r)$ uniformly whenever $\eta_m r \le C_{\mathrm{win}} m^{2\alpha}$, for each fixed $C_{\mathrm{win}} < \infty$. Using their displayed definition of $\mathscr{K}_{pp}$, whose large-time asymptotic is recorded in Paquette et al. (2024, Proposition H.5), and dividing by η_m^2 yields

$$\int q_m(\lambda)^r \nu_m^{\mathcal{K}}(\mathrm{d}\lambda) \asymp \frac{1}{2\alpha} \int_0^1 x^{1-1/(2\alpha)} e^{-2\eta_m r x} \, \mathrm{d}x \asymp h_\alpha(\eta_m r). \tag{C.21}$$

The first comparison is the essential full-kernel input; the separate $\mathscr{K}_{pp}$ asymptotic alone would not imply it. For the second comparison, use the elementary fact

$$\int_0^1 x^a e^{-2\tau x} \, \mathrm{d}x \asymp (1 + \tau)^{-a-1}, \qquad a > -1, \quad \tau \ge 0. \tag{C.22}$$

Indeed, the integral is bounded above and below when $\tau \le 1$; when $\tau \ge 1$, the change of variables $y = \tau x$, followed by integration over $y \in [0, 1]$ for the lower bound and over $y \in [0, \infty)$ for the upper bound, proves the claim. Here $a = 1 - 1/(2\alpha) > -1$ is exactly $\alpha > 1/4$.

We record the forcing input at the same level of precision. In the sub-trace low-source range $0 < \alpha < 1/2$, $\beta < 1/2$, the desired Laplace comparison is already Lemma C.10; below we use the external contour input only for the remaining branches. The same parameter map and functional-calculus identity identify the complete forcing $\mathscr{F}(r)$ of Paquette et al. (2024) with $\int q_m(\lambda)^r \mu_m^{\mathcal{F}}(\mathrm{d}\lambda)$. Their zero-, cap-, and central-contour estimates (Paquette et al., 2024, Propositions F.1–F.3), combined in the complete-forcing comparison (Paquette et al., 2024, Corollary F.1), give

$$\int q_m(\lambda)^r \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \asymp \mathscr{F}_0(m) + \mathscr{F}_{pp}(\eta_m r) + \mathscr{F}_{ac}(\eta_m r, m) \tag{C.23}$$

on the same fixed-multiple cutoff window. Proposition F.3 of Paquette et al. (2024) is the step requiring $\beta < 1 + 2\alpha$: this condition makes the $u \log(1/u)$ remainder negligible. Accordingly, the forcing comparison below assumes $\beta < 1 + 2\alpha$. The gap filling there applies to the complete forcing: by the uniform pointwise margin above, $0 \le q_m \le 1$, so its positive measure representation makes the left-hand side of Eq. (C.23) nonincreasing in r.

The component asymptotics are recorded in Paquette et al. (2024, Propositions H.2–H.4); for completeness, we evaluate the corresponding integrals directly below. The pure-point integral and Eq. (C.22) give

$$\mathscr{F}_{pp}(\tau) = \frac{1}{2\alpha} \int_0^1 x^{(2\beta-1)/(2\alpha)} e^{-2\tau x} \, \mathrm{d}x \asymp (1+\tau)^{-p/(2\alpha)},$$

because $p > 0$. The zero-mode proposition gives

$$\mathscr{F}_0(m) \asymp \begin{cases} m^{-p}, & \beta < 1/2, \\ m^{-2\alpha}, & \beta > 1/2. \end{cases} \tag{C.24}$$

When $\beta < 1/2$, the absolutely-continuous component is absent. When $\beta > 1/2$ and $\alpha < 1/2$, its defining integral gives $\mathscr{F}_{ac} \lesssim m^{-2\alpha}$, so it is absorbed by Eq. (C.24). Finally, when $\alpha > 1/2$ and $\beta > 1/2$, put $b_\alpha := 1 - 1/(2\alpha) > 0$. For $0 \le \tau \le C_{\mathrm{win}} m^{2\alpha}$,

$$\mathscr{F}_{ac}(\tau, m) \asymp \frac{1}{m} \int_{m^{-2\alpha}}^1 x^{-1/(2\alpha)} e^{-2\tau x} \, \mathrm{d}x \asymp m^{-1} (1+\tau)^{-b_\alpha}.$$

For $\tau \le 1$, this follows by dropping the exponential up to fixed constants. For $\tau \ge 1$, the change of variables $y = \tau x$ writes the integral as

$$\tau^{-b_\alpha} \int_{\tau m^{-2\alpha}}^{\tau} y^{b_\alpha - 1} e^{-2y} \, \mathrm{d}y.$$

The last integral is bounded above by the complete gamma integral. For the lower bound, if $\tau m^{-2\alpha} \le 1/2$, integrate over $[1/2, 1]$. Otherwise its lower endpoint lies in $[1/2, C_{\mathrm{win}}]$, and, for all sufficiently large widths along the admissible sequence, the integration interval contains a fixed-length interval starting there. The integrand has a positive minimum on $[1/2, C_{\mathrm{win}} + 1]$, so the lower bound is uniform in the stated window. Thus the sum of the three components is uniformly comparable to $\Phi_{\alpha,\beta}(\tau, m)$ in the corresponding open branch. Outside the sub-trace low-source range, we have proved

$$\int q_m(\lambda)^r \mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \asymp \Phi_{\alpha,\beta}(\eta_m r, m), \qquad 0 \le \eta_m r \le C_{\mathrm{win}} m^{2\alpha}. \tag{C.25}$$

In particular, the cutoff endpoint is included. For the memory input, Proposition G.1 of Paquette et al. (2024) is uniform for $\gamma B r < M d^{2\alpha}$ for every fixed M; choosing $M > C_{\mathrm{win}}$ covers $\eta_m r = C_{\mathrm{win}} m^{2\alpha}$. For the forcing input, Corollary F.1 gives comparability for the complete forcing throughout the transition and uses its monotonicity to fill the intervening cutoff gaps.

Step 2: discrete-to-Laplace transfer. On the uniform stability interval, the ratio $-\log q_m(\lambda)/(\eta_m \lambda)$, with continuous value 2 at $\lambda = 0$, is bounded above and below by positive constants. Hence there are $0 < a_0 \le a_1 < \infty$, independent of m, r, and λ, such that, for every integer $r \ge 0$,

$$e^{-a_1 \eta_m r \lambda} \le q_m(\lambda)^r \le e^{-a_0 \eta_m r \lambda}.$$

In the sub-trace low-source range, integrating this sandwich and applying Lemma C.10 at the two fixed time rescalings proves Eq. (C.25) without any contour input.

For either $\nu = \nu_m^{\mathcal{K}}$ or $\nu = \mu_m^{\mathcal{F}}$, fix $c \in [c_-, c_+]$ and $u > 0$, and choose the proof-local integers

$$\ell_- := \left\lfloor \frac{cu}{a_1 \eta_m} \right\rfloor, \qquad \ell_+ := \left\lceil \frac{cu}{a_0 \eta_m} \right\rceil.$$

Then $a_1 \eta_m \ell_- \le cu \le a_0 \eta_m \ell_+$, and the preceding sandwich gives

$$q_m(\lambda)^{\ell_+} \le e^{-cu\lambda} \le q_m(\lambda)^{\ell_-}.$$

Positivity therefore gives

$$\int q_m(\lambda)^{\ell_+}\nu(\mathrm{d}\lambda) \le \int e^{-cu\lambda}\nu(\mathrm{d}\lambda) \le \int q_m(\lambda)^{\ell_-}\nu(\mathrm{d}\lambda).$$

We first handle a fixed small-time interval. By Eq. (A.7),

$$\nu_m^{\mathcal{K}}([0,\infty)) = \mathrm{tr}(\mathbf{\Lambda}^2) + \frac{(\mathrm{tr}\,\mathbf{\Lambda})^2}{m} \asymp 1, \qquad \mu_m^{\mathcal{F}}([0,\infty)) = \sum_{j=1}^{d} j^{-2(\alpha+\beta)} \asymp 1.$$

The first comparison uses $\alpha > 1/4$; under the additional forcing hypotheses, the second uses $p > 0$. Together with the common support bound Eq. (C.20), these identities show, for each fixed $u_0 < \infty$,

$$\int e^{-cu\lambda}\nu_m^{\mathcal{K}}(\mathrm{d}\lambda) \asymp 1 \asymp h_\alpha(u), \qquad \int e^{-cu\lambda}\mu_m^{\mathcal{F}}(\mathrm{d}\lambda) \asymp 1 \asymp \Phi_{\alpha,\beta}(u,m)$$

uniformly over $0 \le u \le u_0$ and $c \in [c_-, c_+]$.

Choose $u_0 \ge \max\{1, 2a_1\eta_\star/c_-\}$. For $u \ge u_0$, the definitions of ℓ_- and ℓ_+ give fixed positive constants b_-, b_+, depending only on a_0, a_1, c_-, c_+, and u_0, such that

$$b_- u \le \eta_m \ell_- \le \eta_m \ell_+ \le b_+ u.$$

Both h_α and every nonzero summand in $\Phi_{\alpha,\beta}$ are stable, uniformly, under a time rescaling in the fixed interval $[b_-, b_+]$. Consequently,

$$h_\alpha(\eta_m\ell_\pm) \asymp h_\alpha(u), \qquad \Phi_{\alpha,\beta}(\eta_m\ell_\pm, m) \asymp \Phi_{\alpha,\beta}(u,m).$$

It remains to check that rounding does not leave the source window. If $u \le C_0 m^{2\alpha}$, then

$$\eta_m\ell_+ \le \frac{c_+}{a_0} C_0 m^{2\alpha} + \eta_\star.$$

Choose once and for all $C_{\mathrm{win}} > (c_+/a_0)C_0 + \eta_\star$. Then both bracketing integers lie in the enlarged discrete window used in Eqs. (C.21) and (C.25). Applying those two inputs to the preceding integral bracket and clock comparison proves Eqs. (C.17) and (C.18) uniformly through $u = C_0 m^{2\alpha}$. Thus the fixed exponent rescaling and the ceiling error are absorbed before the cutoff endpoint is invoked.

Step 3: kernel mass. Since $q_{\mathcal{K}} = 2 - 1/(2\alpha)$, direct integration gives

$$\int_0^{C_0 m^{2\alpha}} h_\alpha(u)\,\mathrm{d}u \asymp \begin{cases} m^{1-2\alpha}, & 1/4 < \alpha < 1/2, \\ 1, & \alpha > 1/2, \end{cases} = A_{\alpha,m},$$

which is Eq. (C.19). The excluded boundary $\alpha = 1/2$ would instead give a logarithm. □

Lemma C.13 (Class-uniform nonconvolution Volterra reduction)**.** Fix the setting of Lemma C.12. Consider any deterministic schedule with $T_t \le C_0 m^{2\alpha}$, uniformly bounded intrinsic mesh, and satisfying the DE versions of part *(a)* of Assumption C.2 and Assumption C.4. Assume $A_{\alpha,m} \sup_s r_s^{-1} \le \varepsilon_0$. For the joint DE kernel, define

$$(\mathcal{K}^2)_{t,s} := \sum_{u=s+1}^{t-1} \mathcal{K}_{t,u}\mathcal{K}_{u,s}.$$

There is a constant $C_\star$, independent of the schedule, width, and horizon, such that

$$\sup_t \sum_{s<t} \mathcal{K}_{t,s} \le C_\star\varepsilon_0, \qquad (\mathcal{K}^2)_{t,s} \le C_\star\varepsilon_0\,\mathcal{K}_{t,s}. \tag{C.26}$$

Consequently, after choosing ε_0 so that $C_\star\varepsilon_0 < 1$, every nonnegative sequence v satisfies

$$v + \mathcal{K}v \le (I-\mathcal{K})^{-1}v \le v + \frac{1}{1-C_\star\varepsilon_0}\mathcal{K}v. \tag{C.27}$$

Proof. The one-step identity $q_s(\lambda) = (1-\eta_s\lambda)^2 + \eta_s^2\lambda^2/B_s$ and the stated margins give constants $0 < c < C < \infty$, common to the schedule class, for which

$$e^{-C\lambda(T_t-T_s)} \le Q_{s,t}(\lambda) \le e^{-c\lambda(T_t-T_s)}.$$

Thus Lemma C.12 yields, with $a_s := r_s^{-1}\Delta T_s = \eta_s^2/B_s$,

$$\mathcal{K}_{t,s} \asymp a_s h_\alpha(T_t - T_{s+1}). \tag{C.28}$$

Bounded mesh lets endpoint sums be compared with integrals. Hence

$$\sum_{s<t} \mathcal{K}_{t,s} \lesssim (\sup_s r_s^{-1}) \int_0^{T_t} h_\alpha(u)\,\mathrm{d}u \lesssim A_{\alpha,m} \sup_s r_s^{-1}.$$

For $V := T_t - T_{s+1}$, positivity and the same comparison give

$$\begin{aligned}(\mathcal{K}^2)_{t,s} &\lesssim a_s(\sup_\ell r_\ell^{-1}) \int_0^V h_\alpha(V-u)h_\alpha(u)\,\mathrm{d}u \\ &\lesssim a_s(\sup_\ell r_\ell^{-1}) A_{\alpha,m} h_\alpha(V) \\ &\lesssim A_{\alpha,m}(\sup_\ell r_\ell^{-1})\mathcal{K}_{t,s}.\end{aligned}$$

The middle inequality follows by splitting at $V/2$: on each half one factor is at most a fixed multiple of $h_\alpha(V)$, while the other integrates to at most $A_{\alpha,m}$. This proves Eq. (C.26).

Choose ε_0 so that $C_\star\varepsilon_0 < 1$. Positivity gives inductively $\mathcal{K}^n \le (C_\star\varepsilon_0)^{n-1}\mathcal{K}$ for $n \ge 1$. Summing the finite-horizon Neumann series proves Eq. (C.27). □

For the joint-schedule DE quantities in Proposition C.1, suppress the schedule arguments and write $\mathcal{F}_t$, $\mathcal{K}_{t,s}$, and $\mathcal{R}_{\sigma,t}$.

Theorem C.14 (Class-uniform joint-schedule risk bounds in the LM and IM PLRF regimes)**.** Fix $\alpha > 1/4$, $\alpha \ne 1/2$, an admissible PLRF dimension sequence, and a deterministic schedule family satisfying Lemma C.13, with its ratio-peak constant chosen so that $C_\star\varepsilon_0 < 1$. Then, uniformly over the schedule family, widths, horizons, and $0 \le s < t$,

$$\mathcal{K}_{t,s} \asymp \frac{\Delta T_s}{r_s}(1 + T_t - T_{s+1})^{-2+1/(2\alpha)}, \tag{C.29}$$

and

$$\mathcal{R}_{\sigma,t} \asymp \mathcal{F}_t + \sum_{s<t} \frac{\Delta T_s}{r_s}(1 + T_t - T_{s+1})^{-2+1/(2\alpha)}\big[\mathcal{F}_s + \sigma^2\big]. \tag{C.30}$$

The comparison constants may depend on the fixed model and common schedule-family constants, but not on the schedule, width, or horizon.

Proof. The DE recursion gives

$$\mathcal{R}_\sigma = (I - \mathcal{K})^{-1}\big(\mathcal{F} + \sigma^2\mathcal{K}\mathbf{1}\big).$$

Apply Eq. (C.27) directly with $v = \mathcal{F} + \sigma^2\mathcal{K}\mathbf{1}$. Positivity, together with $\mathcal{K}^2\mathbf{1} \le C_\star\varepsilon_0\mathcal{K}\mathbf{1}$, gives

$$\mathcal{F} + \mathcal{K}\mathcal{F} + \sigma^2\mathcal{K}\mathbf{1} \le \mathcal{R}_\sigma \lesssim \mathcal{F} + \mathcal{K}\mathcal{F} + \sigma^2\mathcal{K}\mathbf{1},$$

and inserting Eq. (C.28) proves both displays. □

Corollary C.15 (Explicit LM/IM PLRF DE forcing and risk asymptotics)**.** Under Theorem C.14, suppose additionally that $p = 2\alpha + 2\beta - 1 > 0$, $\beta < 1 + 2\alpha$, and (α, β) lies in one of the six open LM/IM subregimes. Then, uniformly over the same schedule family and source window,

$$\mathcal{F}_t \asymp \Phi_{\alpha,\beta}(T_t, m), \tag{C.31}$$

and hence

$$\mathcal{R}_{\sigma,t} \asymp \Phi_{\alpha,\beta}(T_t, m) + \sum_{s<t} \frac{\Delta T_s}{r_s}(1 + T_t - T_{s+1})^{-2+1/(2\alpha)}\big[\Phi_{\alpha,\beta}(T_s, m) + \sigma^2\big].$$

The comparison remains valid at $T_t \asymp m^{2\alpha}$.

Proof. The two-sided spectral-mode decay bounds in Lemma C.5 and Eq. (C.17) give Eq. (C.31); substitute this comparison in Eq. (C.30). □

C.7 FINITE-BULK DETERMINISTIC-EQUIVALENT RISK ASYMPTOTICS

Unlike the width-independent memory tails in the LM and IM regimes, FB memory is generated by a width-scale band supplied by the PLRF DE measure. Consequently, the kernel and the noisy–clean gap require no target-forcing assumption. Proposition C.18 proves the DE total-risk asymptotic for $\alpha < 1/4$ and $\beta < 1 + 2\alpha$ from Lemmas C.11 and C.16.

Lemma C.16 (Finite-bulk DE band mass)**.** Suppose

$$0 < \alpha < \frac{1}{2}, \qquad 1 < c_- \le \frac{d}{m} \le c_+ < \infty, \qquad \lambda_j = j^{-2\alpha}.$$

There exist constants $0 < a_- < a_+ < \infty$ and $0 < c < C < \infty$, depending only on α, c_-, c_+, such that

$$cm \le \mu_m^{\mathcal{K}}\big([a_- m^{-2\alpha}, a_+ m^{-2\alpha}]\big) \le Cm. \tag{C.32}$$

Proof. Each index $j \in \{m, \dots, d\}$ contributes a fixed positive amount of mass to the same width-scale band, and there are order m such indices. To show this, set $s_m := \mathrm{tr}(\mathbf{\Lambda})/m$. Since $d \asymp m$ and $\alpha < 1/2$, we have $s_m = m^{-1}\sum_{i=1}^{d} i^{-2\alpha} \le C_0 m^{-2\alpha}$. Let $X_{j,m} \sim \mu_{j,m}$, the j-th diagonal measure. By Eq. (A.6), $\mathbb{E}X_{j,m} = \lambda_j$ and $\mathbb{E}X_{j,m}^2 = \lambda_j^2 + s_m\lambda_j$. For $m \le j \le d$, $c_+^{-2\alpha} m^{-2\alpha} \le \lambda_j \le m^{-2\alpha}$, and therefore, uniformly over these indices, $\mathbb{E}X_{j,m}^2/(\mathbb{E}X_{j,m})^2 = 1 + s_m/\lambda_j \le C_1$. Thus the first two moments are uniformly comparable on these indices, and Paley–Zygmund yields

$$\mathbb{P}\left(X_{j,m} \ge \frac{1}{2}\lambda_j\right) \ge \frac{1}{4}\frac{\lambda_j^2}{\mathbb{E}X_{j,m}^2} \ge c_1 > 0.$$

A matching upper cutoff follows from Markov's inequality: for every $A > 0$, $\mathbb{P}(X_{j,m} > Am^{-2\alpha}) \le \lambda_j/(Am^{-2\alpha}) \le 1/A$. Choose A so that $A^{-1} \le c_1/2$, and set $a_- := c_+^{-2\alpha}/2$ and $a_+ := A$. Then

$$\mu_{j,m}\big([a_- m^{-2\alpha}, a_+ m^{-2\alpha}]\big) \ge \frac{c_1}{2}, \qquad m \le j \le d.$$

Thus every selected coordinate contributes at least $c_1/2$ mass to the same band. Summing over the at least $(c_- - 1)m$ indices $j = m, \dots, d$ gives the lower bound in Eq. (C.32); the total mass $\mu_m^{\mathcal{K}}([0,\infty)) = d \le c_+ m$ gives the upper bound. □

Proposition C.17 (Finite-bulk DE kernel and noisy–clean gap)**.** Fix $0 < \alpha < 1/4$ and an admissible PLRF dimension sequence with aspect ratios in a compact subset of $(1, \infty)$, and fix $0 < C_0 < \infty$. Suppose the schedule family satisfies the DE versions of part *(a)* of Assumption C.2 and Assumption C.4, has uniformly bounded intrinsic mesh, and

$$\sup_s \frac{1}{r_s} \le c_\star m^{2\alpha-1}$$

for a sufficiently small common $c_\star$, which may depend on C_0 and the common admissibility constants. Then, uniformly for $0 \le s < t$ with $T_t \le C_0 m^{2\alpha}$,

$$\mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t, s, m) \asymp m^{1-4\alpha}\frac{\Delta T_s}{r_s}. \tag{C.33}$$

Moreover, after decreasing $c_\star$ if necessary, there is one $\rho < 1$, common to the schedule family, such that

$$(\mathcal{K}^2)_{t,s} \le \rho \mathcal{K}_{t,s}. \tag{C.34}$$

For every fixed $\sigma^2 > 0$, on the same window,

$$\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t, m) - \mathcal{R}_{0,\boldsymbol{\eta},\boldsymbol{B}}(t, m) \asymp \sigma^2 m^{1-4\alpha}\sum_{s<t}\frac{\Delta T_s}{r_s}. \tag{C.35}$$

Proof. The two-sided spectral-mode decay bounds in Lemma C.5 and the DE band-mass bound in Lemma C.16 give the lower bound in Eq. (C.33): on the band $\lambda \asymp m^{-2\alpha}$, the age condition $T_t - T_{s+1} \le C_0 m^{2\alpha}$ keeps $Q_{s+1,t}(\lambda)$ bounded below by a positive constant. The upper bound follows from $Q_{s+1,t} \le 1$ and the DE moment identity

$$\int \lambda^2 \, \mu_m^{\mathcal{K}}(\mathrm{d}\lambda) \asymp m^{1-4\alpha}.$$

Consequently,

$$\sup_{u\le t}\sum_{s<u}\mathcal{K}_{u,s}\lesssim m^{1-4\alpha}T_t\sup_s r_s^{-1}.$$

This is at most a fixed constant smaller than one in FB_1, after choosing $c_\star$, and is $o(1)$ in the strict FB_2 window. The two-sided kernel comparison also gives, for $s<t$,

$$(\mathcal{K}^2)_{t,s}=\sum_{r=s+1}^{t-1}\mathcal{K}_{t,r}\mathcal{K}_{r,s}\lesssim\mathcal{K}_{t,s}m^{1-4\alpha}\sum_{r=s+1}^{t-1}\frac{\Delta T_r}{r_r}\le\rho\mathcal{K}_{t,s},$$

with one common $\rho<1$ after decreasing $c_\star$. Thus the same two-term Neumann reduction as in Eq. (C.27) applies.

Set $\mathcal{Z}_t:=\mathcal{R}_{\sigma,t}-\mathcal{R}_{0,t}$. Subtracting the two DE recursions gives $\mathcal{Z}=(I-\mathcal{K})^{-1}\sigma^2\mathcal{K}\mathbf{1}$. Positivity and Eq. (C.34) therefore imply

$$\sigma^2\mathcal{K}\mathbf{1}\le\mathcal{Z}\le\frac{\sigma^2}{1-\rho}\mathcal{K}\mathbf{1}.$$

Summing Eq. (C.33) over $s<t$ proves Eq. (C.35). □

The canonical PLRF target forcing is also controlled throughout the source range in which the central-contour remainder is negligible. This closes the total-risk asymptotic without an additional finite-bulk forcing assumption.

Proposition C.18 (Finite-bulk joint-schedule DE risk asymptotics in the proved source range)**.** Fix an open finite-bulk subregime, $p=2\alpha+2\beta-1>0$, and $\beta<1+2\alpha$, away from the PLRF critical lines. Retain the dimension-sequence and schedule-family hypotheses of Proposition C.17. For fixed $\sigma^2>0$, uniformly for $1\ll T_t\le C_0m^{2\alpha}$ in FB_1,

$$\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m)\asymp m^{-p}+(1+T_t)^{-p/(2\alpha)}+\sigma^2m^{1-4\alpha}\sum_{s<t}\frac{\Delta T_s}{r_s}.\tag{C.36}$$

Uniformly for $1\ll T_t=o(m^{2\alpha})$ in FB_2,

$$\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m)\asymp m^{-2\alpha}+(1+T_t)^{-p/(2\alpha)}+\sigma^2m^{1-4\alpha}\sum_{s<t}\frac{\Delta T_s}{r_s}.\tag{C.37}$$

All constants are uniform over the stated family.

Proof. The two-sided spectral-mode decay bounds in Lemma C.5 and Eq. (C.17) give, in the two finite-bulk branches,

$$\mathcal{F}_t\asymp\begin{cases}m^{-p}+(1+T_t)^{-p/(2\alpha)}, & \mathrm{FB}_1,\\ m^{-2\alpha}+(1+T_t)^{-p/(2\alpha)}, & \mathrm{FB}_2.\end{cases}$$

Write $q=p/(2\alpha)$. In FB_1, $0<q<1$, and

$$(\mathcal{K}\mathcal{F})_t\lesssim m^{1-4\alpha}(\sup_s r_s^{-1})\int_0^{T_t}\left[m^{-p}+(1+u)^{-q}\right]\mathrm{d}u\lesssim m^{-p}+(1+T_t)^{-q}\asymp\mathcal{F}_t.$$

In FB_2, $q>1$; using $\int_0^\infty(1+u)^{-q}\,\mathrm{d}u<\infty$ gives

$$(\mathcal{K}\mathcal{F})_t\lesssim m^{-2\alpha}\left[m^{-2\alpha}T_t+1\right]\lesssim m^{-2\alpha}\lesssim\mathcal{F}_t$$

on the strict window. Positivity and the two-term Neumann reduction from Eq. (C.34) now give

$$\mathcal{F}_t\le\mathcal{R}_{0,t}\le\mathcal{F}_t+\frac{1}{1-\rho}(\mathcal{K}\mathcal{F})_t\lesssim\mathcal{F}_t.$$

Adding the noisy–clean gap in Eq. (C.35) proves Eqs. (C.36) and (C.37). □

Assumption C.19 (High-source FB_2 DE forcing closure)**.** Fix $0<\alpha<1/4$, $\beta>1+2\alpha$, and aspect ratios in a compact subset of $(1,\infty)$. Uniformly over the finite-bulk admissible schedule family,

$$\mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m)\asymp m^{-2\alpha}+(1+T_t)^{-p/(2\alpha)}$$

on the strict FB_2 window $T_t=o(m^{2\alpha})$. The boundary $\beta=1+2\alpha$ is excluded; logarithmic corrections may occur there.

Corollary C.20 (Conditional high-source FB_2 extension)**.** Fix $0 < \alpha < 1/4$, $\beta > 1 + 2\alpha$, and $\sigma^2 > 0$, and let a schedule family satisfy the hypotheses of Proposition C.17. Under the residual forcing input in Assumption C.19, Eq. (C.37) holds uniformly for $1 \ll T_t = o(m^{2\alpha})$. The boundary $\beta = 1 + 2\alpha$ is not included in this pure-power statement.

Proof. The residual assumption supplies the FB_2 forcing comparison. The same bound $\mathcal{K}\mathcal{F} \lesssim \mathcal{F}$, followed by the two-term Neumann reduction and the target-free noisy–clean gap, is identical to the FB_2 part of the preceding proof. □

C.8 Conditional transfer from DE risk asymptotics to realized SGD

The preceding risk asymptotics concern the deterministic equivalent. Comparing their total-risk orders, or the resource infima below, with realized SGD requires the empirical-to-DE conditions stated here.

For the transfer statements below, suppress the fixed schedule and width by writing

$$F_t^{\boldsymbol{W}} := F_t, \qquad K_{t,s}^{\boldsymbol{W}} := K_{t,s}, \qquad \mathcal{F}_t := \mathcal{F}_{\boldsymbol{\eta},\boldsymbol{B}}(t,m), \qquad \mathcal{K}_{t,s} := \mathcal{K}_{\boldsymbol{\eta},\boldsymbol{B}}(t,s,m),$$

and set

$$R_\sigma^{\boldsymbol{W}} := (R_{\sigma,t})_{t\ge 0}, \qquad \mathcal{R}_\sigma := \big(\mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m)\big)_{t\ge 0}.$$

For a deterministic schedule, *fixed-sequence transfer* means that, for each prescribed width–horizon–schedule sequence, there are events of probability tending to one and constants $0 < c < C < \infty$, possibly depending on that sequence but not on m, s, t, such that

$$c\mathcal{F}_t \le F_t^{\boldsymbol{W}} \le C\mathcal{F}_t, \qquad c\mathcal{K}_{t,s} \le K_{t,s}^{\boldsymbol{W}} \le C\mathcal{K}_{t,s}$$

uniformly on its source window. On the same events, for one common $\rho < 1$,

$$(K^{\boldsymbol{W}})_{t,s}^2 \le \rho K_{t,s}^{\boldsymbol{W}}, \qquad \mathcal{K}_{t,s}^2 \le \rho\mathcal{K}_{t,s}.$$

A *class-uniform transfer* means that the PLRF operators are coupled triangularly across resource levels and that there are events E_N, with $\mathbb{P}(E_N) \to 1$, and common c, C, ρ such that, on E_N, the comparisons hold simultaneously for every admissible width, horizon, and schedule whose data or feature-compute budget is at most N.

Assumption A1 (Conditional empirical-to-DE transfer)**.** Assume fixed-sequence transfer for each prescribed width–horizon–schedule sequence under consideration. Whenever comparisons are made simultaneously over an admissible resource-bounded family, assume the class-uniform transfer property instead.

Proposition C.21 (Conditional transfer from joint DE to realized SGD)**.** Under fixed-sequence transfer, for every fixed $\sigma^2 \ge 0$,

$$R_{\sigma,t} \asymp \mathcal{R}_{\sigma,\boldsymbol{\eta},\boldsymbol{B}}(t,m)$$

uniformly along the prescribed sequence. Under class-uniform transfer, the same comparison is uniform over the admissible family; consequently the exact conditional and DE infima at fixed data or feature compute are comparable.

Proof. Let K denote either the exact or the DE positive kernel. The bound $K^2 \le \rho K$ gives, for $v \ge 0$,

$$v + Kv \le (I - K)^{-1}v \le v + \frac{1}{1-\rho}Kv.$$

Applied to forcing and label-noise inputs, componentwise comparison gives

$$R_\sigma^{\boldsymbol{W}} \asymp F^{\boldsymbol{W}} + K^{\boldsymbol{W}}(F^{\boldsymbol{W}} + \sigma^2\mathbf{1}) \asymp \mathcal{F} + \mathcal{K}(\mathcal{F} + \sigma^2\mathbf{1}) \asymp \mathcal{R}_\sigma.$$

Class uniformity permits taking infima on the same event. □

Finite-width Volterra hierarchy. Across the displayed LM, IM, and FB systems, true online SGD closely tracks both the exact conditional finite-W Volterra recursion and its resolvent-DE reduction; see Figures 6 and 7. At fixed $\boldsymbol{W}$, Eq. (C.3) gives the exact conditional population risk, and replacing its forcing and kernel by their resolvent deterministic equivalents gives the second reduction.

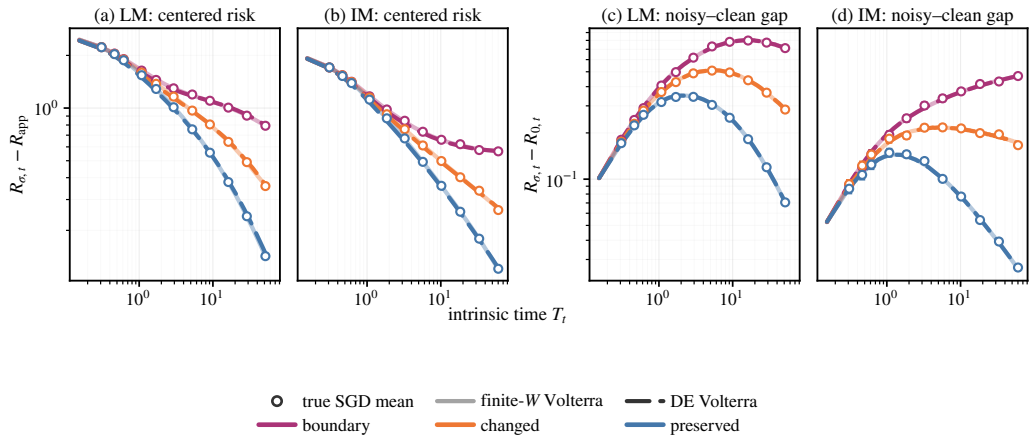


Figure 6: Online SGD, the exact finite-width recursion, and its DE reduction for centered risk and the noisy–clean gap in long and integrable memory. Their agreement supports both approximation steps and the predicted boundary, changed, and preserved schedule responses.

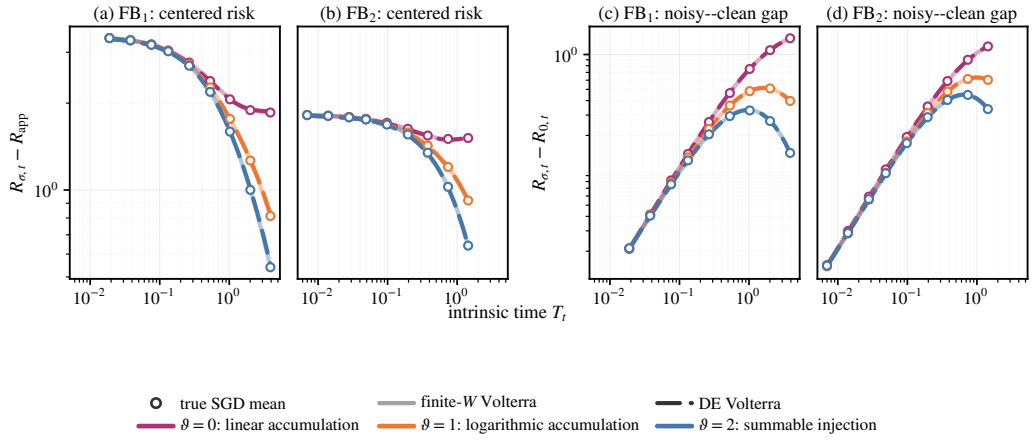


Figure 7: Online SGD, the exact finite-width recursion, and its DE reduction for the three injection responses in FB_1 and FB_2. Their agreement supports the finite-bulk Volterra hierarchy and accumulation laws.

C.9 Optimal ratio and schedule realizations

Represent the discrete ratio as the piecewise-constant intrinsic-time control $r(u) := r_s$ for $T_s \le u < T_{s+1}$, so the injection weights remain exact. If $\sup_s \eta_s \lesssim 1$ and $\sup_s r_s^{-1} \lesssim 1$, their continuum order is

$$\int_0^{T_t} \frac{1}{r(u)} \, (1 + T_t - u)^{-2+1/(2\alpha)} \left[\mathcal{F}(u, m) + \sigma^2\right] \, \mathrm{d}u, \tag{C.38}$$

with finite-bulk analogue $\int_0^{T_t} r(u)^{-1} \, \mathrm{d}u$. It also records exact sample count:

$$D_t := \sum_{s<t} B_s = \sum_{s<t} r_s \Delta T_s = \int_0^{T_t} r(u) \, \mathrm{d}u. \tag{C.39}$$

Use $\mathfrak{f} \asymp m D_t$ as the FLOP proxy.

Setting

$$B_s \equiv B \tag{C.40}$$

gives $T_t = \sum_{u<t} \eta_u$, $r_s = B/\eta_s$, and $\Delta T_s / r_s = \eta_s^2 / B$, recovering every fixed-B risk formula.

The identities $\eta_s^2/B_s = \Delta T_s/r_s$ and $B_s = r_s \Delta T_s$ are algebraic. Bounded-mesh endpoint quadrature, and its vanishing-mesh strengthening, prove respectively the comparison and convergence in Eq. (C.38); summing the second identity proves Eq. (C.39). Under Eq. (C.40), $q_s(z) = 1 - 2\eta_s z + (1 + 1/B)\eta_s^2 z^2$ and $\Delta T_s / r_s = \eta_s^2 / B$ recover every fixed-B coefficient.

Fix terminal horizon $T := T_t$ and budget $D := D_t$. For every LM or IM branch of Theorem C.14, define the surviving injection weight

$$w_{T,m}(u) := (1 + T - u)^{-2+1/(2\alpha)} \left[\mathcal{F}(u, m) + \sigma^2 \right], \qquad 0 \le u \le T. \tag{C.41}$$

The schedule-dependent loss and its budget constraint are then

$$\mathcal{J}_{T,m}(r) := \int_0^T \frac{w_{T,m}(u)}{r(u)} \, \mathrm{d}u, \qquad \int_0^T r(u) \, \mathrm{d}u = D. \tag{C.42}$$

The optimizer is proportional to $\sqrt{w_{T,m}}$, allocating more samples where the weighted noise contribution to the terminal risk is larger; see Theorem C.22. For the box-constrained form, use $\mathrm{clip}(x, a, b) := \min\{b, \max\{a, x\}\}$.

Theorem C.22 (Fixed-horizon continuum ratio control)**.** Suppose $w_{T,m}$ is integrable and positive almost everywhere and $0 < \int_0^T \sqrt{w_{T,m}(u)} \, \mathrm{d}u < \infty$. Among positive controls satisfying the data constraint in Eq. (C.42), the unique minimizer up to null sets and its minimum value are

$$r_T^\star(u) = \frac{D \sqrt{w_{T,m}(u)}}{\displaystyle\int_0^T \sqrt{w_{T,m}(v)} \, \mathrm{d}v}, \qquad \min_r \mathcal{J}_{T,m}(r) = \frac{1}{D} \left(\int_0^T \sqrt{w_{T,m}(u)} \, \mathrm{d}u \right)^2. \tag{C.43}$$

With pointwise bounds $0 < r_- < r_+ < \infty$, the admissible set is nonempty exactly when $r_- T \le D \le r_+ T$. At $D = r_\pm T$, the unique feasible control is $r_T^\star \equiv r_\pm$. Under strict feasibility the unique control is

$$r_T^\star(u) = \mathrm{clip}\left(\sqrt{\frac{w_{T,m}(u)}{\mu}}, r_-, r_+ \right), \qquad \int_0^T r_T^\star(u) \, \mathrm{d}u = D, \tag{C.44}$$

for a multiplier $\mu > 0$ chosen to satisfy the data constraint.

Proof of Theorem C.22. Write $w = w_{T,m}$. For every positive control with $\int_0^T r = D$, Cauchy–Schwarz gives

$$\mathcal{J}_{T,m}(r) D = \left(\int_0^T \frac{w(u)}{r(u)} \, \mathrm{d}u \right) \left(\int_0^T r(u) \, \mathrm{d}u \right) \ge \left(\int_0^T \sqrt{w(u)} \, \mathrm{d}u \right)^2.$$

Equality gives Eq. (C.43); strict convexity of $x \mapsto w(u)/x$ almost everywhere gives uniqueness up to null sets.

For the bounded problem, $r_- T \le D \le r_+ T$ is necessary by integration and sufficient by $r \equiv D/T$. A multiplier $\mu > 0$ solves

$$\int_0^T \mathrm{clip}\left(\sqrt{\frac{w(u)}{\mu}}, r_-, r_+ \right) \mathrm{d}u = D.$$

Its left side is continuous and nonincreasing from $r_+ T$ to $r_- T$, so strict feasibility gives a solution. The clipped control in Eq. (C.44) pointwise minimizes $w(u)/x + \mu x$ on $[r_-, r_+]$, whence

$$\frac{w(u)}{r(u)} + \mu r(u) \ge \frac{w(u)}{r_T^\star(u)} + \mu r_T^\star(u).$$

Integration cancels the resource terms, proving global optimality, while strict convexity gives primal uniqueness; at either resource endpoint the box constraint forces the stated constant control. The multiplier is unique exactly when its resource equation is. Its left side is strictly decreasing through a solution iff the free set

$$\left\{ u : r_- < \sqrt{w(u)/\mu} < r_+ \right\}$$

has positive measure; otherwise it may have a plateau despite primal uniqueness. □

Because $\sqrt{\mathcal{F}(u,m)+\sigma^2}$ accounts for clean and label noise, the full optimizer need not be monotone. When fixed nonzero label noise leads, it reduces to

$$r_T^\star(u) = \mathrm{clip}\left(c(1+T-u)^{-1+1/(4\alpha)}, r_-, r_+\right), \tag{C.45}$$

where c enforces the data constraint. In $\mathrm{FB}_{1,2}$, instead,

$$\sigma^2 m^{1-4\alpha}\int_0^T \frac{\mathrm{d}u}{r(u)}, \qquad r_T^\star(u) \equiv \frac{D}{T}, \tag{C.46}$$

when feasible, with the same $\mathrm{FB}_{1,2}$ proved-source and high-source qualifications stated above.

Two continuum schedule realizations. The positive optimizer fixes r, not its learning-rate/batch factorization. Let τ denote continuum iteration time, so $\mathrm{d}u/\mathrm{d}\tau = \eta(\tau)$ and $r(u(\tau)) = B(\tau)/\eta(\tau)$. For a fixed learning rate η_0, set

$$\eta(\tau) \equiv \eta_0, \qquad u(\tau) = \eta_0\tau, \qquad B(\tau) = \eta_0 r_T^\star(\eta_0\tau), \qquad 0 \le \tau \le \frac{T}{\eta_0}. \tag{C.47}$$

This is the fixed-learning-rate realization of Wang et al. (2026). For a fixed continuum batch $B_0 > 0$, let $u(0) = 0$ solve

$$\frac{\mathrm{d}u}{\mathrm{d}\tau} = \frac{B_0}{r_T^\star(u(\tau))}, \qquad B(\tau) \equiv B_0, \qquad \eta(\tau) = \frac{B_0}{r_T^\star(u(\tau))}, \qquad 0 \le \tau \le \frac{D}{B_0}. \tag{C.48}$$

Both forms process D samples and reach intrinsic time T; the second is the fixed-batch realization of Bordelon & Mori (2026). If, on an unclipped terminal segment,

$$r_T^\star(u) = c(1+T-u)^{-1+1/(4\alpha)},$$

then its fixed-batch realization on the corresponding terminal iteration-time segment is

$$\eta(\tau) = \frac{B_0}{c}\left[1 + \frac{D - B_0\tau}{4\alpha c}\right]^{4\alpha-1}. \tag{C.49}$$

Thus the intrinsic square-root profile becomes iteration-time decay of exponent $4\alpha-1$, while clipping becomes learning-rate or batch plateaus. With $p = 2\alpha + 2\beta - 1$, the corresponding terminal-decay, peak, risk, and source exponents are, respectively,

$$4\alpha - 1, \qquad \frac{\beta}{\alpha+\beta}, \qquad \frac{p}{p+1}, \qquad \frac{p}{2\alpha}. \tag{C.50}$$

When fixed label noise leads, Eq. (C.45) coincides, after reparameterization by intrinsic time, with the joint learning-rate–batch-size optimizer of Bordelon & Mori (2026). Thus, when fixed label noise leads, the $\beta > 0$ branch recovers the power-decay shape and the peak and risk exponents of Bordelon & Mori (2026) and Li et al. (2026a). For $\beta < 0$, lower-ratio clipping becomes a maximum-learning-rate plateau followed by the same terminal power decay, recovering their WSD structure and hard-branch risk exponent. The easy–hard boundary is $\beta = 0$. These comparisons are theorem-level only in the trace-class regime $\alpha > 1/2$, away from the excluded boundary; below trace class the correspondence is algebraic. The constant-learning-rate factorization similarly matches Wang et al. (2026). Related work optimizes batch given learning rate (Li et al., 2026b) or proves their finite-sample equivalence (Meterez et al., 2026a); here ratio optimization follows after the spectral forcing and memory asymptotics of Theorem A.4.

From continuum control to integer SGD. Here *admissible* means the full class of Definition D2, with common constants fixed as budgets grow and no finite upper ratio cap in the scaling infima. Equal-sample quantiles discretize the continuum profile, using $B_s \equiv 1$ for every upper bound. Let $\mathcal{R}_\sigma^\star(D)$ be the terminal DE-risk infimum subject to $\sum_{s<t} B_s \le D$, and define $\mathcal{R}_\sigma^\star(\mathfrak{f})$ analogously under $m\sum_{s<t} B_s \le \mathfrak{f}$. Integer realization loses no exponent and no admissible schedule improves the orders, as proved below.

Table 4: Optimal data and feature-compute rates across the PLRF propagation regimes. Each entry is an $\asymp$-order for the deterministic-equivalent risk, with $p = 2\alpha + 2\beta - 1$.

Subregime and optimizer branch	$\mathcal{R}^\star_\sigma(D)$	$\mathcal{R}^\star_\sigma(\mathfrak{f})$
LM/IM orders: $\alpha > 1/4$		
$\mathrm{IM}_1,\ \beta < 0$	$D^{-p/(2\alpha)}$	$\mathfrak{f}^{-p/(1+2\alpha)}$
$\mathrm{LM}_1, \mathrm{LM}_2;\ \mathrm{IM}_1,\ \beta > 0$	$D^{-p/(1+p)}$	$\mathfrak{f}^{-p/(2+p)}$
LM_3	$D^{-p/(1+p)}$	$\mathfrak{f}^{-2\alpha p/[p+2\alpha(1+p)]}$
$\mathrm{IM}_{2,3},\ 1/2 < \alpha \le 1$	$D^{-p/(1+p)}$	$\mathfrak{f}^{-p/(1+p+2\beta)}$
$\mathrm{IM}_{2,3},\ \alpha > 1$	$D^{-p/(1+p)}$	$\mathfrak{f}^{-\alpha/(1+\alpha)}$
Finite-bulk orders: $0 < \alpha < 1/4$		
FB_1	$D^{-p/(1+p)}$	$\mathfrak{f}^{-p/(2+p)}$
FB_2	$D^{-2\alpha p/[p(1-2\alpha)+8\alpha^2]}$	$\mathfrak{f}^{-2\alpha p/[p(2-2\alpha)+8\alpha^2]}$

Corollary C.23 (Matching integer-schedule scaling across the PLRF propagation subregimes)**.** Fix $\sigma^2 > 0$ and retain the full class in Definition D2. In the explicit LM/IM source range of Corollary C.15, the DE optima at fixed data and feature compute have the LM/IM orders in Table 4. In the FB source range $\beta < 1 + 2\alpha$, the DE optima have the displayed FB orders. Every proved upper bound is attained, up to a constant, by an admissible deterministic schedule with $B_s \equiv 1$, and the corresponding matching lower bound holds for every admissible schedule. For high-source FB_2, $\beta > 1 + 2\alpha$, the displayed FB_2 order extends conditionally under Assumption C.19. The boundary $\beta = 1 + 2\alpha$ is excluded.

The matching orders are summarized in Table 4.

The subregime fixes the clean-risk and memory scalings, r allocates noise, and the compute budget determines the trade-off between width and data. Unit-batch schedules attain every proved order. At $\alpha = 1$, the two $\mathrm{IM}_{2,3}$ compute expressions agree at $\mathfrak{f}^{-1/2}$. For $\alpha > 1$, the optimizer reaches $T \asymp m^{2\alpha}$, a source-window saturation rather than a new subregime. The lower-ratio floor is active in IM_1 for $\beta < 0$; at fixed data its $\beta > 0$ and $\beta < 0$ branches have peaks

$$D^{\,1-\frac{1}{2(p+1)}} \quad \text{and} \quad D^{\,\frac{1}{2}(1+p/(2\alpha))},$$

respectively, so any cap must exceed the relevant order. At $\beta = 0$ the neighboring exponents agree, but open-branch statements exclude the crossover.

We now prove Corollary C.23 and the consequences in Appendix C.9: specialize the continuum solution, realize its profiles with integer $B_s = 1$ schedules, and match the DE lower objectives under data and FLOP budgets.

Subregime profiles. Substitution of Eq. (C.41) into the square-root optimizer gives, on the LM and IM branches,

$$r^\star_T(u) \propto (1 + T - u)^{-1+1/(4\alpha)} \sqrt{\mathcal{F}(u, m) + \sigma^2}.$$

Its fixed nonzero label-noise component yields Eq. (C.45). In FB_1 and FB_2,

$$\left(\int_0^T \frac{\mathrm{d}u}{r(u)}\right)\left(\int_0^T r(u)\,\mathrm{d}u\right) \ge T^2$$

instead gives the constant optimizer Eq. (C.46).

The next lemma turns the required ratio profiles into genuine SGD schedules; it is also the achievability step in Corollary C.23.

For integers $D \ge 1$, a horizon $T > 0$, and $\alpha > 1/4$, define

$$I_\alpha(T) := \int_0^T (1 + T - u)^{-1+1/(4\alpha)}\,\mathrm{d}u, \qquad \bar{r}(u) := \frac{D}{I_\alpha(T)}(1 + T - u)^{-1+1/(4\alpha)}.$$

Let $0 = u_0 < \cdots < u_D = T$ be the data-quantile partition determined by $\int_0^{u_s} \bar{r}(v)\,\mathrm{d}v = s$, and set $B_s := 1$ and $\eta_s := u_{s+1} - u_s$.

Lemma C.24 (Integer realization of the subregime optimizers)**.** Let $D \to \infty$ through integers and let $T \to \infty$ with $T/D \lesssim 1$. Fix $\alpha > 0$, $\alpha \neq 1/4$. For $\alpha > 1/4$, the preceding deterministic integer-batch schedule uses exactly D samples, reaches intrinsic time T, and satisfies

$$\sum_{s<D} \eta_s^2 \big(1 + T - u_{s+1}\big)^{-2+1/(2\alpha)} \asymp \frac{T^{1/(2\alpha)}}{D}.$$

For $1/4 < \alpha < 1/2$, the same order is attained by the simpler constant choice $B_s = 1$, $\eta_s = T/D$. For $0 < \alpha < 1/4$, that constant choice satisfies $\sum_{s<D} \eta_s^2 = T^2/D$.

In the $\beta < 0$ branch of subregime IM_1, there is a horizon $T \asymp D$ and an admissible unit-batch schedule for which

$$\sum_{s<D} \eta_s^2 \big(1 + T - u_{s+1}\big)^{-2+1/(2\alpha)} = o\Big(D^{-p/(2\alpha)}\Big).$$

These schedules satisfy the peak and bounded-batch conditions at the subregime-wise horizons and widths used in the resource bounds.

Proof. Assume first $\alpha > 1/4$. The data-quantile definition gives

$$\sum_{s<D} B_s = D, \qquad \sum_{s<D} \eta_s = T, \qquad \frac{1}{\eta_s} = \frac{1}{u_{s+1} - u_s} \int_{u_s}^{u_{s+1}} \bar{r}(v)\,\mathrm{d}v.$$

Thus the discrete ratio B_s/η_s is the cell average of $\bar{r}$. Moreover, $\inf_u \bar{r}(u) \asymp D/T$, so $\eta_s \lesssim T/D$. The shifted powers $\bar{r}(u)$ and $(1 + T - u)^{-2+1/(2\alpha)}$ vary by only constant factors on each unit-data cell. Consequently the discrete injection sum is comparable to

$$\int_0^T \frac{(1 + T - u)^{-2+1/(2\alpha)}}{\bar{r}(u)}\,\mathrm{d}u \asymp \frac{I_\alpha(T)^2}{D} \asymp \frac{T^{1/(2\alpha)}}{D}.$$

For a constant step, direct summation gives the same long-memory order when $1/4 < \alpha < 1/2$.

If $0 < \alpha < 1/4$, take $B_s = 1$ and $\eta_s = T/D$ directly. Then

$$\sum_{s<D} B_s = D, \qquad \sum_{s<D} \eta_s = T, \qquad \sum_{s<D} \eta_s^2 = \frac{T^2}{D},$$

which proves the finite-bulk assertion.

For the $\beta < 0$ branch of subregime IM_1, put

$$s_0 := \frac{p}{2\alpha}, \qquad h := 1 - \frac{1}{2\alpha}, \qquad L := D^{(1+s_0)/(1+h)},$$

choose a sufficiently large fixed r_0, and set

$$\bar{r}(u) := r_0 \max\left\{1, \left(\frac{1 + L}{1 + T - u}\right)^{1-1/(4\alpha)}\right\}, \qquad \int_0^T \bar{r}(u)\,\mathrm{d}u = D.$$

Then $T \asymp D$, and the same quantile construction has

$$\sum_{s<D} \eta_s^2 \big(1 + T - u_{s+1}\big)^{-2+1/(2\alpha)} \lesssim L^{-h} = o\Big(D^{-p/(2\alpha)}\Big),$$

because $s_0 < h$ and $h(1 + s_0)/(1 + h) > s_0$. Here $\eta_s \le 1/r_0$, so the clipped profile and shifted kernel are uniformly comparable on every cell; hence the continuum L^{-h} bound transfers to the displayed discrete sum. Finally, $B_s \equiv 1$ supplies the bounded-batch propagation alternative.

It remains to check admissibility at these scales. Put $x = T^{1/(2\alpha)}$. The largest step is $O(T/D)$, except that the $\beta < 0$ IM_1 construction directly has $\eta_s \le 1/r_0$. We have $x \asymp m$ in both IM_1 branches, FB_1, LM_1, LM_2, and the $\alpha > 1$ source-saturated IM_2–IM_3 branch; whereas $x/m \to 0$

in FB_2, LM_3, and the $1/2 < \alpha \le 1$ interior IM_2–IM_3 branch. Thus every source window holds, strictly in FB_2 because $p > 2\alpha$. For $\alpha < 1/2$, the $\mathrm{FB}_1, \mathrm{LM}_1, \mathrm{LM}_2$ choices satisfy

$$\frac{T/D}{m^{2\alpha-1}} \asymp x^{-p} \longrightarrow 0,$$

while the corresponding ratios in FB_2 and LM_3 are $x^{-2\alpha} \to 0$ and vanish by the defining inequalities, respectively. Since the normalized peak tends to zero, the peak condition holds for all sufficiently large widths. The row-mass certificate Eq. (C.10) for $\alpha > 1/2$ and Lemma C.13 below trace class give the Volterra margin; the strict FB_2 window has vanishing row mass. Rounding m, D changes only fixed factors. □

Two continuum schedule parameterizations. For fixed learning rate, $u(\tau) = \eta_0\tau$ and $B(\tau) = \eta_0 r_T^\star(\eta_0\tau)$, so

$$\int_0^{T/\eta_0} B(\tau)\,\mathrm{d}\tau = \int_0^T r_T^\star(u)\,\mathrm{d}u = D.$$

This proves Eq. (C.47). For fixed batch, Eq. (C.48) gives $\mathrm{d}\tau/\mathrm{d}u = r_T^\star(u)/B_0$. Hence the terminal iteration time is D/B_0, and $\int_0^{D/B_0} B_0\,\mathrm{d}\tau = D$.

On a terminal, subsequently unclipped segment with $r_T^\star(u) = c(1+T-u)^{-1+1/(4\alpha)}$, put $x = T-u$. Throughout its constant-batch-size time coordinate,

$$\frac{D}{B_0} - \tau = \frac{1}{B_0}\int_u^T r_T^\star(v)\,\mathrm{d}v = \frac{4\alpha c}{B_0}\left[(1+x)^{1/(4\alpha)} - 1\right].$$

Solving for x and substituting into $\eta = B_0/r_T^\star$ gives Eq. (C.49) with decay exponent $4\alpha - 1$. In other clipped configurations we assert only the plateau correspondence: lower/upper ratio clipping becomes maximum/minimum learning-rate plateaus in the constant-batch-size parameterization and minimum/maximum batch plateaus in the constant-learning-rate parameterization. Direct substitution gives Eq. (C.50). In the external conventions, the spectral parameter is 2α, the source parameter of Li et al. (2026a) is $p/(2\alpha)$, and the target-decay parameter of Bordelon & Mori (2026) is $p+1$. Hence their decay, peak, and risk exponents match the displayed native exponents, while both trace-class conditions reduce to $\alpha > 1/2$.

Matching lower bounds. Let $D_t := \sum_{s<t} B_s \le D$. Positivity of the DE Volterra recursion and Eq. (C.29) give, for $\alpha > 1/4$,

$$\mathcal{R}_{\sigma,t} \gtrsim \Phi_{\alpha,\beta}(T,m) + \sigma^2 \sum_{s<t} \frac{\eta_s^2}{B_s}\big(1+T-T_{s+1}\big)^{-2+1/(2\alpha)}.$$

Discrete Cauchy–Schwarz and the right Riemann sum imply

$$D_t \sum_{s<t} \frac{\eta_s^2}{B_s}\big(1+T-T_{s+1}\big)^{-2+1/(2\alpha)} \ge \left[\sum_{s<t} \eta_s\big(1+T-T_{s+1}\big)^{-1+1/(4\alpha)}\right]^2 \gtrsim T^{1/(2\alpha)}.$$

For $0 < \alpha < 1/4$, the target-free gap bound in Proposition C.17, together with the proved forcing comparison when $p > 0$, $\beta < 1 + 2\alpha$, and the parameters are off the critical lines, instead gives

$$\mathcal{R}_{\sigma,t} \gtrsim \Phi_{\alpha,\beta}(T,m) + \sigma^2 m^{1-4\alpha} \sum_{s<t} \frac{\eta_s^2}{B_s} \gtrsim \Phi_{\alpha,\beta}(T,m) + \sigma^2 m^{1-4\alpha}\frac{T^2}{D}.$$

These bounds use the integer sample count, not a continuum factorization. In the $\beta < 0$ branch of subregime IM_1, the lower ratio floor also gives $T \lesssim D$.

For compactness put $x := T^{1/(2\alpha)}$, so that the source window is $x \lesssim m$. Ignoring fixed constants, the preceding inequalities give the following lower objectives:

$$\begin{aligned}
&m^{-p} + x^{-p} + x/D && \text{in subregimes } \mathrm{LM}_1, \mathrm{LM}_2, \text{ and } \mathrm{IM}_1,\\
&m^{-2\alpha} + x^{-p} + x/D && \text{in subregime } \mathrm{LM}_3,\\
&m^{-2\alpha} + x^{-p} + m^{-1}x^{-(2\alpha-1)} + x/D && \text{in subregimes } \mathrm{IM}_2 \text{ and } \mathrm{IM}_3,\\
&m^{-p} + x^{-p} + m^{1-4\alpha}x^{4\alpha}/D && \text{in subregime } \mathrm{FB}_1,\\
&m^{-2\alpha} + x^{-p} + m^{1-4\alpha}x^{4\alpha}/D && \text{in subregime } \mathrm{FB}_2.
\end{aligned}$$

The only compute balance that changes inside the IM region is the following one. We isolate it here because both the unrestricted schedule problem and the fixed-batch schedule comparison use the same four terms.

Lemma C.25 (Joint IM_2–IM_3 compute balance)**.** Let $\alpha > 1/2$, $\beta > 1/2$, $p = 2\alpha + 2\beta - 1$, and $F \to \infty$. Then

$$\inf_{1 \ll x \lesssim m} \left\{ x^{-p} + m^{-2\alpha} + m^{-1}x^{-(2\alpha-1)} + \frac{mx}{F} \right\} \asymp \begin{cases} F^{-p/(1+p+2\beta)}, & 1/2 < \alpha \le 1, \\ F^{-\alpha/(1+\alpha)}, & \alpha > 1. \end{cases}$$

For $1/2 < \alpha \le 1$, one attaining choice is

$$x \asymp F^{1/(1+p+2\beta)}, \qquad m \asymp x^{2\beta}.$$

For $\alpha > 1$, the optimum reaches the source endpoint:

$$x \asymp m \asymp F^{1/[2(1+\alpha)]}.$$

At $\alpha = 1$, the rate is $F^{-1/2}$, but the allocation is not unique: every choice satisfying

$$mx \asymp F^{1/2}, \qquad F^{1/(2p)} \lesssim x \lesssim F^{1/4}$$

has the same order.

Proof. For $1/2 < \alpha \le 1$, apply weighted AM–GM to the forcing, label-noise, and feature-distortion terms with weights

$$\frac{2(1-\alpha)}{1+p+2\beta}, \qquad \frac{p}{1+p+2\beta}, \qquad \frac{p}{1+p+2\beta}.$$

They are nonnegative, sum to one, and their weighted product is $F^{-p/(1+p+2\beta)}$. At the displayed choice of x and m, those three terms match. Moreover, $4\alpha\beta - p = (2\alpha-1)(2\beta-1) > 0$, so the width floor is lower order, and $x \lesssim m$.

For $\alpha > 1$, use instead the label-noise, feature-distortion, and width-floor weights

$$\frac{\alpha}{1+\alpha}, \qquad \frac{\alpha}{(2\alpha-1)(1+\alpha)}, \qquad \frac{\alpha-1}{(2\alpha-1)(1+\alpha)}.$$

Their weighted product is $F^{-\alpha/(1+\alpha)}$. Taking $x \asymp m \asymp F^{1/[2(1+\alpha)]}$ matches these three terms, while $p > 2\alpha$ makes x^{-p} lower order. When $\alpha = 1$, direct substitution gives the stated family of attaining allocations. □

Attaining scales. Because $\sigma^2 > 0$ is fixed and the clean forcing is bounded, the upper FSL has the same order as these label-noise objectives for the schedules in Lemma C.24. At fixed data, the matching choices are

$$\begin{aligned}
(\mathrm{IM}_1,\ \beta < 0): &\quad m \asymp D^{1/(2\alpha)}, && T \asymp D;\\
(\mathrm{IM}_1,\ \beta > 0),\ \mathrm{FB}_1,\ \mathrm{LM}_1,\ \mathrm{LM}_2: &\quad m \asymp D^{1/(1+p)}, && T \asymp D^{2\alpha/(1+p)};\\
\mathrm{FB}_2: &\quad m \asymp D^{\frac{p}{p(1-2\alpha)+8\alpha^2}}, && T \asymp D^{\frac{4\alpha^2}{p(1-2\alpha)+8\alpha^2}};\\
\mathrm{LM}_3: &\quad m \asymp D^{\frac{p}{2\alpha(1+p)}}, && T \asymp D^{2\alpha/(1+p)};\\
\mathrm{IM}_2,\ \mathrm{IM}_3: &\quad m \asymp D^{2\beta/(1+p)}, && T \asymp D^{2\alpha/(1+p)}.
\end{aligned}$$

The first uses the plateau–ramp construction; the second and last use the data-quantile square-root schedule in IM, while all listed FB and LM cases use the constant $B_s = 1$ construction. Direct substitution gives the fixed-data column of Table 4.

On $mD \le \mathfrak{f}$, the matching choices for, respectively, $(\mathrm{IM}_1, \beta < 0)$; the shared $(\mathrm{IM}_1, \beta > 0), \mathrm{FB}_1, \mathrm{LM}_1, \mathrm{LM}_2$ class; FB_2; and LM_3, are

$$\begin{aligned}
&m \asymp \mathfrak{f}^{1/(1+2\alpha)}, && D \asymp \mathfrak{f}^{2\alpha/(1+2\alpha)}, && T \asymp D;\\
&m \asymp \mathfrak{f}^{1/(2+p)}, && D \asymp \mathfrak{f}^{(1+p)/(2+p)}, && T \asymp \mathfrak{f}^{2\alpha/(2+p)};\\
&m \asymp \mathfrak{f}^{\frac{p}{p(2-2\alpha)+8\alpha^2}}, && D \asymp \mathfrak{f}^{\frac{p(1-2\alpha)+8\alpha^2}{p(2-2\alpha)+8\alpha^2}}, && T \asymp \mathfrak{f}^{\frac{4\alpha^2}{p(2-2\alpha)+8\alpha^2}};\\
&m \asymp \mathfrak{f}^{\frac{p}{p+2\alpha(1+p)}}, && D \asymp \mathfrak{f}^{\frac{2\alpha(1+p)}{p+2\alpha(1+p)}}, && T \asymp \mathfrak{f}^{\frac{4\alpha^2}{p+2\alpha(1+p)}}.
\end{aligned}$$

Balancing proves the corresponding table entries. Indeed, $x \lesssim m$ and $1 - 4\alpha > 0$ lower-bound the FB_1 noise by x/D, explaining its shared rate, while $p > 2\alpha$ makes the FB_2 source window strict. If ε bounds all three FB_2 terms, then $m \gtrsim \varepsilon^{-1/(2\alpha)}$ and $x \gtrsim \varepsilon^{-1/p}$; substitution into the noise term gives

$$\varepsilon \gtrsim D^{-\frac{2\alpha p}{p(1-2\alpha)+8\alpha^2}} \quad \text{or} \quad \varepsilon \gtrsim \mathfrak{f}^{-\frac{2\alpha p}{p(2-2\alpha)+8\alpha^2}},$$

under the data and compute budgets, respectively.

For IM_2 and IM_3, apply Lemma C.25 with $F = \mathfrak{f}$. Together with $D = \mathfrak{f}/m$ and $T = x^{2\alpha}$, its two attaining choices give exactly the two compute branches in Table 4. Both use the data-quantile $B_s = 1$ schedule from Lemma C.24; at $\alpha = 1$ the rates agree at $\mathfrak{f}^{-1/2}$, with no forced logarithmic correction and no unique width allocation.

All upper constructions have $B_s \equiv 1$, so batch integrality and the bounded-batch propagation alternative are exact; rounding m, D preserves orders. The uniform FSL, these lower bounds, and Lemma C.24 prove Corollary C.23 and Table 4 over the fixed admissible class and the stated source range. The high-source FB_2 continuation separately invokes Assumption C.19.

D Sharp transfer for regularly varying joint schedules

This appendix gives the full regular-variation version of the schedule results stated in the main text. Appendix D.1 converts the low-spectrum tail into a two-time memory kernel. Appendix D.2 uses examples and counterexamples to show what the spectral, uniformity, and stability assumptions permit and exclude. Appendix D.3 derives the noisy–clean response in long memory, including the effect of feedback, while Appendix D.4 treats the two endpoint contributions that arise under integrable memory. Appendix D.5 proves the converse that recovers B/η from the observed gap below the memory ceiling. Appendix D.6 combines the gap with the clean risk to obtain the preserve–change–destroy classification, and Appendix D.7 translates the intrinsic-time results to polynomial schedules in iteration time.

We first state the regular-variation condition shared by these results.

Definition D3 (Width–time triangular sequence)**.** A width–time triangular sequence is a family indexed by the width m. At width m, fix a finite system $\boldsymbol{W}_m$, a deterministic schedule $\{(\eta_{m,s}, B_{m,s})\}_{s<t_m}$, and a terminal index t_m, and set

$$T_m := \sum_{s<t_m} \eta_{m,s}.$$

Both the finite system and the observation horizon may vary with m, and the limit is taken with $m, t_m, T_m \to \infty$. Uniform statements along the sequence use constants and bounds independent of m and t_m. Below, we write t and T_t for the terminal pair t_m and T_m when no confusion can arise.

By contrast, a fixed infinite-spectrum system keeps the dynamics and schedule fixed and sends only the terminal time to infinity.

Assumption D.1 (Regularly varying schedule path)**.** Consider either one fixed infinite-spectrum system or a width–time triangular sequence in Definition D3; omit the width index in the fixed system. A schedule path satisfies the required regularity if:

(a) For some $\vartheta \in \mathbb{R}$, $1/r_m$ is eventually positive and monotone, and

$$\frac{r_m(T_t)}{r_m(xT_t)} \longrightarrow x^{-\vartheta} \tag{D.1}$$

for every $x > 0$.

(b) For every sufficiently small $\epsilon > 0$, there is a finite constant C_ϵ, independent of the width and horizon, such that eventually

$$\frac{r_m(T_t)}{r_m(xT_t)} \le C_\epsilon \max\{x^{-\vartheta-\epsilon}, x^{-\vartheta+\epsilon}\}, \qquad x > 0. \tag{D.2}$$

In a triangular sequence, the convergence in part *(a)* is locally uniform and eventual monotonicity begins beyond a common intrinsic time.

These conditions permit width-dependent slow variation but exclude hidden oscillations or bursts; polynomial width amplitudes must be tracked separately.

Lemma D.2 (Uniform Karamata consequence)**.** Under Assumption D.1, if $\vartheta < 1$, then

$$\int_0^{xT_t} \frac{\mathrm{d}u}{r_m(u)} \sim \frac{x^{1-\vartheta}}{1-\vartheta} \frac{T_t}{r_m(T_t)},$$

locally uniformly for $x > 0$.

Proof. Choose $\epsilon < 1 - \vartheta$ in Eq. (D.2). After the change of variables $u = yT_t$, the Potter bound supplies an integrable envelope, so dominated convergence gives the result. □

We use the following two-time notation throughout this section. For the schedule coordinates in Eq. (2.1), let

$$H_{t,s} := T_t - T_{s+1}, \qquad \Psi_{t,s} := \sum_j \widehat{\lambda}_j^2 Q_{s+1,t}(\widehat{\lambda}_j).$$

Here $H_{t,s}$ is the age of the injection at time s, while $\Psi_{t,s}$ is its surviving variance before weighting by the injected noise; thus $K_{t,s} = (T_{s+1} - T_s)\Psi_{t,s}/r_s$. We use the piecewise-constant interpolation $r(u) = r_s$ for $T_s \le u < T_{s+1}$, and write $h_t := \max_{s<t}(T_{s+1} - T_s)$ for the largest intrinsic-time step.

D.1 From the empirical spectral tail to two-time memory

The schedule-transfer results use the following two-time kernel condition. The theorem below derives it from the empirical low-spectrum tail.

Assumption D.3 (Long-memory two-time kernel control)**.** Consider either one fixed infinite-spectrum system or a width–time triangular sequence in Definition D3; omit the width index in the fixed system. The two-time kernel has the following properties for some $0 < q_{\mathcal{K}} < 1$:

(a) The reference profile satisfies

$$k(v) \sim v^{-q_{\mathcal{K}}} L_{\mathcal{K}}(v), \tag{D.3}$$

where $L_{\mathcal{K}}$ is eventually positive and slowly varying.

(b) For every fixed $\varepsilon > 0$, the bulk-age kernel satisfies

$$\sup_{s<t:\ H_{t,s} \ge \varepsilon T_t} \left| \frac{\Psi_{t,s}}{k(H_{t,s})} - 1 \right| \longrightarrow 0. \tag{D.4}$$

(c) There are nonincreasing envelopes G_m such that

$$0 \le \Psi_{t,s} \le G_m(H_{t,s}), \qquad \sup_m G_m(0) < \infty, \tag{D.5}$$

and, for every sufficiently small $\xi > 0$, there is a finite constant C_ξ, independent of m, t, such that

$$\limsup_{m,t\to\infty} \frac{\int_0^{\varepsilon T_t} G_m(v)\,\mathrm{d}v}{T_t k(T_t)} \le C_\xi \varepsilon^{1-q_{\mathcal{K}}-\xi}, \qquad 0 < \varepsilon < \frac{1}{2}. \tag{D.6}$$

(d) The intrinsic mesh satisfies

$$h_t = o(T_t k(T_t)) . \tag{D.7}$$

Since $Tk(T)$ has positive index $1 - q_{\mathcal{K}}$, it diverges, so a uniformly bounded learning rate satisfies Eq. (D.7). These are the main text's precise fine-mesh long-memory conditions.

The disjoint age intervals and monotonicity of G_m give the repeatedly used recent-endpoint bound

$$\sum_{H_{t,s} \le x} (T_{s+1} - T_s) G_m(H_{t,s}) \le h_t G_m(0) + \int_0^x G_m(v)\, \mathrm{d}v. \tag{D.8}$$

The next conditional, forward theorem transfers one-time empirical tails to two-time memory; global Potter control is needed at recent injection.

Theorem D.4 (Sharp empirical spectral transfer to two-time memory)**.** Condition on a deterministic empirical width–time sequence indexed by m, with terminal index $t = t_m$ and $T_t \to \infty$. Fix $0 < q_{\mathcal{K}} < 1$ and an eventually positive slowly varying function $L_{\mathcal{K}}$. Suppose, locally uniformly for $c > 0$,

$$\sum_{0 < \widehat{\lambda}_j \le c/T_t} \widehat{\lambda}_j^2 \sim \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)} c^{q_{\mathcal{K}}} T_t^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T_t), \tag{D.9}$$

and assume the corresponding global Potter/no-escape bound Eq. (B.1) for every sufficiently small slack $\xi \in (0, \min\{q_{\mathcal{K}}, 1 - q_{\mathcal{K}}\})$, with $T = T_t$ and the normalization in Eq. (D.9). Under part *(a)* of Assumption C.2 and Eq. (D.7), for every fixed $\varepsilon > 0$,

$$\sup_{s<t:\ H_{t,s} \ge \varepsilon T_t} \left| \frac{\Psi_{t,s}}{H_{t,s}^{-q_{\mathcal{K}}} L_{\mathcal{K}}(H_{t,s})} - 1 \right| \longrightarrow 0. \tag{D.10}$$

If, in addition, $\sup_m \sum_j \widehat{\lambda}_j^2 < \infty$ on the conditioned sequence, then there are envelopes satisfying Eqs. (D.5) and (D.6) for the profile in Eq. (D.3). Hence Assumption D.3 holds under the additional trace bound.

Proof. Transfer the cutoff over bulk ages, then construct the recent-endpoint envelope. Write $H = H_{t,s}$. Since $H/T_t \in [\varepsilon, 1]$, Eq. (D.9), the uniform convergence theorem for slowly varying functions, and the global Potter bound give, locally uniformly for $c > 0$,

$$\sum_{0 < \widehat{\lambda}_j \le c/H} \widehat{\lambda}_j^2 \sim \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)} c^{q_{\mathcal{K}}} H^{-q_{\mathcal{K}}} L_{\mathcal{K}}(H),$$

uniformly over all such ages.

Joint pointwise stability gives $0 \le Q_{s+1,t}(\lambda) \le e^{-\delta H \lambda}$ for its fixed margin $\delta > 0$. Moreover, Eq. (D.7) and $k(T) \to 0$ imply $h_t/T_t \to 0$. On every compact critical band $a \le H\lambda \le b$, expanding the one-step factors gives $Q_{s+1,t}(\lambda) = e^{-2H\lambda}(1 + o(1))$ uniformly. If the supremum in Eq. (D.10) did not converge to zero, there would be a sequence of injection ages along which the normalized error remained bounded away from zero. Along this sequence, the critical-band and cutoff limits give the Laplace–Stieltjes limit on compact rescaled intervals, while Potter control and the exponential bound remove both endpoints as in Theorem B.1. Hence

$$\frac{\Psi_{t,s}}{H^{-q_{\mathcal{K}}} L_{\mathcal{K}}(H)} \longrightarrow \frac{2^{q_{\mathcal{K}}}}{\Gamma(q_{\mathcal{K}} + 1)} \Gamma(q_{\mathcal{K}} + 1) 2^{-q_{\mathcal{K}}} = 1,$$

a contradiction, proving Eq. (D.10) and, with Eq. (D.3), Eq. (D.4).

For the endpoint, take $G_m(v) := \sum_j \widehat{\lambda}_j^2 e^{-\delta v \widehat{\lambda}_j}$. It is nonincreasing, bounds $\Psi_{t,s}$, and is uniformly bounded at zero. The global Potter envelope yields $G_m(v) \lesssim k(T)(T/v)^{q_{\mathcal{K}} + \xi}$ for $0 < v \le \varepsilon T$: integration by parts writes $G_m(v) = \delta v \int_0^\infty e^{-\delta v x} A_m(x)\, \mathrm{d}x$, where $A_m(x) = \sum_{0 < \widehat{\lambda}_j \le x} \widehat{\lambda}_j^2$, and the two branches of the Potter bound give the displayed estimate. Integration, using $q_{\mathcal{K}} + \xi < 1$, yields

$$\int_0^{\varepsilon T} G_m(v)\, \mathrm{d}v \lesssim \varepsilon^{1 - q_{\mathcal{K}} - \xi} T k(T).$$

This is Eq. (D.6); the remaining envelope and mesh conditions were assumed. □

For PLRF with $d/m \to c > 1$ and $1/4 < \alpha < 1/2$, the endpoint trace is tight because

$$\mathbb{E}_{\boldsymbol{W}} \operatorname{tr}(\widehat{\boldsymbol{H}}^2) = \left(1 + \frac{1}{m}\right) \operatorname{tr}(\boldsymbol{\Lambda}^2) + \frac{(\operatorname{tr} \boldsymbol{\Lambda})^2}{m} = O(1).$$

Thus $\operatorname{tr}(\widehat{\boldsymbol{H}}^2) = O_{\mathbb{P}}(1)$, excluding polynomial endpoint blowup.

D.2 Examples and counterexamples for the transfer assumptions

The spectral criterion concerns weighted mass near zero rather than a polynomial formula for individual eigenvalues. The following table collects two component-power examples and two elementary failures.

Stretched-exponential spectra. We examine how stretched-exponential spectral decay affects the forcing and memory component asymptotics. The calligraphic components below are specialization-specific resolvent-DE quantities.

Definition D4 (Stretched-exponential spectral specialization)**.** Fix $0 < \gamma < 1$, $c > 1$, and $\beta \in \mathbb{R}$, let $d = \lceil cm \rceil$, and set

$$\lambda_j = e^{-j^\gamma}, \qquad |\theta_j^\star| \asymp j^{-\beta}.$$

The eigenvalue sequence in Definition D4 is not regularly varying in the index j. Since $\gamma < 1$, the ratio of successive eigenvalues satisfies $\lambda_{j+1}/\lambda_j \to 1$. Its small-eigenvalue cumulative energies have the same order as regularly varying reference functions, so training retains power laws with logarithmic corrections. For the constant schedule below, write $T := \eta t$.

Theorem D.5 (Stretched-exponential forcing and memory asymptotics)**.** Under Definition D4, let the constant schedule $\eta_t \equiv \eta$, $B_t \equiv B$ satisfy Assumption A.2, with $\eta \asymp 1$. Uniformly on every strict intermediate subwindow

$$1 \ll T, \qquad \log T \le (1 - \varepsilon) m^\gamma$$

with fixed $\varepsilon > 0$, the aligned pure-point components satisfy

$$\mathcal{F}_{pp}(t, m) \asymp T^{-1} (\log T)^{(1-\gamma-2\beta)/\gamma}, \tag{D.11}$$

$$\frac{1}{\eta} \mathcal{K}_{pp}(t, m) \asymp \frac{\eta}{B} T^{-2} (\log T)^{(1-\gamma)/\gamma}, \tag{D.12}$$

$$\sum_{s=t}^{\infty} \mathcal{K}_{pp}(s, m) \asymp \frac{\eta}{B} T^{-1} (\log T)^{(1-\gamma)/\gamma}. \tag{D.13}$$

The proof uses the following tail and Laplace-sum estimates.

Stretched-exponential sum calculus.

Lemma D.6 (Tail and Laplace sums)**.** For fixed $q > 0$ and $a \in \mathbb{R}$,

$$\sum_{j \ge n} j^{-a} e^{-q j^\gamma} \sim \frac{1}{\gamma q} n^{1-\gamma-a} e^{-q n^\gamma}.$$

$$\sum_{j \ge 1} j^{-a} e^{-q j^\gamma} \exp\{-2u e^{-j^\gamma}\} \sim \frac{\Gamma(q)}{\gamma (2u)^q} \big(\log(2u)\big)^{(1-a)/\gamma - 1}.$$

Proof. For the tail, eventual monotonicity and variation scale $n^{1-\gamma} \to \infty$ permit sum–integral comparison; $y = q x^\gamma$ gives

$$\int_n^\infty x^{-a} e^{-q x^\gamma} \, dx = \frac{q^{(a-1)/\gamma}}{\gamma} \int_{q n^\gamma}^\infty y^{(1-a)/\gamma - 1} e^{-y} \, dy.$$

The incomplete-Gamma tail expansion proves the first claim.

For the second, the summand is concentrated at $x \asymp (\log u)^{1/\gamma}$ on a window of width $\asymp (\log u)^{(1-\gamma)/\gamma} \to \infty$, so sum–integral comparison applies. With $x^\gamma = z$ and $y = 2ue^{-z}$, the continuous proxy becomes

$$\int_1^\infty x^{-a} e^{-qx^\gamma} e^{-2ue^{-x^\gamma}}\, \mathrm{d}x = \frac{(2u)^{-q}}{\gamma} \int_0^{2u/e} y^{q-1} e^{-y} \left(\log \frac{2u}{y}\right)^{(1-a)/\gamma - 1} dy.$$

After division by $(\log(2u))^{(1-a)/\gamma-1}$, the logarithmic factor tends to one on compact subsets of $y > 0$. Near zero, $q > 0$ absorbs the remaining logarithmic factor. On $1 \le y \le \sqrt{u}$, the normalized factor is uniformly bounded, while on $[\sqrt{u}, 2u/e]$ the exponential tail dominates its at-most-polylogarithmic growth. Dominated convergence therefore yields $\Gamma(q)$. □

The condition $\gamma < 1$ both makes the local index width $j^{1-\gamma}$ diverge, validating the integral approximation, and gives $\lambda_{j+1}/\lambda_j = \exp\{-[(j+1)^\gamma - j^\gamma]\} \to 1$. At $\gamma = 1$, geometric lacunarity can introduce log-periodic corrections and destroy regular variation.

Proof of Theorem D.5. We first compute the cumulative forcing and memory energies, then apply the Laplace-sum estimate on the active spectral shell. Finally, the finite-width cutoff verifies uniformity on the stated intermediate subwindow.

Cumulative energies and component transforms.

Let $x \downarrow 0$ and set $n_x = (\log(1/x))^{1/\gamma}$. Applying the first part of Lemma D.6 gives

$$\sum_{\lambda_j \le x} \lambda_j (\theta_j^\star)^2 \asymp x \left(\log \frac{1}{x}\right)^{(1-\gamma-2\beta)/\gamma}$$

$$\frac{\eta^2}{B} \sum_{\lambda_j \le x} \lambda_j^2 \sim \frac{\eta^2}{2\gamma B} x^2 \left(\log \frac{1}{x}\right)^{(1-\gamma)/\gamma}.$$

Thus the kernel cumulative energy is regularly varying at zero with index 2. Under the stated two-sided target comparison, the forcing cumulative energy is only comparable to an index-one regularly varying reference; it is itself regularly varying when $\theta_j^\star \sim C_\theta j^{-\beta}$, in which case its first display above strengthens to

$$\sum_{\lambda_j \le x} \lambda_j (\theta_j^\star)^2 \sim \frac{C_\theta^2}{\gamma} x \left(\log \frac{1}{x}\right)^{(1-\gamma-2\beta)/\gamma}.$$

Moreover,

$$\sum_j \lambda_j < \infty, \qquad \sum_j \lambda_j^2 < \infty, \qquad \sum_j \lambda_j (\theta_j^\star)^2 < \infty$$

for every fixed β. Hence the covariance is trace class and $\eta \asymp 1$, whereas $\boldsymbol{\theta}^\star \in \ell^2$ iff $\beta > 1/2$.

On the main shell $\lambda \asymp T^{-1}$, the remainder in $t \log q_\eta(\lambda)$ is $O(T\lambda^2) = o(1)$, so $q_\eta(\lambda)^t = \exp\{-2T\lambda\}(1 + o(1))$. In Lemma D.6, $(a, q) = (2\beta, 1)$ proves Eq. (D.11) and $(a, q) = (0, 2)$ proves Eq. (D.12). If $\theta_j^\star \sim C_\theta j^{-\beta}$, the corresponding leading constants are

$$\mathcal{F}_{pp}(t, m) \sim \frac{C_\theta^2}{2\gamma} T^{-1} \big(\log(2T)\big)^{(1-\gamma-2\beta)/\gamma},$$

$$\mathcal{K}_{pp}(t, m) \sim \frac{\eta^2}{4\gamma B} T^{-2} \big(\log(2T)\big)^{(1-\gamma)/\gamma}.$$

The modal geometric sum gives the terminal kernel tail directly:

$$\sum_{s=t}^\infty \mathcal{K}_{pp}(s, m) = \frac{\eta}{B} \sum_j \frac{\lambda_j}{2 - \eta(1 + 1/B)\lambda_j} q_\eta(\lambda_j)^t.$$

The denominator tends to 2 on the active shell, so the Laplace-sum formula yields

$$\sum_{s=t}^\infty \mathcal{K}_{pp}(s, m) \sim \frac{\eta}{4\gamma B} T^{-1} \big(\log(2T)\big)^{(1-\gamma)/\gamma},$$

which proves Eq. (D.13).

Finite-width cutoff.

The critical learned index solves $Te^{-j_T^\gamma} \asymp 1$. Thus

$$j_T \asymp (\log T)^{1/\gamma}, \qquad 1 \ll j_T \ll m \iff 1 \ll T,\ \log T \ll m^\gamma,$$
$$T_{\text{term}} \asymp \lambda_m^{-1} = e^{m^\gamma}, \qquad \log T \le (1-\varepsilon)m^\gamma \quad \text{on a uniform strict subwindow.}$$

For completeness, for fixed $q > 0$ and $a \in \mathbb{R}$, the truncated continuous proxy satisfies

$$\int_1^m x^{-a} e^{-qx^\gamma} e^{-2Te^{-x^\gamma}}\,\mathrm{d}x = \frac{(2T)^{-q}}{\gamma} \int_{2Te^{-m^\gamma}}^{2T/e} y^{q-1} e^{-y} \left(\log \frac{2T}{y}\right)^{(1-a)/\gamma - 1} dy.$$

The logarithmic factor is kept on this real-valued domain. After it is extracted at the required asymptotic precision, the exponential tail permits the remaining Gamma-type integral to be extended to infinity. Thus $2Te^{-m^\gamma} \to 0$, $= \Theta(1)$, and $\to \infty$ give, respectively, the intermediate power law, finite-width crossover, and exit from the regularly varying window. At fixed m, eventual convergence is geometric.

On the strict subwindow in the theorem, $Te^{-m^\gamma} \le e^{-\varepsilon m^\gamma} \to 0$. Hence the preceding sum asymptotics hold uniformly there, and Eqs. (D.11) to (D.13) complete the proof. □

Irregular targets with the canonical forcing and memory exponents. Fix $\alpha > 1/4$, let $p := 2\alpha + 2\beta - 1 > 0$, and set $\lambda_j = j^{-2\alpha}$ and $|\theta_j^\star|^2 = j^{-2\beta} c_j$. Suppose $c_j \ge 0$ and, with $C_n := \sum_{j \le n} c_j$, assume $C_n/n \to 1$. For $N_x := \lceil x^{-1/(2\alpha)} \rceil$, summation by parts gives

$$\sum_{\lambda_j \le x} \lambda_j |\theta_j^\star|^2 = \sum_{j \ge N_x} j^{-(p+1)} c_j \sim \frac{1}{p} N_x^{-p} \sim \frac{1}{p} x^{p/(2\alpha)}.$$

Indeed, for $a_j = j^{-(p+1)}$,

$$\sum_{j=N}^{\infty} c_j a_j = -C_{N-1} a_N + \sum_{j=N}^{\infty} C_j (a_j - a_{j+1}) \sim -N^{-p} + \frac{p+1}{p} N^{-p} = \frac{1}{p} N^{-p}.$$

Thus the forcing tail has the same leading power and constant as for the canonical choice $c_j \equiv 1$, although the target coordinates need not obey any fixed power envelope. Three examples satisfy the averaging condition:

$$\begin{aligned} &c_{2k} = 2, \quad c_{2k-1} = (2k-1)^{-2\varepsilon}, \quad \varepsilon > 0, && \text{alternating two-rate;} \\ &c_{k^2} = 2k-1, \quad c_j = 0 \text{ otherwise}, && \text{mass-compensated sparse;} \\ &c_j = Z_j^2, \quad Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), && \text{frozen random amplitude.} \end{aligned}$$

For the first example, the even terms contribute $n + O(1)$ to C_n and the odd terms contribute $o(n)$. For the second, $C_n = \lfloor \sqrt{n} \rfloor^2$. For the third, $C_n/n \to 1$ almost surely by the strong law; the infinite sequence is drawn once and then held fixed. The first construction has different powers on its even and odd subsequences, the second vanishes off the squares, and in the third the amplitudes are almost surely neither bounded above nor bounded away from zero. Consequently, each construction violates $|\theta_j^\star| \asymp j^{-\rho}$ for every fixed ρ, almost surely in the random case.

These population tails can also be realized as exact empirical tails by a deterministic aligned truncation. Take $d = 2m$, $\boldsymbol{W}_m = (I_m, 0)^\top$, and truncate one infinite target sequence to its first $2m$ coordinates. Then $\widehat{\boldsymbol{H}}_m = \mathrm{diag}(\lambda_1, \dots, \lambda_m, 0, \dots, 0)$. If $T \to \infty$ with $T = o(m^{2\alpha})$, set $N_T := \lceil T^{1/(2\alpha)} \rceil$. Then

$$\sum_{N_T \le j \le m} j^{-(p+1)} c_j \sim \frac{1}{p} T^{-p/(2\alpha)}, \qquad \sum_{N_T \le j \le m} j^{-4\alpha} \sim \frac{1}{4\alpha - 1} T^{-(4\alpha-1)/(2\alpha)}.$$

The omitted tails beyond m are lower order, and the same estimates hold locally uniformly at constant-factor cutoffs. Regular variation of the infinite tails gives the Potter envelope, and truncation can only decrease the cutoff mass. Hence, for a fixed constant schedule and some fixed $\delta > 0$, if $\eta(1 + 1/B) \le 2 - \delta$, these tails satisfy the stable no-escape hypotheses of Theorem A.4. The resulting forcing asymptotic matches the canonical diagonal choice $c_j \equiv 1$; the memory response is unchanged because $K_{\boldsymbol{W}_m}$ is target-independent. The random-case claims hold almost surely with respect to the single frozen target draw.

Proposition D.7 (Power-law spectrum with faster-than-polynomial forcing)**.** Fix $\alpha > 1/4$ and let

$$\lambda_j = j^{-2\alpha}, \qquad |\theta_j^\star|^2 = e^{-j}.$$

The target-weighted spectral tail satisfies

$$\sum_{\lambda_j \le x} \lambda_j |\theta_j^\star|^2 \asymp x \exp\Big(-x^{-1/(2\alpha)}\Big), \qquad x \downarrow 0.$$

For the deterministic aligned truncations, take $d = 2m$, $\boldsymbol{W}_m = (I_m, 0)^\top$, and truncate the displayed spectrum and target to their first $2m$ coordinates. Fix a constant schedule $\eta_t \equiv \eta$, $B_t \equiv B \ge 1$, satisfying

$$\eta\left(1 + \frac{1}{B}\right) \le 2 - \delta$$

for some $\delta > 0$. Along any joint limit $m, t \to \infty$ such that $T = \eta t \to \infty$ and $T = o(m^{2\alpha+1})$, the exact learnable forcing obeys

$$-\log F_{\boldsymbol{W}_m, >0}(t) \sim \frac{2\alpha+1}{2\alpha}(4\alpha)^{1/(2\alpha+1)} T^{1/(2\alpha+1)}.$$

Consequently,

$$F_{\boldsymbol{W}_m, >0}(t) = o(T^{-q}) \qquad \text{for every } q > 0.$$

Proof. If $N_x := \lceil x^{-1/(2\alpha)} \rceil$, then $\sum_{\lambda_j \le x} \lambda_j |\theta_j^\star|^2 = \sum_{j \ge N_x} j^{-2\alpha} e^{-j}$. Successive summands have ratio at most e^{-1}, so this tail is comparable to its first term:

$$\sum_{j \ge N_x} j^{-2\alpha} e^{-j} \asymp N_x^{-2\alpha} e^{-N_x} \asymp x e^{-x^{-1/(2\alpha)}}.$$

For the aligned truncation,

$$F_{\boldsymbol{W}_m, >0}(t) = \sum_{j=1}^{m} j^{-2\alpha} e^{-j} q_\eta(j^{-2\alpha})^t.$$

Since the schedule is fixed, $(-\log q_\eta(\lambda))/(\eta\lambda) \to 2$ as $\lambda \downarrow 0$. Thus, for every $0 < \varepsilon < 2$, there is a fixed J_ε such that, for $j \ge J_\varepsilon$,

$$(2-\varepsilon)\eta j^{-2\alpha} \le -\log q_\eta(j^{-2\alpha}) \le (2+\varepsilon)\eta j^{-2\alpha}.$$

The stability margin makes the finitely many terms $j < J_\varepsilon$ exponentially small in T.

For $A > 0$, an elementary discrete Laplace estimate gives

$$-\log \sum_{j=J_\varepsilon}^{m} j^{-2\alpha} \exp\big(-j - ATj^{-2\alpha}\big) \sim \frac{2\alpha+1}{2\alpha}(2\alpha A)^{1/(2\alpha+1)} T^{1/(2\alpha+1)}$$

whenever $T^{1/(2\alpha+1)} = o(m)$. Indeed, the exponent $x + ATx^{-2\alpha}$ is minimized at $x = (2\alpha AT)^{1/(2\alpha+1)}$, where its value is

$$\frac{2\alpha+1}{2\alpha}(2\alpha A)^{1/(2\alpha+1)} T^{1/(2\alpha+1)}.$$

A nearest integer gives the matching lower bound, while splitting the sum below, near, and above this scale gives the upper bound; the polynomial factor contributes only $O(\log T)$ to the logarithm.

Applying this estimate with $A = 2 + \varepsilon$ and $A = 2 - \varepsilon$ yields

$$\frac{2\alpha+1}{2\alpha}\big(2\alpha(2-\varepsilon)\big)^{1/(2\alpha+1)} \le \liminf \frac{-\log F_{\boldsymbol{W}_m, >0}(t)}{T^{1/(2\alpha+1)}}$$

and

$$\limsup \frac{-\log F_{\boldsymbol{W}_m, >0}(t)}{T^{1/(2\alpha+1)}} \le \frac{2\alpha+1}{2\alpha}\big(2\alpha(2+\varepsilon)\big)^{1/(2\alpha+1)}.$$

Letting $\varepsilon \downarrow 0$ proves the stated constant. Since $T^{1/(2\alpha+1)}/\log T \to \infty$, the forcing is smaller than every inverse power of T. The final memory statement follows directly from the definition of $K_{\boldsymbol{W}_m}$. □

The stretched-exponential row is understood on the strict intermediate window of Theorem D.5. Width-dependent no-escape is a separate issue: target-weighted mass can move toward zero with width and change the forcing-transform constant, as shown in Proposition D.10. Beyond frozen dynamics, Barkeshli et al. (2026) find dataset- and model-size scaling laws on Erdős–Rényi random-walk data without explicit power-law structure, showing that loss scaling need not directly inherit a raw-data power law.

The following deterministic empirical operators separate the spectral and feedback hypotheses.

For $1/2 < a < 1$, $b > 1$, and width m, consider the deterministic spectrum $\widehat{\lambda}_j = j^{-a}$, $1 \le j \le m$, and denote its target-alignment weights by $w_j := |\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star\rangle|^2 = j^{-b}$. For this example, write

$$q_{\mathcal{K}} := 2 - \frac{1}{a} \in (0,1), \qquad c_{\mathcal{K}} := \frac{\Gamma(q_{\mathcal{K}}+1)}{2^{q_{\mathcal{K}}}(2a-1)}, \qquad L_{\mathcal{K}}(T) \equiv c_{\mathcal{K}}.$$

Proposition D.8 (A deterministic long-memory example)**.** Along every sequence $T = T_m \to \infty$ with $T = o(m^a)$, locally uniformly for $c > 0$,

$$\sum_{0<\widehat{\lambda}_j \le c/T} w_j \sim \frac{c^{(b-1)/a}}{b-1} T^{-(b-1)/a},$$

$$\sum_{0<\widehat{\lambda}_j \le c/T} \widehat{\lambda}_j^2 \sim \frac{c^{2-1/a}}{2a-1} T^{-(2-1/a)}.$$

Both tails satisfy the uniform Potter/no-escape envelope. Consequently, bounded pointwise-stable steps satisfying Eq. (D.7) verify the pure-power kernel asymptotic and the two-time conditions of Theorem D.13 through Theorem D.4, with the preceding $q_{\mathcal{K}}$ and $L_{\mathcal{K}}$. By contrast, a constant schedule with η/B bounded away from zero has no row-stability margin uniform in m.

Proof. Integral comparison, with the active index $n_x = \lceil x^{-1/a} \rceil$, gives

$$\sum_{j \ge n_x} j^{-b} \sim \frac{x^{(b-1)/a}}{b-1}, \qquad \sum_{j \ge n_x} j^{-2a} \sim \frac{x^{2-1/a}}{2a-1}.$$

Since $T^{1/a} = o(m)$, tails beyond m are negligible at $x = c/T$ uniformly on compact c-sets. The same integral comparisons give Eq. (B.1) and $\sup_m \sum_{j \le m} \widehat{\lambda}_j^2 < \infty$.

For the last claim, let $q(\lambda) = 1 - 2\eta\lambda + (1 + 1/B)\eta^2\lambda^2$. Under constant pointwise-stable η, B, the limiting row mass satisfies

$$\lim_{t\to\infty} \sum_{s<t} K_{t,s} = \frac{\eta^2}{B} \sum_{j \le m} \frac{\widehat{\lambda}_j^2}{1 - q(\widehat{\lambda}_j)} \ge \frac{\eta}{2B} \sum_{j \le m} \widehat{\lambda}_j \asymp m^{1-a}.$$

Thus two-time long memory does not imply constant-ratio row stability; one must additionally shrink η/B or raise the ratio path. □

Proposition D.9 (Stable no-escape for a non-power spectrum)**.** Let

$$\widehat{\boldsymbol{H}}_m = \mathrm{diag}_{1 \le j \le m}(e^{-\sqrt{j}}), \qquad w_j = e^{-\sqrt{j}}, \qquad B = 1, \qquad \eta = \frac{1}{4}.$$

If $T_m = \eta t_m \to \infty$ and $\log T_m \le (1-\rho)\sqrt{m}$ for fixed $\rho \in (0,1)$, then, locally uniformly for $c > 0$,

$$\sum_{\substack{j \le m \\ e^{-\sqrt{j}} \le c/T_m}} w_j \sim 2cT_m^{-1} \log T_m, \tag{D.14}$$

$$\sum_{\substack{j \le m \\ e^{-\sqrt{j}} \le c/T_m}} e^{-2\sqrt{j}} \sim c^2 T_m^{-2} \log T_m. \tag{D.15}$$

Both branches satisfy stable no-escape, and

$$F_{\boldsymbol{W}_m,>0}(t_m) \sim T_m^{-1} \log T_m, \qquad K_{\boldsymbol{W}_m}(t_m) \sim \frac{\eta^2}{2} T_m^{-2} \log T_m.$$

The pointwise and row-stability margins are uniform in m.

Proof. The tail estimates

$$\sum_{j\ge n} e^{-\sqrt{j}} \sim 2\sqrt{n}\, e^{-\sqrt{n}}, \qquad \sum_{j\ge n} e^{-2\sqrt{j}} \sim \sqrt{n}\, e^{-2\sqrt{n}}$$

are the $\gamma = 1/2$ cases of Lemma D.6. Substituting $n = (\log(T_m/c))^2$ gives Eqs. (D.14) and (D.15). The strict window makes mass beyond m negligible, since its two omitted-to-leading ratios are bounded by

$$\frac{T_m \sqrt{m}\, e^{-\sqrt{m}}}{\log T_m} \quad \text{and} \quad \frac{T_m^2 \sqrt{m}\, e^{-2\sqrt{m}}}{\log T_m},$$

and vanish. Infinite-tail Potter control and bounded mass away from zero give finite-width no-escape; transform constants follow from Theorem B.1.

Finally, $\lambda_{\max} = e^{-1}$, so the pointwise condition holds with $\delta = 1$. Monotone integral comparison yields $\operatorname{tr}(\widehat{\boldsymbol{H}}_m) \le e^{-1} + \int_1^\infty e^{-\sqrt{x}}\, \mathrm{d}x = 5/e$. The certificate Eq. (C.10) is therefore at most $5/(4e) < 1$, uniformly in m. □

The next proposition separates three spectral-transfer failures.

Proposition D.10 (Three distinct spectral boundaries)**.** The following conditioned systems isolate three failures.

(i) If every positive eigenvalue is at least $\lambda_\star > 0$, fixed $\eta > 0, B$ with a pointwise margin give exponentially decaying forcing and memory, rather than a nondegenerate low-spectrum power law.

(ii) For $\lambda_j = e^{-j}$ and target-weighted mass $w_j = (1 - e^{-\tau})e^{-\tau j}$, $\tau > 0$, the cutoff tail is bounded above and below by constant multiples of x^τ, but is not regularly varying.

(iii) There is a triangular target-weighted spectral array for the forcing branch in which every fixed constant-factor cutoff has the candidate power-law limit, yet a moving packet violates the no-escape envelope and changes the training-transform constant.

Proof. For (i), pointwise stability gives

$$q_\eta(\lambda) = 1 - \eta\lambda\left[2 - \left(1 + \frac{1}{B}\right)\eta\lambda\right] \le 1 - \delta\eta\lambda_\star =: \varrho < 1.$$

Thus, when the corresponding total weighted masses are bounded, $F_{\boldsymbol{W},>0}(t) \lesssim \varrho^t$ and $K_{\boldsymbol{W}}(t) \lesssim \varrho^t$, whereas both cutoff masses vanish for $x < \lambda_\star$. A gap therefore violates the nondegenerate tail asymptotic, not the no-escape upper envelope.

For (ii), for all sufficiently small x,

$$A(x) := \sum_{\lambda_j \le x} w_j = e^{-\tau\lceil \log(1/x)\rceil}, \qquad e^{-\tau} x^\tau \le A(x) \le x^\tau.$$

But, with $T_n = e^{n+\phi}$ and $0 < \phi < 1$,

$$\frac{A(1/T_n)}{T_n^{-\tau}} = e^{-\tau(1-\phi)}.$$

The limit depends on the logarithmic phase ϕ, so $T^\tau A(1/T)$ is log-periodic rather than slowly varying. This is the geometric-lacunarity boundary $\gamma = 1$ of Theorem D.5.

For (iii), fix $\tau > 0$, set $T_n = n$, $a_n = n^{-\tau}$, and define background eigenvalues and target-alignment weights

$$\lambda_{n,k} = \frac{k}{n^2}, \qquad w_{n,k} = \left(\frac{k}{n^2}\right)^\tau - \left(\frac{k-1}{n^2}\right)^\tau, \qquad 1 \le k \le n^2.$$

Add one mode at

$$y_n = \frac{\tau}{4}\log n, \qquad \lambda_n^\star = \frac{y_n}{n}, \qquad w_n^\star = n^{-\tau/2}.$$

Writing A_n for the cumulative weighted mass, every compact $K \subset (0, \infty)$, for all sufficiently large n, satisfies

$$\sup_{c \in K} \left| \frac{A_n(c/n)}{a_n} - c^\tau \right| = \sup_{c \in K} \left| \left(\frac{\lfloor cn \rfloor}{n} \right)^\tau - c^\tau \right| \longrightarrow 0,$$

because the moving mode lies above c/n. At its own scale, however,

$$\frac{A_n(y_n/n)}{a_n} \geq n^{\tau/2},$$

which exceeds every bound allowed by the Potter envelope, $C y_n^{\tau+\epsilon} = O((\log n)^{\tau+\epsilon})$. Moreover,

$$\frac{1}{a_n} \int e^{-2n\lambda} \, \mathrm{d}A_n(\lambda) \longrightarrow \Gamma(\tau+1) 2^{-\tau} + 1 :$$

the background gives the Tauberian constant, while the packet contributes $a_n^{-1} w_n^\star e^{-2y_n} = 1$. For the exact stable-SGD version, take $B = 1$, $\eta = 1/2$, and $t_n = 2n$, so $T_n = \eta t_n = n$ and the pointwise margin can be chosen uniformly. The background constant follows from Theorem B.1; for the packet,

$$q_\eta(\lambda_n^\star)^{t_n} = e^{-2n h_\eta(\lambda_n^\star)}, \qquad h_\eta(\lambda) := -\frac{\log q_\eta(\lambda)}{2\eta} = \lambda + O(\lambda^2),$$

and $n h_\eta(\lambda_n^\star) = y_n + o(1)$, again giving the extra 1. □

Proposition D.11 (Pointwise contraction does not imply row stability)**.** Even a one-mode constant schedule can satisfy the pointwise condition with a fixed margin while its Volterra rows exceed one. In particular, take

$$B = 1, \qquad \widehat{\lambda} = 1, \qquad \eta = \frac{7}{10}.$$

Then part *(a)* of Assumption C.2 holds with $\delta = 1/2$, but

$$\sum_{s<t} K_{t,s} \longrightarrow \frac{7}{6} > 1,$$

and, for every $\sigma^2 > 0$, the exact noisy risk grows exponentially.

Proof. Writing $x = \eta\widehat{\lambda}$, the one-mode decay factor and limiting row mass are

$$q = 1 - 2x + \left(1 + \frac{1}{B}\right) x^2, \qquad \lim_{t\to\infty} \sum_{s<t} K_{t,s} = \frac{x/B}{2 - (1 + 1/B)x}.$$

Pointwise contraction requires $x < 2B/(B+1)$, whereas row stability requires the strictly stronger $x < 2B/(B+2)$. At the displayed numerical choice,

$$q = \frac{29}{50}, \qquad \eta\left(1 + \frac{1}{B}\right)\widehat{\lambda} = \frac{7}{5} \leq \frac{3}{2} = 2 - \frac{1}{2}, \qquad \sum_{s<t} K_{t,s} = \frac{7}{6}\left[1 - \left(\frac{29}{50}\right)^t\right].$$

The exact recursion Eq. (C.2) reduces to

$$R_{\sigma,t+1} = \frac{107}{100} R_{\sigma,t} + \frac{49}{100}\sigma^2,$$

so the full second moment diverges whenever $R_{\sigma,0} + \sigma^2 > 0$, although every homogeneous mode contracts. □

Thus a gap removes the low-spectrum mechanism, lacunarity removes regular variation, and spectral mass that shifts with width can invalidate the uniform comparison in the joint width–time limit; pointwise and row stability respectively control modes and accumulated feedback.

Finite-width realizations of the six constructions. We complement the preceding componentwise results with literal finite-width minibatch-SGD trajectories on fresh Gaussian covariates. The left column of each figure reports the population excess risk, evaluated exactly from the parameter error, while the middle and right columns report the corresponding finite-spectrum forcing and one-injection memory together with their cumulative spectral masses. We use intrinsic time $T = \eta t$ and normalize every target so that $\sum_j \lambda_j |\theta_j^\star|^2 = 1$. Multiple widths in the first two rows expose the finite-width cutoff, and multiple batch sizes in the rapid-target row probe the memory-driven tail. At any fixed width, all positive modes eventually relax exponentially; the power-law descriptions below concern the displayed pre-cutoff width–time window.

For the mass-compensated sparse target, set $c_{k^2} = 2k - 1$ for $k \geq 1$ and $c_j = 0$ otherwise. The four rows of Figure 8 have the displayed risk forms

$$\begin{aligned}
\lambda_j = j^{-0.8}, \quad |\theta_j^\star|^2 \propto j^{-0.6} \quad &\Longrightarrow \quad R(T) \asymp T^{-1/2}, \\
\lambda_j = j^{-0.8}, \quad |\theta_j^\star|^2 \propto j^{-0.6} c_j \quad &\Longrightarrow \quad R(T) \asymp T^{-1/2}, \\
\lambda_j = e^{1-\sqrt{j}}, \quad |\theta_j^\star|^2 \propto j^{-3/2} \quad &\Longrightarrow \quad R(T) \asymp T^{-1} (\log T)^{-2}, \\
\lambda_j = e^{-1.4(j-1)}, \quad |\theta_j^\star|^2 \propto \lambda_j^{0.75} \quad &\Longrightarrow \quad R(T) = T^{-7/4} \Phi(\log T),
\end{aligned}$$

where Φ is bounded, nonconstant, and log-periodic. Thus all four risk curves show an overall power-law decrease on the resolved finite-width window, although geometric spacing prevents the last row from having a pure-power asymptotic.

The two rows of Figure 9 isolate the remaining boundary mechanisms:

$$\lambda_j = j^{-0.8}, \quad |\theta_j^\star|^2 \propto e^{-j} \quad \Longrightarrow \quad F(T) = \exp\{-\Theta(T^{5/9})\}, \qquad R(T) \asymp T^{-3/4}.$$

For the ideal gapped construction, set

$$(\lambda_j, |\theta_j^\star|^2) = \begin{cases} (10^{-(j-1)/63}, C j^{-1}), & 1 \leq j \leq 64, \\ (0, 0), & j > 64. \end{cases}$$

Its risk satisfies $R(T) = O(e^{-cT})$ for some $c > 0$. The first row removes slow-direction target energy without removing the target-independent memory tail, whereas the second removes the low-spectrum mechanism itself. Together, the two figures illustrate that the cumulative weighted spectral masses in Theorem 3.1, rather than the pointwise form of either the spectrum or the target alone, determine the long-time curve shape.

Together, Figures 8 and 9 isolate the component-level laws for F and K; the next subsection shows how schedule-dependent injection and Volterra feedback transform these component laws into the observed risk.

D.3 Long-memory schedule transfer with feedback

Having separated the spectral and stability assumptions, we now derive the response to label noise.

Define the first label-noise injection and the exact label-noise excess by $N_t := \sigma^2 \sum_{s<t} K_{t,s}$ and $Z_t := R_{\sigma,t} - R_{0,t}$, respectively. Subtracting the clean and noisy versions of Eq. (C.3) gives the exact identity

$$Z_t = N_t + \sum_{s<t} K_{t,s} Z_s. \tag{D.16}$$

For the continuum comparison used below, define the counterpart of N_t by

$$N_{\mathrm{c}}(T) := \sigma^2 \int_0^T \frac{k(T-u)}{r(u)} \, \mathrm{d}u. \tag{D.17}$$

First isolate a feedback estimate shared by forward transfer and converse.

Lemma D.12 (Lower-order feedback below the memory ceiling)**.** Assume Eqs. (D.3), (D.4) and (D.16) with $K_{t,s} \geq 0$. For $0 < q < q_K$, if either N_t or Z_t is asymptotic to $T_t^{-q} L(T_t)$, where L is eventually positive and slowly varying, then

$$Z_t \sim N_t \sim T_t^{-q} L(T_t).$$

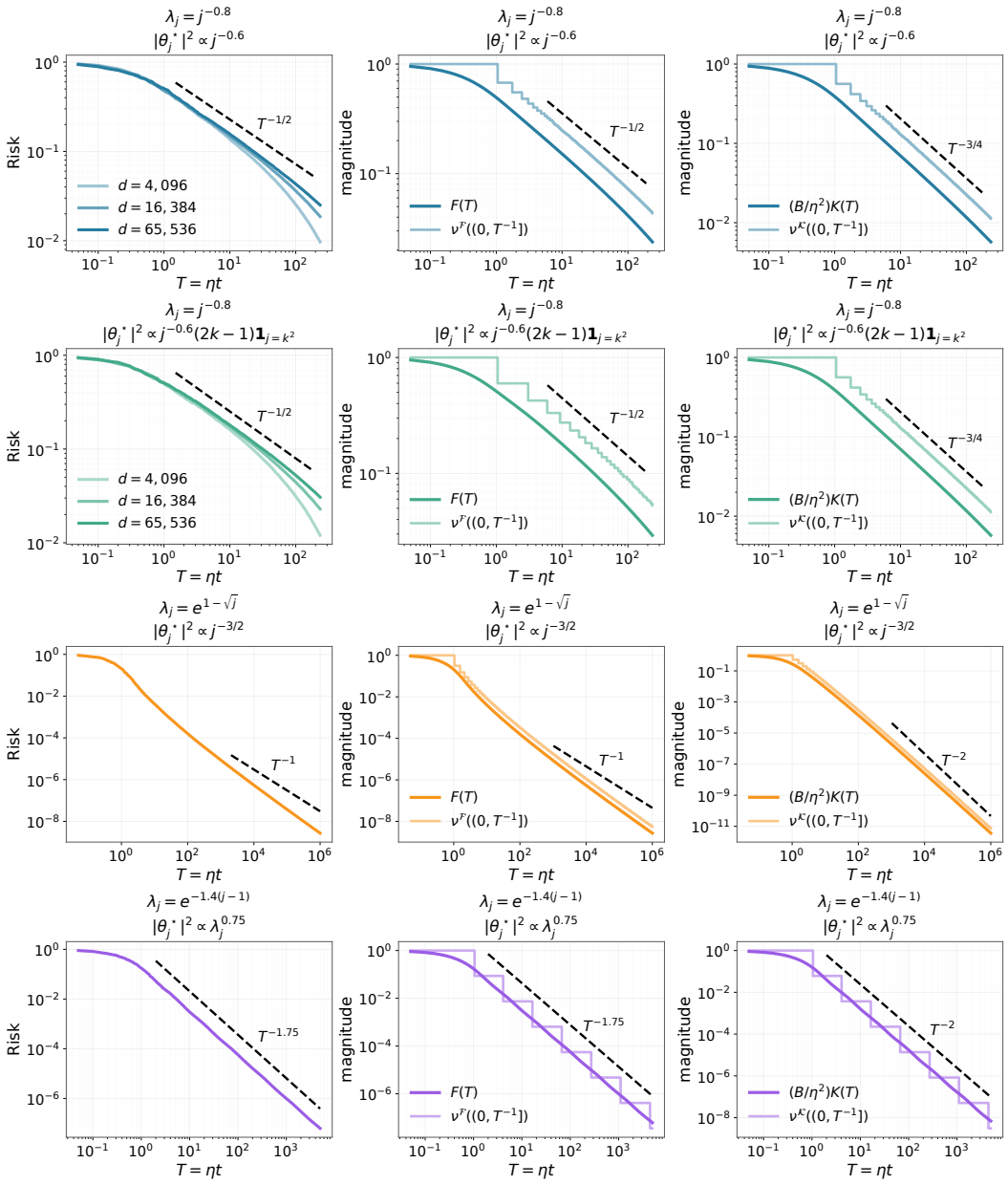


Figure 8: Finite-width realizations of minibatch-SGD excess risk. Across the four constructions, the risk curves exhibit an overall power-law decay over the displayed finite-width window, including logarithmic corrections and log-periodic modulation.

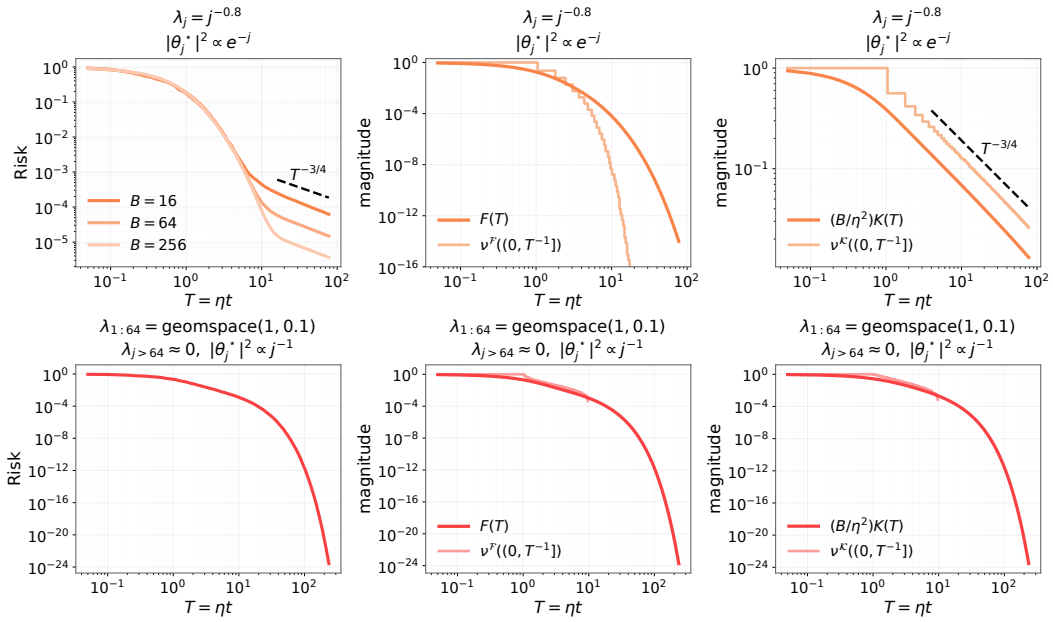


Figure 9: Finite-width realizations of minibatch-SGD excess risk. The rapidly decaying target develops a memory-driven $T^{-3/4}$ tail, whereas the spectrally gapped construction relaxes exponentially.

Proof. Positivity in Eq. (D.16) gives $0 \le N_t \le Z_t$. For fixed S, sharp bulk convergence and regular variation of k give $\sum_{s<S} K_{t,s} Z_s = O_S(k(T_t))$. The remaining feedback satisfies

$$0 \le Z_t - N_t \le O_S(k(T_t)) + \frac{\sup_{s \ge S} Z_s}{\sigma^2} N_t. \tag{D.18}$$

If $Z_t \sim T_t^{-q} L(T_t)$, then $Z_t \to 0$, $k(T_t) = o(Z_t)$, and $N_t \le Z_t$. Divide Eq. (D.18) by Z_t, then send $t \to \infty$ and $S \to \infty$, obtaining $Z_t - N_t = o(Z_t)$.

If instead the equivalent holds for N_t, its row sum tends to zero; after a finite prefix it is at most $1/2$, so induction bounds Z, and $Z_t \le N_t + (N_t/\sigma^2) \sup_s Z_s = O(N_t)$. Thus $Z_t \to 0$, and division of Eq. (D.18) by N_t, using $k(T_t) = o(N_t)$, finishes the proof. □

For an eventually positive ratio r that is regularly varying with exponent ϑ, write its slowly varying factor and cumulative injection as $L_r(u) := u^{-\vartheta} r(u)$, $V(T) := \int_0^T r(u)^{-1}\, \mathrm{d}u$. For the fixed-system endpoint branches below, also write $V_\infty := \lim_{T\to\infty} V(T)$ and denote the extended-valued endpoint amplitude by

$$C := \sum_{s \ge 0} \frac{T_{s+1} - T_s}{r_s} \left(\sigma^2 + Z_s\right) \in [0, \infty].$$

Theorem D.13 (Sharp discrete schedule transfer)**.** Fix $\sigma^2 > 0$, assume Assumptions D.1 and D.3, and retain $r(u) = u^\vartheta L_r(u)$ and the preceding V. In a fixed infinite-spectrum system, also retain V_∞ and C.

If $\vartheta < 1$, then

$$N_t \sim \sigma^2 \operatorname{Beta}(1-\vartheta, 1-q_K) \frac{T_t k(T_t)}{r(T_t)} \sim \sigma^2 \operatorname{Beta}(1-\vartheta, 1-q_K) T_t^{-(\vartheta+q_K-1)} \frac{L_K(T_t)}{L_r(T_t)}. \tag{D.19}$$

For a fixed infinite-spectrum system, if $\vartheta = 1$ and $V(T) \to \infty$, then

$$N_t \sim \sigma^2 k(T_t) V(T_t). \tag{D.20}$$

For such a fixed system, if instead $V_\infty < \infty$, including every $\vartheta > 1$,

$$N_t \sim \sigma^2 V_\infty k(T_t). \tag{D.21}$$

For the full response, assume in addition part *(b)* of Assumption C.2. If $1 - q_{\mathcal{K}} < \vartheta < 1$, then

$$Z_t \sim N_t. \tag{D.22}$$

For a fixed infinite-spectrum system, the same conclusion holds when $\vartheta = 1$ and $V(T) \to \infty$. For such a fixed system, if $V_\infty < \infty$, including every $\vartheta > 1$, then

$$Z_t \sim Ck(T_t), \qquad C < \infty. \tag{D.23}$$

Thus every fixed-system branch with $\vartheta \geq 1$ has noise exponent $q_{\mathcal{K}}$, although at $\vartheta = 1$ its slowly varying factor depends on whether V converges.

Proof. For $\vartheta < 1$, split N_t at $T = T_t$ into

$$T_{s+1} \leq \varepsilon T, \qquad \varepsilon T < T_{s+1} < (1-\varepsilon)T, \qquad T_s \geq (1-\varepsilon)T.$$

On the middle region, sharp bulk transfer, schedule convergence, and fine mesh give

$$\frac{r(T)}{Tk(T)} \sum_{\varepsilon T < T_{s+1} < (1-\varepsilon)T} \frac{T_{s+1} - T_s}{r_s} \Psi_{t,s} \longrightarrow \int_\varepsilon^{1-\varepsilon} x^{-\vartheta}(1-x)^{-q_{\mathcal{K}}} \, \mathrm{d}x. \tag{D.24}$$

For the early region, Eq. (D.4) and the uniform Potter bounds give

$$\sum_{T_{s+1} \leq \varepsilon T} \frac{T_{s+1} - T_s}{r_s} \Psi_{t,s} \lesssim k(T) \int_0^{\varepsilon T} \frac{\mathrm{d}u}{r(u)}.$$

Lemma D.2 yields, for every sufficiently small $\xi > 0$,

$$\limsup_{t \to \infty} \frac{r(T)\,\text{early}}{Tk(T)} \lesssim \varepsilon^{1-\vartheta-\xi}. \tag{D.25}$$

On the recent region, uniform regular variation and eventual monotonicity give $r_s^{-1} \lesssim r(T)^{-1}$. Applying Eq. (D.8) with $x = \varepsilon T$ therefore yields

$$\limsup_{t \to \infty} \frac{r(T)\,\text{recent}}{Tk(T)} \lesssim \varepsilon^{1-q_{\mathcal{K}}-\xi}, \tag{D.26}$$

with the mesh contribution vanishing by Eq. (D.7). Sending first $t \to \infty$, then $\varepsilon \downarrow 0$, in Eqs. (D.24) to (D.26) gives

$$\int_0^1 x^{-\vartheta}(1-x)^{-q_{\mathcal{K}}} \, \mathrm{d}x = \mathrm{Beta}(1-\vartheta, 1-q_{\mathcal{K}}),$$

proves Eq. (D.19).

For $1 - q_{\mathcal{K}} < \vartheta < 1$, Eq. (D.19) has exponent $-q$, $q = \vartheta + q_{\mathcal{K}} - 1 \in (0, q_{\mathcal{K}})$, so Lemma D.12 gives Eq. (D.22).

For the remaining fixed-system branches, first let $\vartheta = 1$. The cumulative injection $V(T) = \int_0^T \mathrm{d}u/r(u)$ is slowly varying. If $V(T) \to \infty$, boundary Karamata theory gives

$$\frac{T}{r(T)V(T)} \longrightarrow 0, \qquad \frac{V(\varepsilon T)}{V(T)} \longrightarrow 1.$$

In the same split, the early part normalized by $k(T)V(T)$ tends to one as $\varepsilon \downarrow 0$; bulk and recent parts vanish because $T/[r(T)V(T)] \to 0$. This proves Eq. (D.20). The boundedness argument of Lemma D.12, followed by Eq. (D.18), gives $Z_t \to 0$ and $Z_t \sim N_t$ because $k(T_t) = o(N_t)$.

If $V_\infty < \infty$, the following prefix–tail argument with $Z_s = 0$ gives Eq. (D.21). Row stability bounds Z, so $(T_{s+1} - T_s)(\sigma^2 + Z_s)/r_s$ is summable. For every fixed s, Eq. (D.4) gives $\Psi_{t,s}/k(T_t) \to 1$. For the remaining tail, first keep $T_{s+1} \leq (1-\varepsilon)T_t$, where Potter bounds multiply its small mass by an ε-dependent constant; the recent endpoint is $o(k(T_t))$ since $T/r(T) \to 0$. Prefix–tail limits followed by $\varepsilon \downarrow 0$ prove Eq. (D.23), also for $\vartheta > 1$. □

At $\vartheta = 1 - q_{\mathcal{K}}$, Eq. (D.19) is slowly varying and dominates positive-power clean decay unless a triangular normalization vanishes. Below it, the row mass N_t/σ^2 grows polynomially, contradicting part *(b)* of Assumption C.2; this is inadmissible, not a stable growing-loss regime.

D.4 INTEGRABLE-MEMORY SCHEDULE TRANSFER

For $q_{\mathcal{K}} > 1$, recent injections see total memory mass while early ones see the tail of k. We prove the sharp continuum endpoint asymptotic, then transfer its order to the exact gap. The discrete step requires all-age memory because low-spectrum tails do not determine bounded-age mass; these are the quantified conditions underlying the integrable-memory part of Theorem 5.1.

For a nonnegative locally bounded memory profile k and injection density $1/r$, write their total masses as $\bar{k} := \int_0^\infty k(v)\,\mathrm{d}v$ and $V_\infty := \int_0^\infty \mathrm{d}u/r(u)$, allowing the value $+\infty$ before integrability is assumed. The proposition gives exact constants for N_{c}; the discrete theorem uses only its order consequence.

Proposition D.14 (Two-endpoint direct response under integrable memory)**.** Let $k \geq 0$ be locally bounded and eventually positive, and let r be positive with $1/r$ locally bounded and eventually positive. Suppose k is eventually nonincreasing,

$$k(T) \sim T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T), \qquad q_{\mathcal{K}} > 1, \qquad \bar{k} \in (0, \infty),$$

and suppose $1/r$ is eventually monotone and regularly varying with exponent $-\vartheta$. Then

$$N_{\mathrm{c}}(T) \sim \sigma^2 \begin{cases} \bar{k}/r(T), & \vartheta < q_{\mathcal{K}}, \\ \bar{k}/r(T) + V_\infty k(T), & \vartheta = q_{\mathcal{K}}, \\ V_\infty k(T), & \vartheta > q_{\mathcal{K}}, \end{cases} \tag{D.27}$$

where $V_\infty < \infty$ in the last two branches. In particular,

$$N_{\mathrm{c}}(T) \asymp \frac{1}{r(T)} + k(T) \int_0^T \frac{\mathrm{d}u}{r(u)}.$$

Thus the direct continuum response has power exponent $\min\{\vartheta, q_{\mathcal{K}}\}$. At $\vartheta = 0$, the direct response has zero power exponent. If $\vartheta < 0$, it grows with a positive power.

Proof. Split the direct convolution into recent and old endpoints. In age $v = T - u$, fix $A > 0$; uniform regular variation gives

$$\int_0^A \frac{k(v)}{r(T-v)}\,\mathrm{d}v \sim \frac{\bar{k}_A}{r(T)}, \qquad \bar{k}_A := \int_0^A k(v)\,\mathrm{d}v.$$

On $A \leq v \leq T/2$, Potter bounds and integrability of k give an upper bound of order $r(T)^{-1} \int_A^\infty k(v)\,\mathrm{d}v$. On the old-injection half, regular variation of k gives

$$\int_0^{T/2} \frac{k(T-u)}{r(u)}\,\mathrm{d}u \lesssim k(T) \int_0^{T/2} \frac{\mathrm{d}u}{r(u)}.$$

For $\vartheta < 1$, Karamata and $Tk(T) \to 0$ make this $o(1/r(T))$. At $\vartheta = 1$, it is $T^{1-q_{\mathcal{K}}}$ times a slow factor and still vanishes. For $1 < \vartheta < q_{\mathcal{K}}$, $V_\infty < \infty$ and $k(T) = o(1/r(T))$. Letting $A \to \infty$ proves the first branch of Eq. (D.27).

For $\vartheta > q_{\mathcal{K}}$, exchanging $(u, 1/r, V_\infty)$ with $(v, k, \bar{k})$ gives $V_\infty k(T)$. At equality both integrable endpoints remain, while Potter bounds remove the middle, giving $\bar{k}/r(T) + V_\infty k(T)$. Multiplication by σ^2 proves Eq. (D.27); the order-level display follows by the same three cases.

At $\vartheta = 0$, Eq. (D.27) gives zero power exponent, while for $\vartheta < 0$ it makes $N_{\mathrm{c}}(T)/\sigma^2$, the continuum row mass, diverge. □

We now quantify the discrete integrable-memory part of Theorem 5.1, using $(H_{t,s}, \Psi_{t,s}, N_t, Z_t)$ defined above. For an eventually positive regularly varying ratio of index ϑ, retain the slowly varying factor $L_r(u) := u^{-\vartheta} r(u)$.

Assumption D.15 (Integrable-memory all-age kernel control)**.** Consider either one fixed limiting dynamics or a width–time triangular sequence in Definition D3. The two-time kernel has the following properties:

(a) For some $q_{\mathcal{K}} > 1$, the reference profile $k : [0,\infty) \to (0,\infty)$ is locally bounded and nonincreasing, with

$$k(T) \sim T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T), \qquad 0 < \int_0^\infty k(v)\,\mathrm{d}v < \infty.$$

Here $L_{\mathcal{K}}$ is eventually positive and slowly varying.

(b) For constants $0 < c < C < \infty$,

$$c\,k(H_{t,s}) \le \Psi_{t,s} \le C\,k(H_{t,s}), \qquad 0 \le s < t. \tag{D.28}$$

(c) The intrinsic mesh is bounded:

$$h_t = \max_{s<t}(T_{s+1} - T_s) \le h_\star < \infty. \tag{D.29}$$

In a triangular sequence, c, C, and $h_\star$ are uniform over the width and horizon.

Theorem D.16 (Exact discrete transfer under integrable memory)**.** Fix $\sigma^2 > 0$, assume Assumptions D.1 and D.15, and use the slowly varying factor L_r defined immediately above. Assume also that $1/r$ is locally bounded. For the exponent conclusions below, additionally require

$$\frac{\log(r_m(T_t)^{-1})}{\log T_t} \longrightarrow -\vartheta, \qquad \frac{\log\int_0^{T_t} \mathrm{d}u/r_m(u)}{\log T_t} \longrightarrow \max\{1-\vartheta, 0\}$$

along a triangular sequence; no extra normalization is needed for a fixed system. Finally assume part *(b)* of Assumption C.2, with its constant $\kappa < 1$. Then, with all comparisons uniform in the triangular setting,

$$Z_t \asymp N_t \asymp \sigma^2\left[\frac{1}{r(T_t)} + k(T_t)\int_0^{T_t}\frac{\mathrm{d}u}{r(u)}\right] \asymp \sigma^2\left[\frac{T_t^{-\vartheta}}{L_r(T_t)} + T_t^{-q_{\mathcal{K}}}L_{\mathcal{K}}(T_t)\int_0^{T_t}\frac{\mathrm{d}u}{r(u)}\right]. \tag{D.30}$$

Consequently, for a fixed system, and along a triangular sequence under the displayed normalization conditions,

$$Z_t = T_t^{-\min\{\vartheta, q_{\mathcal{K}}\}+o(1)}, \qquad \vartheta > 0.$$

At $\vartheta = 0$ the exact noisy–clean gap has no positive power exponent. For a fixed system, $\vartheta < 0$ makes the row mass grow and contradicts part *(b)* of Assumption C.2; the same holds along a triangular sequence whenever its width-dependent normalization does not cancel that growth.

Proof. Compare the cell sum with the continuum convolution, sandwich the exact gap, then read off the exponent. Regular variation and positivity of nonincreasing k give $C_\star < \infty$ such that $k(v) \le C_\star k(v+h)$ for $v \ge 0$ and $0 \le h \le h_\star$. because the ratio tends uniformly to one at infinity and positivity controls compact intervals. For $u \in [T_s, T_{s+1})$, age $T_t - u$ lies between $H_{t,s}$ and $H_{t,s} + h_\star$. Since $1/r$ is constant on this cell, summing the resulting upper and lower bounds gives

$$\sum_{s<t}\frac{T_{s+1}-T_s}{r_s}k(H_{t,s}) \asymp \int_0^{T_t}\frac{k(T_t-u)}{r(u)}\,\mathrm{d}u.$$

Together with Eq. (D.28), this proves

$$N_t \asymp \sigma^2\int_0^{T_t}\frac{k(T_t-u)}{r(u)}\,\mathrm{d}u.$$

For a fixed system, Proposition D.14 gives the second comparison in Eq. (D.30). Uniformly in the triangular setting, splitting at $T_t/2$ and applying common Potter bounds gives

$$\int_0^{T_t}\frac{k(T_t-u)}{r(u)}\,\mathrm{d}u \lesssim \frac{1}{r(T_t)} + k(T_t)\int_0^{T_t}\frac{\mathrm{d}u}{r(u)}.$$

A fixed recent-age interval gives the $1/r(T_t)$ lower bound. If $1/r$ decreases, the old half gives the second lower bound since $\int_0^{T_t/2}\mathrm{d}u/r(u) \gtrsim \int_0^{T_t}\mathrm{d}u/r(u)$. If it increases, then $\vartheta \le 0$, and Lemma D.2 with $Tk(T) \to 0$ absorbs the second term into the first. Constants are width-independent.

For the exact gap, Eq. (D.16) and positivity give $Z_t \geq N_t$. Let $C_0 := \kappa\sigma^2/(1-\kappa)$. An induction using part *(b)* of Assumption C.2 gives $0 \leq Z_t \leq C_0$ for every t. Hence

$$Z_t = \sum_{s<t} K_{t,s}(\sigma^2 + Z_s) \leq (\sigma^2 + C_0) \sum_{s<t} K_{t,s} = \frac{1}{1-\kappa} N_t.$$

Thus $N_t \leq Z_t \leq N_t/(1-\kappa)$, proving the first comparison.

For $\vartheta > 0$, recent injection leads below $q_{\mathcal{K}}$, old injection above it, and both remain at equality, giving exponent $\min\{\vartheta, q_{\mathcal{K}}\}$. At $\vartheta = 0$ the recent term excludes positive-power decay; below zero it grows and contradicts the row bound under the stated normalization. □

If the centered clean loss has exponent $q_0 > 0$, positivity of $R_\sigma - R_{\mathrm{app}} = (R_0 - R_{\mathrm{app}}) + Z$ and Eq. (D.30) give total exponent $\min\{q_0, q_{\mathcal{N}}(\vartheta)\}$. This is the integrable-memory input used in Theorem 5.1.

Under the corresponding row-stable fixed-dynamics pure-power hypotheses, the following table collects the schedule boundaries.

For the PLRF IM regime, where $\alpha > 1/2$, the DE theorem for the LM and IM regimes supplies the all-age profile $k(v) = (1+v)^{-2+1/(2\alpha)}$ at the resolvent-DE level. Applying the exact finite-width result additionally requires the fixed-sequence empirical-to-DE transfer in Assumption A1, together with bounded mesh and empirical row stability.

D.5 From the exact noisy–clean gap back to B/η

Return to $0 < q_{\mathcal{K}} < 1$ and remove schedule regular variation from the reverse direction.

If needed, extend k nonnegatively and locally integrably over bounded ages, without changing its tail, and use Eq. (D.17).

With Lemma D.12, the next output-driven lemma supplies the discrete step, without assuming regular variation of $1/r$.

Lemma D.17 (Discrete-to-continuous convolution comparison from the output asymptotics)**.** Work with a fixed limiting sequence and assume Assumption D.3, with $T_t \uparrow \infty$. Let r be positive and suppose its step interpolation $1/r$ is eventually nonincreasing. If, for some $0 < q < q_{\mathcal{K}}$ and eventually positive slowly varying L,

$$N_t \sim T_t^{-q} L(T_t), \tag{D.31}$$

then the continuum comparison in Eq. (D.17) satisfies

$$N_{\mathrm{c}}(T) \sim T^{-q} L(T) \qquad (T \to \infty). \tag{D.32}$$

Proof. Use the output to bound recent schedule height, remove both recent endpoints, compare old cells with the continuum integral, then interpolate. Write $\widetilde{N}_t = N_t/\sigma^2$, $T = T_t$; since $k(T) \to 0$, the mesh condition gives $h_t/T_t \to 0$. For $0 < \varepsilon < 1/8$, cells with $T_s \in [T/4, T/2]$ have length $T/4 + o(T)$. There, Eq. (D.4) and the uniform convergence theorem for regular variation give $\Psi_{t,s} \geq ck(T)$ for some $c > 0$. Eventual monotonicity yields the output-driven bound

$$\frac{Tk(T)}{r((1-2\varepsilon)T)} \leq C_0 \widetilde{N}_t. \tag{D.33}$$

Let $\widetilde{N}_t^{\mathrm{rec}}$ be the contribution of the exact cells with $H_{t,s} < \varepsilon T$. Their left endpoints eventually exceed $(1-2\varepsilon)T$, so Eq. (D.8) with $x = 2\varepsilon T$ gives

$$\widetilde{N}_t^{\mathrm{rec}} \leq \frac{1}{r((1-2\varepsilon)T)} \left\{ h_t G_m(0) + \int_0^{2\varepsilon T} G_m(v)\,\mathrm{d}v \right\}.$$

Combining Eqs. (D.6), (D.7) and (D.33), for every sufficiently small $\xi > 0$, gives

$$\limsup_{t\to\infty} \frac{\widetilde{N}_t^{\mathrm{rec}}}{\widetilde{N}_t} \leq C_\xi \varepsilon^{1-q_{\mathcal{K}}-\xi}.$$

Karamata integration of k, together with Eq. (D.33), gives the same estimate for the recent part of the continuum comparison:

$$\limsup_{t\to\infty} \frac{\int_{(1-\varepsilon)T}^{T} k(T-u)\,\mathrm{d}u/r(u)}{\widetilde{N}_t} \le C_\xi \varepsilon^{1-q_{\mathcal{K}}-\xi}. \tag{D.34}$$

For old parts, sharp bulk convergence gives $\Psi_{t,s}/k(H_{t,s}) \to 1$ uniformly when $H_{t,s} \ge \varepsilon T$, and

$$\sup_{H_{t,s}\ge \varepsilon T} \left| \frac{\int_{T_s}^{T_{s+1}} k(T-u)\,\mathrm{d}u}{(T_{s+1}-T_s)k(H_{t,s})} - 1 \right| \longrightarrow 0,$$

because $h_t/T \to 0$ and convergence is uniform away from zero age. Multiplication by $1/r_s \ge 0$ makes the old-part difference $o(\widetilde{N}_t)$, except at most one crossing cell. The remaining contribution from this cell is bounded by

$$\frac{h_t}{r((1-2\varepsilon)T)} \sup_{v\in[\varepsilon T - h_t, \varepsilon T + h_t]} k(v) = o(\widetilde{N}_t)$$

for fixed ε, by the output bound, mesh condition, and regular variation. The two recent estimates followed by $\varepsilon \downarrow 0$ prove

$$N_{\mathrm{c}}(T_t) \sim N_t. \tag{D.35}$$

For $T_t \le U < T_{t+1}$, the mesh condition gives

$$T_{t+1} - T_t \le h_{t+1} = o(T_{t+1}k(T_{t+1})) = o(T_{t+1}),$$

so $U/T_t \to 1$; on ages at least εT_t, $k(U-u)/k(T_t-u) \to 1$ uniformly. The new interval is bounded by

$$\frac{1}{r_t}\int_0^{U-T_t} k(v)\,\mathrm{d}v \le \frac{1}{r((1-2\varepsilon)T_t)} \int_0^{2\varepsilon T_t} k(v)\,\mathrm{d}v \le C_\xi \varepsilon^{1-q_{\mathcal{K}}-\xi}\widetilde{N}_t,$$

using Eq. (D.33). The other recent pieces obey the same estimate as Eq. (D.34). Therefore

$$\lim_{t\to\infty} \sup_{T_t \le U < T_{t+1}} \left| \frac{N_{\mathrm{c}}(U)}{N_{\mathrm{c}}(T_t)} - 1 \right| = 0.$$

Combining this with Eqs. (D.31) and (D.35) and uniform convergence of the slowly varying factor proves Eq. (D.32). □

Proposition D.18 (Gap-to-ratio Tauberian converse)**.** Let $k \ge 0$ be locally integrable, have finite Laplace transform for every positive argument, and satisfy $k(T) \sim T^{-q_{\mathcal{K}}} L_{\mathcal{K}}(T)$, where $0 < q_{\mathcal{K}} < 1$ and $L_{\mathcal{K}}$ is eventually positive and slowly varying. Let r be positive, with $1/r$ locally integrable and eventually nonincreasing, and suppose $1/r$ has a finite Laplace transform for every positive argument. Retain the continuum response N_{c} from Eq. (D.17). For $0 < q < q_{\mathcal{K}}$ and eventually positive slowly varying L,

$$N_{\mathrm{c}}(T) \sim T^{-q}L(T)$$

if and only if

$$\frac{1}{r(T)} \sim \frac{1}{\sigma^2\,\mathrm{Beta}(q_{\mathcal{K}}-q, 1-q_{\mathcal{K}})} T^{q_{\mathcal{K}}-1-q} \frac{L(T)}{L_{\mathcal{K}}(T)}. \tag{D.36}$$

Equivalently,

$$r(T) \sim \sigma^2\,\mathrm{Beta}(q_{\mathcal{K}}-q, 1-q_{\mathcal{K}}) T^{1-q_{\mathcal{K}}+q} \frac{L_{\mathcal{K}}(T)}{L(T)}. \tag{D.37}$$

At the ceiling $q = q_{\mathcal{K}}$, the converse fails.

Proof. The reverse implication is beta convolution. Conversely, positivity and Tonelli give, with hats denoting Laplace transforms,

$$\widehat{N}_{\mathrm{c}}(s) = \sigma^2 \widehat{(1/r)}(s)\widehat{k}(s).$$

Karamata's Laplace theorem gives, as $s \downarrow 0$,

$$\widehat{N}_{\mathrm{c}}(s) \sim \Gamma(1-q)s^{q-1}L(1/s),$$
$$\widehat{k}(s) \sim \Gamma(1-q_{\mathcal{K}})s^{q_{\mathcal{K}}-1}L_{\mathcal{K}}(1/s).$$

Consequently,

$$\widehat{(1/r)}(s) \sim \frac{\Gamma(1-q)}{\sigma^2\Gamma(1-q_{\mathcal{K}})}s^{q-q_{\mathcal{K}}}\frac{L(1/s)}{L_{\mathcal{K}}(1/s)}.$$

For $V(T) = \int_0^T \mathrm{d}u/r(u)$, the Laplace–Stieltjes Tauberian theorem yields

$$V(T) \sim \frac{\Gamma(1-q)}{\sigma^2\Gamma(1-q_{\mathcal{K}})\Gamma(1+q_{\mathcal{K}}-q)}T^{q_{\mathcal{K}}-q}\frac{L(T)}{L_{\mathcal{K}}(T)}.$$

The monotone-density theorem then gives Eq. (D.36); the beta identity gives Eq. (D.37). These standard regular-variation results are collected in Bingham et al. (1989, Theorems 1.7.1–1.7.2).

At $q = q_{\mathcal{K}}$, the family $r_a(u)^{-1} = (a-1)(1+u)^{-a}$, $a > 1$, has unit total injection and distinct tails but $N_{\mathrm{c},a}(T) \sim \sigma^2 k(T)$; the ceiling is not identifiable. □

Theorem D.19 (Exact fixed-limit discrete gap-to-ratio identification)**.** Fix one infinite-spectrum limiting sequence with $T_t \uparrow \infty$, $0 < T_{s+1} - T_s < \infty$, and $\sigma^2 > 0$. Assume Assumption D.3 and Eq. (D.16), with the width index suppressed and all limits taken as $t \to \infty$. Assume $K_{t,s} \geq 0$, and let r be a positive step interpolation such that $1/r$ is finite, locally integrable, and eventually nonincreasing. For $0 < q < q_{\mathcal{K}}$ and eventually positive slowly varying L,

$$Z_t \sim T_t^{-q}L(T_t) \quad \Longleftrightarrow \quad \frac{1}{r(T)} \sim \frac{1}{\sigma^2\,\mathrm{Beta}(q_{\mathcal{K}}-q, 1-q_{\mathcal{K}})}T^{q_{\mathcal{K}}-1-q}\frac{L(T)}{L_{\mathcal{K}}(T)}. \tag{D.38}$$

Proof. If the equivalent holds for Z_t, Lemmas D.12 and D.17 transfer it first to N_t, then to $N_{\mathrm{c}}(T)$. Local integrability, the kernel tail, and eventual schedule monotonicity give finite Laplace transforms, so Proposition D.18 yields the displayed asymptotic for $1/r$.

Conversely, this asymptotic makes r regularly varying with exponent $1 - q_{\mathcal{K}} + q \in (1-q_{\mathcal{K}}, 1)$, supplying the fixed-sequence uniform convergence, Potter, and Karamata inputs to Theorem D.13. Its direct part and the beta identity give $N_t \sim T_t^{-q}L(T_t)$, and Lemma D.12 gives the same for Z_t. □

Remark D.20 (Why gap-to-ratio identification is regime-qualified)**.** Eventual monotonicity is essential: sparse bounded spikes can destroy pointwise regular variation of $1/r$ while adding only $O(k(T)) = o(T^{-q})$. So is the fixed limit: a triangular diagonal $r_m(u)^{-1} = a_m f(u/T_m)$, with arbitrary positive decreasing f, can be normalized to the same terminal response. Triangular identification needs locally uniform output at all constant-factor subhorizons.

D.6 Full preserve–change–destroy classification

Combining the gap estimates with the clean-risk asymptotics gives the preserve/change/destroy classification.

For the classification below, abbreviate T_t by T. For a regularly varying ratio of index ϑ, retain $r(T) = T^{\vartheta}L_r(T)$ and $V(T) = \int_0^T r(u)^{-1}\,\mathrm{d}u$ from the preceding notation.

Theorem D.21 (Full regular-variation preserve–change–destroy classification)**.** Fix one infinite-spectrum dynamics and suppose

$$R_0(T) - R_{\mathrm{app}} \sim T^{-q_0}L_0(T), \qquad q_0 > 0,$$

where L_0 is eventually positive and slowly varying.

In long memory, assume the hypotheses of Theorem D.13. For $1 - q_{\mathcal{K}} < \vartheta < 1$,

$$R_\sigma(T) - R_0(T) \sim \sigma^2\,\mathrm{Beta}(1-\vartheta, 1-q_{\mathcal{K}})T^{-(\vartheta+q_{\mathcal{K}}-1)}\frac{L_{\mathcal{K}}(T)}{L_r(T)}.$$

For $\vartheta = 1$, the gap is asymptotic to $\sigma^2 k(T)V(T)$ if $V(T) \to \infty$, and to $Ck(T)$, with the positive constant C from Eq. (D.23), if $V_\infty < \infty$; the latter also holds for $\vartheta > 1$. Thus

$$q_{\mathcal{N}}(\vartheta) = \begin{cases} \vartheta + q_{\mathcal{K}} - 1, & 1 - q_{\mathcal{K}} < \vartheta < 1, \\ q_{\mathcal{K}}, & \vartheta \geq 1. \end{cases}$$

In integrable memory, assume the hypotheses of Theorem D.16. For $\vartheta > 0$,

$$R_\sigma(T) - R_0(T) \asymp \sigma^2 \left[\frac{1}{r(T)} + k(T) \int_0^T \frac{\mathrm{d}u}{r(u)} \right], \qquad q_{\mathcal{N}}(\vartheta) = \min\{\vartheta, q_{\mathcal{K}}\}.$$

In every decaying branch,

$$R_\sigma(T) - R_{\mathrm{app}} = T^{-\min\{q_0, q_{\mathcal{N}}(\vartheta)\} + o(1)}.$$

Hence the schedule changes, preserves, or is critical according as $q_{\mathcal{N}}(\vartheta)$ is below, above, or equal to q_0. At equality positivity prevents cancellation; in branches with exact equivalents, the slow factors add. At LM-interior equality $q_0 = q_{\mathcal{N}}(\vartheta)$, for example,

$$R_\sigma(T) - R_{\mathrm{app}} \sim T^{-q_0} \left[L_0(T) + \sigma^2 \operatorname{Beta}(1 - \vartheta, 1 - q_{\mathcal{K}}) \frac{L_{\mathcal{K}}(T)}{L_r(T)} \right].$$

At an IM equality, the available order-level statement is

$$R_\sigma(T) - R_{\mathrm{app}} \asymp T^{-q_0} L_0(T) + \sigma^2 \left[\frac{1}{r(T)} + k(T) \int_0^T \frac{\mathrm{d}u}{r(u)} \right].$$

At the LM edge $\vartheta = 1 - q_{\mathcal{K}}$ and the IM edge $\vartheta = 0$, no positive-power gap decay is possible, so positive-power convergence of the total centered loss to R_{app} is destroyed. Below either edge, row stability fails.

The long-memory interior equivalent and the integrable-memory order comparison extend to width–time triangular sequences under the uniform conditions of Theorems D.13 and D.16. The long-memory endpoint equivalents for $\vartheta \geq 1$ are fixed-system statements. An exponent-only triangular classification additionally requires subpower diagonal normalization of the clean factor and of the relevant slow factor; without it, width-dependent amplitudes must remain explicit.

Proof of Theorem D.21. Write $Z(T_t) := Z_t$, with all asymptotics on the grid $T = T_t$. For common clean/noisy floor R_{app}, suppose

$$R_0(T) - R_{\mathrm{app}} \sim T^{-q_0} L_0(T), \qquad q_0 > 0,$$

with L_0 eventually positive and slowly varying. The positive decomposition

$$R_\sigma(T) - R_{\mathrm{app}} = R_0(T) - R_{\mathrm{app}} + Z(T),$$

reduces the result to exponent comparison and the stability boundary.

For $1 - q_{\mathcal{K}} < \vartheta < 1$, write $r(u) = u^\vartheta L_r(u)$, and set $q_{\mathcal{N}}(\vartheta) := \vartheta + q_{\mathcal{K}} - 1$. By Eqs. (D.19) and (D.22),

$$Z(T) \sim \sigma^2 \operatorname{Beta}(1 - \vartheta, 1 - q_{\mathcal{K}}) T^{-q_{\mathcal{N}}(\vartheta)} \frac{L_{\mathcal{K}}(T)}{L_r(T)}.$$

Thus the noisy exponent is $\min\{q_0, q_{\mathcal{N}}(\vartheta)\}$: it changes below q_0, preserves the clean equivalent above, and at equality

$$R_\sigma(T) - R_{\mathrm{app}} \sim T^{-q_0} \left[L_0(T) + \sigma^2 \operatorname{Beta}(1 - \vartheta, 1 - q_{\mathcal{K}}) \frac{L_{\mathcal{K}}(T)}{L_r(T)} \right].$$

preserves the exponent, with a renormalized slow factor if the gap matters.

For $\vartheta = 1$, put $V(T) = \int_0^T r(u)^{-1} \, \mathrm{d}u$. The two branches are $Z(T) \sim \sigma^2 k(T) V(T)$ if $V(T) \to \infty$, and $Z(T) \sim Ck(T)$ if $V_\infty < \infty$, by Eqs. (D.20), (D.22) and (D.23). Both have exponent $q_{\mathcal{K}}$; e.g.

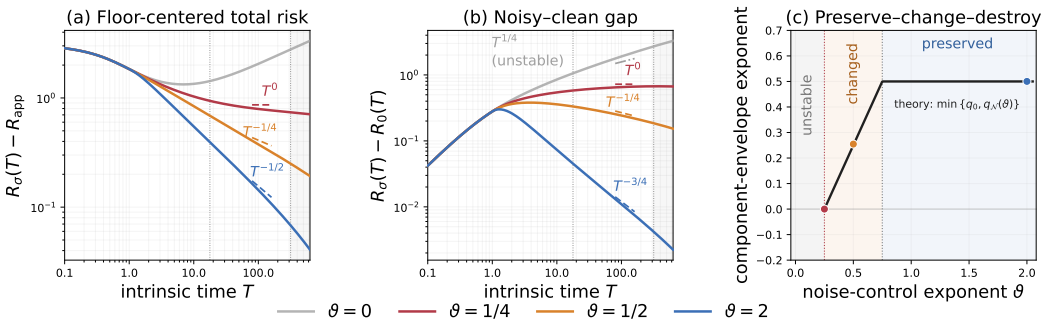


Figure 10: Floor-centered risk, noisy–clean gap, and predicted exponents across four noise-control schedules. The transition from growth through destruction and change to clean-law preservation supports Theorem 5.1.

$r(u) \sim cu$ gives $V(T) \sim c^{-1} \log T$, changing only normalization. For $\vartheta > 1$, Eq. (D.23) gives the same exponent. Thus throughout $\vartheta \geq 1$, the centered total exponent is $\min\{q_0, q_{\mathcal{K}}\}$, giving change, criticality, or preservation according as $q_{\mathcal{K}} < q_0$, $q_{\mathcal{K}} = q_0$, or $q_{\mathcal{K}} > q_0$.

At $\vartheta = 1 - q_{\mathcal{K}}$, Eq. (D.19) becomes

$$N_t \sim \sigma^2 \operatorname{Beta}(q_{\mathcal{K}}, 1 - q_{\mathcal{K}}) \frac{L_{\mathcal{K}}(T_t)}{L_r(T_t)}.$$

Since $Z_t \geq N_t$, this slow lower bound excludes positive-power decay; below the boundary, polynomial growth of $N_t/\sigma^2 = \sum_{s<t} K_{t,s}$ contradicts row stability for the fixed dynamics considered here. By the positive decomposition, the same lower bound rules out positive-power convergence of the total centered loss to R_{app}; this is the destroy boundary.

For $q_{\mathcal{K}} > 1$, Eq. (D.30) gives

$$Z(T) = T^{-q_{\mathcal{N}}(\vartheta)+o(1)}, \qquad q_{\mathcal{N}}(\vartheta) = \min\{\vartheta, q_{\mathcal{K}}\}, \qquad \vartheta > 0.$$

The same positive decomposition therefore gives total exponent $\min\{q_0, q_{\mathcal{N}}(\vartheta)\}$, and hence the stated change, criticality, and preservation alternatives. At $\vartheta = 0$, recent injection rules out positive-power decay; for $\vartheta < 0$, the row mass grows and contradicts row stability. The positive decomposition transfers this obstruction to the total centered loss. Together with the long-memory argument above, this proves Theorem D.21 and its pure-power main-text specialization Theorem 5.1. □

Figure 10 illustrates the classification in a finite-width analytic-spectrum continuum: increasing ϑ moves the response from the unstable side of the destroy boundary, through schedule-controlled change, and toward preservation of the clean-loss decay.

Proposition D.22 (Clean masking and memory saturation)**.** Work in the fixed-limit setting above, and suppose the schedules being compared have centered clean losses with the same leading asymptotic $R_{0,i}(T) - R_{\mathrm{app}} \sim T^{-q_0} L_0(T)$, where L_0 is eventually positive and slowly varying. The gap-based converse in Theorem D.19 transfers to total loss without a clean baseline only in the noise-dominated range $q < q_0$. This restriction is necessary for a uniform converse based on the full leading total-loss equivalent, while $q < q_{\mathcal{K}}$ is necessary for exponent-level schedule identification.

(i) If $q_0 < q_{\mathcal{K}}$, distinct gap exponents in $(q_0, q_{\mathcal{K}})$ can be generated by distinct admissible schedules, while their corresponding clean and total losses share the same leading clean-loss asymptotic.

(ii) If the clean loss has exponent $q_0 < q_{\mathcal{K}}$ and $Z(T) \sim T^{-q_0} L(T)$ for an eventually positive slowly varying L, then

$$R_\sigma(T) - R_{\mathrm{app}} \sim T^{-q_0}\big(L_0(T) + L(T)\big). \tag{D.39}$$

(iii) At the memory ceiling, integrable regularly varying injections with different tail exponents all produce stable full gaps with exponent $q_{\mathcal{K}}$.

Proof. For (i), choose distinct $q_i \in (q_0, q_{\mathcal{K}})$. Applying Eq. (D.38) with constant slow factors gives ratios of power $1 - q_{\mathcal{K}} + q_i$ and gaps $Z_i(T) \sim a_i T^{-q_i}$, with $a_i > 0$ chosen small enough for uniform stability. Since $q_i > q_0$, $Z_i = o(R_{0,i} - R_{\mathrm{app}})$, hence $R_{\sigma,i} - R_{\mathrm{app}} \sim R_{0,i} - R_{\mathrm{app}} \sim T^{-q_0} L_0(T)$ for both schedules.

Adding the positive equivalents proves Eq. (D.39) and the three cases in (ii). For example, when $L_0 \equiv 1$, $L_i(T) = (\log T)^{-i}$, $i = 1, 2$, yield different ratio tails through Eq. (D.38), but both leading total-risk asymptotics equal the clean-loss asymptotic by the common clean-asymptotic assumption.

For (iii), choose distinct integrable tail exponents $\vartheta_i > 1$ with row-stable amplitudes. From Eq. (D.23), we have $Z_i(T) \sim C_i k(T)$, so both gaps have exponent $q_{\mathcal{K}}$. The unit-mass family in Proposition D.18 more strongly gives distinct integrable tails with the same direct equivalent. □

D.7 Polynomial schedules in iteration time

Corollary D.23 (Polynomial schedules in iteration time)**.** Assume the corresponding fixed-system hypotheses of Theorem D.21, including its regularly varying clean-loss asymptotic with exponent $q_0 > 0$, and impose its row-stability condition only on the decaying branches. For $n \geq 1$, let

$$\eta_n = c_\eta n^{-\ell}, \qquad B_n = \lceil c_B n^b \rceil, \qquad c_\eta, c_B > 0, \quad 0 \leq \ell < 1, \quad b \geq 0.$$

Then

$$T_n \asymp n^{1-\ell}, \qquad r(T_n) = \frac{B_n}{\eta_n} \asymp T_n^{\vartheta}, \qquad \vartheta = \frac{\ell + b}{1 - \ell}.$$

Consequently, on every decaying branch, the noisy–clean gap exponent with respect to iteration index is

$$q_{\mathcal{N}}^{\mathrm{it}}(\ell, b) := (1 - \ell) q_{\mathcal{N}} \left(\frac{\ell + b}{1 - \ell} \right).$$

For the PLRF long-memory specialization

$$\frac{1}{4} < \alpha < \frac{1}{2}, \qquad q_{\mathcal{K}} = 2 - \frac{1}{2\alpha},$$

the following alternatives hold:

1. If $\ell + 2\alpha b < 1 - 2\alpha$, the row-stability bound cannot hold.
2. If $\ell + 2\alpha b = 1 - 2\alpha$, the noisy–clean gap has no positive-power decay.
3. If
$$\ell + 2\alpha b > 1 - 2\alpha, \qquad 2\ell + b < 1,$$
then
$$q_{\mathcal{N}}^{\mathrm{it}}(\ell, b) = \frac{\ell + 2\alpha b - (1 - 2\alpha)}{2\alpha}.$$
4. If $2\ell + b \geq 1$, then
$$q_{\mathcal{N}}^{\mathrm{it}}(\ell, b) = (1 - \ell) q_{\mathcal{K}}.$$
At $2\ell + b = 1$, the gap additionally carries the logarithmic correction inherited from the long-memory ceiling.

In integrable memory, where $q_{\mathcal{K}} > 1$, if $\ell + b > 0$, then

$$q_{\mathcal{N}}^{\mathrm{it}}(\ell, b) = \min\{\ell + b, (1 - \ell) q_{\mathcal{K}}\}.$$

If $\ell = b = 0$, the noisy–clean gap has no positive-power decay.

The clean-risk asymptotic in Theorem D.21 gives

$$R_0(T_n) - R_{\mathrm{app}} = n^{-(1-\ell)q_0 + o(1)}.$$

Hence, on every decaying branch,

$$R_\sigma(T_n) - R_{\mathrm{app}} = n^{-\min\{(1-\ell)q_0, q_{\mathcal{N}}^{\mathrm{it}}(\ell,b)\} + o(1)}.$$

Proof. Changing finitely many initial schedule values does not affect the stated asymptotic exponents, and discrete Karamata gives

$$T_n = T_1 + \sum_{s=1}^{n-1} c_\eta s^{-\ell} \sim \frac{c_\eta}{1-\ell} n^{1-\ell}.$$

Moreover, $B_n \sim \bar{c}_B n^b$, where $\bar{c}_B = c_B$ for $b > 0$ and $\bar{c}_B = \lceil c_B \rceil$ for $b = 0$. Therefore

$$r(T_n) = \frac{B_n}{\eta_n} \sim \frac{\bar{c}_B}{c_\eta} n^{b+\ell}.$$

Since $T_{n+1}/T_n \to 1$, the piecewise-constant intrinsic-time interpolation of r is eventually nondecreasing and regularly varying with index

$$\vartheta = \frac{\ell + b}{1-\ell}.$$

In PLRF long memory,

$$1 - q_{\mathcal{K}} = \frac{1-2\alpha}{2\alpha}.$$

Thus

$$\vartheta = 1 - q_{\mathcal{K}} \quad \Longleftrightarrow \quad \ell + 2\alpha b = 1 - 2\alpha,$$

which gives the stability/no-decay boundary. Likewise,

$$\vartheta = 1 \quad \Longleftrightarrow \quad 2\ell + b = 1,$$

which gives the memory ceiling.

On the interior long-memory branch $1 - q_{\mathcal{K}} < \vartheta < 1$, Theorem D.21 gives intrinsic-time gap exponent $\vartheta + q_{\mathcal{K}} - 1$. Composing with $T_n \asymp n^{1-\ell}$ yields

$$(1-\ell)(\vartheta + q_{\mathcal{K}} - 1) = (1-\ell)\left(\frac{\ell+b}{1-\ell} + 1 - \frac{1}{2\alpha}\right) = \frac{\ell + 2\alpha b - (1-2\alpha)}{2\alpha}.$$

For $\vartheta > 1$, the intrinsic-time exponent saturates at $q_{\mathcal{K}}$, giving iteration-time exponent $(1-\ell)q_{\mathcal{K}}$. At $\vartheta = 1$, the polynomial schedule gives $r(u) \sim cu$ for some $c > 0$, so $V(T) \sim c^{-1} \log T$; the long-memory boundary asymptotic in Theorem D.21 therefore gives the stated ceiling logarithm.

In integrable memory, the intrinsic-time exponent is $\min\{\vartheta, q_{\mathcal{K}}\}$ whenever $\vartheta > 0$. Hence

$$(1-\ell)\min\{\vartheta, q_{\mathcal{K}}\} = \min\{\ell + b, (1-\ell)q_{\mathcal{K}}\}.$$

When $\ell = b = 0$, one has $\vartheta = 0$, and the gap has no positive-power decay.

Finally, the regularly varying clean-risk asymptotic in Theorem D.21 composed with $T_n \asymp n^{1-\ell}$ gives clean iteration-time exponent $(1-\ell)q_0$. Positivity then gives the displayed minimum of the clean and gap exponents and the change/critical/preserve alternatives. □

E CONSTANT LEARNING-RATE AND BATCH-SIZE SCHEDULES

This appendix analyzes constant learning-rate and batch-size schedules under a feature-compute budget. Appendix E.1 collects the forcing, memory, and localization inputs needed to formulate the optimization problem. Appendix E.2 studies the exact discrete-time infinite-width problem and establishes the scale of its fixed-noise optimum. Appendix E.3 derives the source-window-optimal rates below trace class and the conditional plateau within a fixed finite-time window above trace class, and Appendix E.4 proves these results. Appendix E.5 compares the clean and noisy optima and derives the compute and noise scales at which the transition occurs.

Within the analyzed source and finite-time windows, label noise shifts compute toward wider models and fewer iterations.

Unless a statement explicitly invokes the exact conditional recursion and uses Roman risk notation, every calligraphic forcing, kernel, and risk in this section is a resolvent deterministic equivalent (DE).

Write $P^\star_{\sigma,\mathrm{const}} := \lim_{\mathfrak{f}\to\infty} \mathcal{R}^\star_{\sigma,\mathrm{const}}(\mathfrak{f})$ for the constant-schedule compute-optimal DE plateau, distinct from the infinite-width DE curve $\mathcal{R}_{\sigma,\infty}(t)$.

E.1 Scaling inputs and optimization setup

Write $p := 2\alpha + 2\beta - 1 > 0$.

Here $T = \eta t$, and stability gives $q_\eta(\lambda)^t \asymp e^{-cT\lambda}$. Thus $\lambda_j = j^{-2\alpha}$ has cutoff

$$j_T \asymp T^{1/(2\alpha)}: \qquad j \ll j_T \text{ is substantially learned}, \qquad j \gg j_T \text{ is substantially unlearned.}$$

The power-law window $1 \ll \eta t \lesssim m^{2\alpha}$ is therefore past the initial transient but before resolution of the smallest feature modes.

The four clean-loss components used below have the following scales. On the LM and IM DE branches, assume $\alpha > 1/4$, $\beta < 1+2\alpha$, and stay away from critical lines; the FB total-risk asymptotic uses the same source restriction with $0 < \alpha < 1/4$ and $p > 0$. In the window $1 \ll \eta t \lesssim m^{2\alpha}$, absent components are set to zero and

$$\mathcal{F}_{pp}(t,m) \asymp (\eta t)^{-p/(2\alpha)}, \tag{E.1}$$

$$\mathcal{F}_0(m) \asymp \begin{cases} m^{-p}, & \beta < \frac{1}{2}, \\ m^{-2\alpha}, & \beta > \frac{1}{2}, \end{cases}$$

$$\mathcal{F}_{ac}(t,m) \asymp m^{-1}(\eta t)^{-1+1/(2\alpha)} \quad \left(\alpha > \frac{1}{2},\ \beta > \frac{1}{2}\right), \tag{E.2}$$

$$\frac{1}{\eta}\mathcal{K}_{pp}(t,m) \asymp \frac{\eta}{B}(\eta t)^{-2+1/(2\alpha)} \quad \left(\alpha > \frac{1}{4}\right). \tag{E.3}$$

Here $\mathcal{F}_{pp}$ is time-limited unlearned aligned-target error, $\mathcal{F}_0$ is width-limited permanent null-space error, and $\mathcal{F}_{ac}$ and $\mathcal{K}_{pp}/\eta$ depend on both width and time, representing finite-width spectral distortion and aligned spectral memory from transient SGD sampling noise, respectively. The next corollary shows that these four terms recover the clean $4+3$ decomposition. This closure is not an additional assumption: the forcing and kernel estimates of Paquette et al. (2024, Appendices G–H), together with their Volterra approximation theorem and subexponential kernel estimate, establish the decomposition in the canonical PLRF model. We record its translation to the present notation.

Corollary E.1 (Recovery of the $4+3$ spectral decomposition)**.** Suppose $\alpha > 1/4$, $2\alpha + 2\beta > 1$, $\beta < 1 + 2\alpha$, with parameters away from the critical lines. Let the constant schedule satisfy Assumption A.2. Under the notation map

$$d_{4+3} = m, \qquad v_{4+3} = d, \qquad B_{4+3} = B, \qquad \gamma_{4+3} = \frac{\eta}{B},$$

uniformly for $1 \ll \eta t \lesssim m^{2\alpha}$,

$$\mathcal{R}_0(t,m) \asymp \mathcal{F}_0(m) + \mathcal{F}_{pp}(t,m) + \mathcal{F}_{ac}(t,m) + \frac{1}{\eta}\mathcal{K}_{pp}(t,m). \tag{E.4}$$

Proof. Under the displayed map, $\gamma_{4+3}B_{4+3} = \eta$ and $\gamma_{4+3}^2 B_{4+3} = \eta^2/B$, matching both the intrinsic clock and the kernel prefactor. The forcing estimates of Paquette et al. (2024, Appendix H) give $\mathcal{F} \asymp \mathcal{F}_0 + \mathcal{F}_{pp} + \mathcal{F}_{ac}$, including their error control. Their complete-kernel comparison and Volterra approximation theorem give the remaining $\eta^{-1}\mathcal{K}_{pp}$ contribution, proving the claim. □

When $\alpha < 1/4$, the infinite pure-point kernel is not summable. The finite-bulk DE kernel and noisy–clean gap are nevertheless controlled by Proposition C.17, and the canonical PLRF constant-schedule risk asymptotic is the specialization of Proposition C.18 for $p > 0$, $\beta < 1 + 2\alpha$, off the critical lines. For comparison, write $F_{ac}^{\boldsymbol{W}}$ and $F_0^{\boldsymbol{W}}$ for the absolutely-continuous and zero-mode components of the exact empirical forcing. The corresponding finite-dimensional proposition of Paquette et al. (2024) gives

$$F_{ac}^{\boldsymbol{W}}(t,m) = 0 \quad \left(\beta < \frac{1}{2}\right), \qquad 0 \le F_{ac}^{\boldsymbol{W}}(t,m) \le C_{\alpha,\beta,c} F_0^{\boldsymbol{W}}(m) \quad \left(\beta > \frac{1}{2},\ \alpha < \frac{1}{2}\right).$$

Here $c = \lim d/m \in (1,\infty)$, uniformly on strict localized windows.

Label noise is controlled by the cumulative kernel. For $\alpha > 1/4$, the PLRF mode estimates and Eq. (A.1) give

$$\mathcal{S}_t(m) \asymp \frac{\eta}{B} \sum_{j \le m} j^{-2\alpha} \left(1 - \exp\{-c_1 \eta t j^{-2\alpha}\}\right), \tag{E.5}$$

for a constant $c_1 > 0$. In the trace-class region $\alpha > 1/2$, write $\mathcal{S}_\infty(m) := \lim_{t\to\infty} \mathcal{S}_t(m) \in (0, \kappa]$. Then

$$\mathcal{S}_t(m) = \begin{cases} \mathcal{S}_\infty(m) - \Theta\left(t^{-1+1/(2\alpha)}\right), & \alpha > \frac{1}{2}, \\ \Theta\left(\left(\frac{t}{Bm}\right)^{(1-2\alpha)/(2\alpha)}\right), & \frac{1}{4} < \alpha < \frac{1}{2}, \\ \Theta\left(\frac{t}{Bm}\right), & 0 < \alpha < \frac{1}{4}, \end{cases} \tag{E.6}$$

The last line is the target-independent finite-bulk DE kernel estimate from Proposition C.17; the exact empirical analogue is Lemma C.8. Above trace class, accumulated noise saturates. Below trace class, the width-dependent learning rate can drive it to zero.

One iteration costs order Bm, so total compute is $\mathfrak{f} := tBm$. For reference, define the full-budget optimum

$$\mathcal{R}^\star_{\sigma,\mathrm{const}}(\mathfrak{f}) := \min_{1 \le m \le \mathfrak{f}/B} \mathcal{R}_\sigma\left(\frac{\mathfrak{f}}{Bm}, m\right),$$

and let $m^\star_{\sigma,\mathrm{const}}(\mathfrak{f})$ be any minimizer. To separate a possible noise floor from the part that still improves, define

$$P^\star_{\sigma,\mathrm{const}} := \lim_{\mathfrak{f}\to\infty} \mathcal{R}^\star_{\sigma,\mathrm{const}}(\mathfrak{f}),$$
$$\Delta^\star_{\sigma,\mathrm{const}}(\mathfrak{f}) := \mathcal{R}^\star_{\sigma,\mathrm{const}}(\mathfrak{f}) - P^\star_{\sigma,\mathrm{const}}.$$

Below trace class, the power laws hold only in the source window. We therefore optimize only over the concrete interior class

$$\mathcal{M}_{\mathrm{src}}(\mathfrak{f}) := \left\{1 \le m \le \frac{\mathfrak{f}}{B} : \log \mathfrak{f} \le \frac{\eta \mathfrak{f}}{Bm} \le \frac{m^{2\alpha}}{\log \mathfrak{f}}\right\}.$$

For sufficiently large $\mathfrak{f}$, define

$$\mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) := \min_{m \in \mathcal{M}_{\mathrm{src}}(\mathfrak{f})} \mathcal{R}_\sigma\left(\frac{\mathfrak{f}}{Bm}, m\right), \qquad m^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \in \operatorname*{argmin}_{m \in \mathcal{M}_{\mathrm{src}}(\mathfrak{f})} \mathcal{R}_\sigma\left(\frac{\mathfrak{f}}{Bm}, m\right).$$

Set $t^\star_{\sigma,\mathrm{src},\mathfrak{f}} := \mathfrak{f}/(Bm^\star_{\sigma,\mathrm{src}}(\mathfrak{f}))$. The logarithmic margins select a strict source window; the attaining scales below lie polynomially far from both boundaries.

In the trace-class region $\alpha > 1/2$, define $\mathcal{R}_{\sigma,\infty}(t) := \lim_{m\to\infty} \mathcal{R}_\sigma(t, m)$ and

$$a_\sigma = \begin{cases} p, & \beta < \frac{1}{2}, \\ 1, & \beta > \frac{1}{2}. \end{cases} \tag{E.7}$$

Assumption E.2 (Trace-class local optimum and DE expansion)**.** Let $U \subset (1, \infty)$ be compact.

(a) The continuously interpolated curve $\mathcal{R}_{\sigma,\infty}$ has a unique nondegenerate minimizer $t^\star_{\sigma,\mathrm{const}}$ in the interior of U.

(b) Uniformly for $t \in U$,

$$\mathcal{R}_\sigma\left(t, \frac{\mathfrak{f}}{Bt}\right) = \mathcal{R}_{\sigma,\infty}(t) + \left(\frac{\mathfrak{f}}{Bt}\right)^{-a_\sigma} G_\sigma(t) + o(\mathfrak{f}^{-a_\sigma}), \qquad G_\sigma(t^\star_{\sigma,\mathrm{const}}) > 0. \tag{E.8}$$

Under Assumption E.2, define

$$\mathcal{R}^\star_{\sigma,U}(\mathfrak{f}) := \min_{t \in U} \mathcal{R}_\sigma\left(t, \frac{\mathfrak{f}}{Bt}\right), \qquad t^\star_{\sigma,U,\mathfrak{f}} \in \operatorname*{argmin}_{t \in U} \mathcal{R}_\sigma\left(t, \frac{\mathfrak{f}}{Bt}\right),$$

and set $m^\star_{\sigma,U,\mathfrak{f}} := \mathfrak{f}/(Bt^\star_{\sigma,U,\mathfrak{f}})$, $P^\star_{\sigma,U} := \mathcal{R}_{\sigma,\infty}(t^\star_{\sigma,\mathrm{const}})$, and $\Delta^\star_{\sigma,U}(\mathfrak{f}) := \mathcal{R}^\star_{\sigma,U}(\mathfrak{f}) - P^\star_{\sigma,U}$.

E.2 Discrete infinite-width fixed-noise optimum

Before optimizing over continuous compute diagonals, introduce the exact integer-time objects at infinite width. For trace-class parameters, let

$$\kappa_\infty := \frac{\eta}{B}\sum_{j\ge 1}\frac{j^{-2\alpha}}{2-\eta\left(1+\frac{1}{B}\right)j^{-2\alpha}}.$$

For $|w|<1$, set

$$\widetilde{\mathcal{F}}_\infty(w) := \sum_{j\ge 1}\frac{j^{-2(\alpha+\beta)}}{1-q_\eta(j^{-2\alpha})w}, \qquad \widetilde{\mathcal{K}}_\infty(w) := \frac{\eta^2}{B}\sum_{j\ge 1}\frac{j^{-4\alpha}}{1-q_\eta(j^{-2\alpha})w}.$$

Thus $\widetilde{\mathcal{K}}_\infty(1)=\kappa_\infty$. Write $[w^t]$ for formal coefficient extraction and, for $\sigma^2>0$, define

$$\mathcal{R}^{\mathrm{disc}}_{\sigma,\infty}(t) := [w^t]\,\frac{\widetilde{\mathcal{F}}_\infty(w)+\sigma^2 w\widetilde{\mathcal{K}}_\infty(w)/(1-w)}{1-w\widetilde{\mathcal{K}}_\infty(w)}, \qquad \mathcal{R}^{\star,\mathrm{disc}}_{\sigma,\infty} := \inf_{t\in\mathbb{N},\ t\ge 1}\mathcal{R}^{\mathrm{disc}}_{\sigma,\infty}(t).$$

The next proposition checks this discrete problem. It shows that its best fixed-noise risk is uniformly of order σ^2. The proof is in Appendix E.2.

Proposition E.3 (Uniform scale of the discrete trace-class optimum)**.** Fix $\sigma^2>0$, $\delta\in(0,1)$, $\underline{\eta}>0$, and $\overline{\kappa}\in(0,1)$, and restrict to

$$\alpha\ge\tfrac{1}{2}+\delta, \qquad \eta\ge\underline{\eta}, \qquad \eta\left(1+\frac{1}{B}\right)\le 2-\delta, \qquad \kappa_\infty\le\overline{\kappa}.$$

Then, uniformly on this strict region,

$$\sigma^2\frac{\eta^2}{B}\zeta(4\alpha)\le\mathcal{R}^{\star,\mathrm{disc}}_{\sigma,\infty}\le\frac{\sigma^2\kappa_\infty}{1-\kappa_\infty}, \qquad \mathcal{R}^{\star,\mathrm{disc}}_{\sigma,\infty}\asymp\sigma^2\frac{\eta^2}{B}\zeta(4\alpha)\asymp\sigma^2.$$

For genuinely empirical finite-step statements, the Gaussian propagation in Appendix C.1 specializes to

$$\mathbb{E}_t[\rho_j^2(t+1)] = q_\eta(\widehat{\lambda}_j)\rho_j^2(t)+\frac{\eta^2}{B}\widehat{\lambda}_j^2\left(\|\boldsymbol{e}_t\|_2^2+\sigma^2\right). \tag{E.9}$$

Proof of Proposition E.3. For fixed integer time,

$$\mathbb{E}\|\widehat{\boldsymbol{H}}-\boldsymbol{\Lambda}\|_{\mathrm{HS}}^2=\frac{(\operatorname{tr}\boldsymbol{\Lambda})^2+\operatorname{tr}(\boldsymbol{\Lambda}^2)}{m}=O(m^{-1}).$$

Finite induction in Eq. (E.9) gives the stated infinite-width series, while positivity gives

$$\mathcal{R}^{\mathrm{disc}}_{\sigma,\infty}(t)\ge\sigma^2\widetilde{\mathcal{K}}_\infty(0)=\sigma^2\frac{\eta^2}{B}\zeta(4\alpha).$$

The kernel geometric series gives κ_∞; hence the positive Volterra resolvent converges to $\sigma^2\kappa_\infty/(1-\kappa_\infty)$, which bounds the integer-time infimum from above. Therefore

$$1\le\frac{\mathcal{R}^{\star,\mathrm{disc}}_{\sigma,\infty}}{\sigma^2\frac{\eta^2}{B}\zeta(4\alpha)}\le\frac{\overline{\kappa}}{(1-\overline{\kappa})\underline{\eta}^2}, \quad \text{and} \quad \underline{\eta}^2\le\frac{\eta^2}{B}\zeta(4\alpha)\le\zeta(2+4\delta),$$

proving both uniform comparisons. □

E.3 Source-window rates and trace-class plateau

The next theorem gives two distinct optimization statements across the eight open PLRF propagation subregimes of Definition D1. Above trace class, it concerns the optimum within the fixed finite-time window U. Below trace class, it concerns the source-window-restricted optimum, which balances unlearned target error against accumulated label noise. The proof is in Appendix E.4; the plateau consequence is stated separately in Proposition E.5.

For fixed $\sigma^2 > 0$, denote the relevant loss and width exponents by $(\bar{\rho}_\sigma, \bar{\xi}_\sigma)$, using

$$\begin{cases} \mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \asymp \mathfrak{f}^{-\bar{\rho}_\sigma}, & m^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \asymp \mathfrak{f}^{\bar{\xi}_\sigma}, & \alpha < 1/2, \\ \Delta^\star_{\sigma,U}(\mathfrak{f}) \asymp \mathfrak{f}^{-\bar{\rho}_\sigma}, & m^\star_{\sigma,U,\mathfrak{f}} \asymp \mathfrak{f}^{\bar{\xi}_\sigma}, & \alpha > 1/2. \end{cases}$$

On the trace-class branch, $t^\star_{\sigma,U,\mathfrak{f}}$ denotes the optimizing iteration within U.

Theorem E.4 (Fixed-noise compute rates)**.** Fix $\sigma^2 > 0$, suppose $2\alpha + 2\beta > 1$, and assume (α, β) lies in an open PLRF propagation subregime. Assume Assumption A.2 and the cumulative-memory asymptotics in Eq. (E.6).

On the LM and IM branches, assume $\beta < 1 + 2\alpha$, stay off the critical lines, and use Corollary E.1. On the FB branches, assume $\beta < 1 + 2\alpha$, stay off the critical lines, and use Proposition C.18; the high-source FB_2 extension additionally uses Assumption C.19 and Corollary C.20.

For $\alpha < 1/2$, optimize over $\mathcal{M}_{\mathrm{src}}(\mathfrak{f})$. For $\alpha > 1/2$, optimize over U and assume Assumption E.2. As $\mathfrak{f} \to \infty$ with fixed $\sigma^2 > 0$, these exponents are well defined and satisfy

$$(\bar{\rho}_\sigma, \bar{\xi}_\sigma) = \begin{cases} (p, 1)\,, & \alpha > \frac{1}{2},\ \beta < \frac{1}{2}, \\ (1, 1), & \alpha > \frac{1}{2},\ \beta > \frac{1}{2}, \\ \left(\dfrac{(1-2\alpha)p}{2[\alpha(1-2\alpha) + 2\beta(1-\alpha)]}, \dfrac{\beta}{\alpha(1-2\alpha) + 2\beta(1-\alpha)} \right), & \frac{1}{4} < \alpha < \frac{1}{2}, \\ \left(\dfrac{\alpha p}{p(1-\alpha) + 2\alpha}, \dfrac{p + 2\alpha}{2[p(1-\alpha) + 2\alpha]} \right), & 0 < \alpha < \frac{1}{4}. \end{cases} \tag{E.10}$$

For the five sub-trace subregimes, the corresponding loss and width scalings are

$$\mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \asymp \begin{cases} (\sigma^2)^{\frac{(1-\alpha)p}{\alpha(1-2\alpha)+2\beta(1-\alpha)}} \mathfrak{f}^{-\frac{(1-2\alpha)p}{2[\alpha(1-2\alpha)+2\beta(1-\alpha)]}}, & \frac{1}{4} < \alpha < \frac{1}{2}, \\ (\sigma^2)^{\frac{(1-\alpha)p}{p(1-\alpha)+2\alpha}} \mathfrak{f}^{-\frac{\alpha p}{p(1-\alpha)+2\alpha}}, & 0 < \alpha < \frac{1}{4}, \end{cases}$$

$$m^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \asymp \begin{cases} (\sigma^2)^{\frac{\alpha}{\alpha(1-2\alpha)+2\beta(1-\alpha)}} \mathfrak{f}^{\frac{\beta}{\alpha(1-2\alpha)+2\beta(1-\alpha)}}, & \frac{1}{4} < \alpha < \frac{1}{2}, \\ (\sigma^2)^{\frac{\alpha}{p(1-\alpha)+2\alpha}} \mathfrak{f}^{\frac{p+2\alpha}{2[p(1-\alpha)+2\alpha]}}, & 0 < \alpha < \frac{1}{4}. \end{cases} \tag{E.11}$$

In the three IM subregimes, every such choice satisfies

$$t^\star_{\sigma,U,\mathfrak{f}} \longrightarrow t^\star_{\sigma,\mathrm{const}} = \Theta(1), \qquad m^\star_{\sigma,U,\mathfrak{f}} \longrightarrow \infty.$$

In the IM regime, the optimal number of iterations remains finite. Within the source window, the five FB–LM subregimes balance $\mathcal{F}_{pp} \asymp \sigma^2 \mathcal{S}_t$.

Figure 11 visualizes the finite-time stopping mechanism in the three IM subregimes using finite-width deterministic-equivalent Volterra dynamics with $B = 1$, $\eta = 0.05$, and $d/m = 2$.

To compare with clean training, define

$$\mathcal{R}^\star_{0,\mathrm{const}}(\mathfrak{f}) := \min_{1 \le m \le \mathfrak{f}/B} \mathcal{R}_0\left(\frac{\mathfrak{f}}{Bm}, m\right), \qquad \mathcal{R}^\star_{0,\mathrm{const}}(\mathfrak{f}) \asymp \mathfrak{f}^{-\rho_0}, \qquad m^\star_{0,\mathrm{const}}(\mathfrak{f}) \asymp \mathfrak{f}^{\xi_0},$$

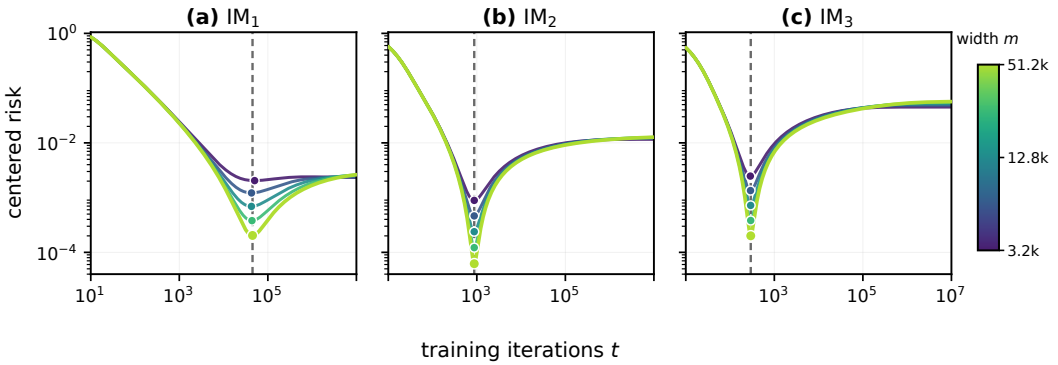


Figure 11: Centered finite-width DE risk across widths in the three integrable-memory subregimes. The stabilization of the minimizing iteration supports the finite-time local optimum in Theorem E.4.

where $m^\star_{0,\mathrm{const}}(\mathfrak{f})$ is any minimizer in the definition of $\mathcal{R}^\star_{0,\mathrm{const}}$. The clean $4+3$ calculation gives the reusable baseline

$$(\rho_0,\xi_0)=\begin{cases}\left(\dfrac{p}{2\alpha+1},\dfrac{1}{2\alpha+1}\right), & \text{4+3 Phase Ia,}\\ \left(\dfrac{p}{2(\alpha+\beta)},\dfrac{\beta}{\alpha+\beta}\right), & \text{4+3 Phase II,}\\ \left(\dfrac{4\alpha-1}{4\alpha},\dfrac{1}{2}\right), & \text{4+3 Phase III,}\\ \left(\dfrac{p}{2},\dfrac{1}{2}\right), & \text{4+3 Phase Ib,}\\ \left(\dfrac{\alpha p}{2\beta-1+3\alpha-2\alpha\beta},\dfrac{p}{2(2\beta-1+3\alpha-2\alpha\beta)}\right), & \text{4+3 Phase Ic,}\\ \left(\alpha,\dfrac{1}{2}\right), & \text{4+3 Phase IVa,}\\ \left(\dfrac{(1-2\alpha)p}{2[2\beta(1-\alpha)-\alpha]},\dfrac{\beta-\alpha}{2\beta(1-\alpha)-\alpha}\right), & \text{4+3 Phase IVb.}\end{cases} \tag{E.12}$$

Combining Eqs. (E.10) and (E.12) gives the following comparisons, with the noisy exponents interpreted as source-window-restricted below trace class and U-restricted above trace class:

$$\bar\rho_\sigma\begin{cases}>\rho_0, & \text{in 4+3 Phases Ia, II, and III,}\\ <\rho_0, & \text{in 4+3 Phases Ib, Ic, IVa, and IVb, '}\end{cases}$$

$$\bar\xi_\sigma>\xi_0 \quad \text{in all seven open 4+3 comparison phases.}$$

The larger exponent in 4+3 Phases Ia, II, and III does not mean that noise improves the uncentered risk. Clean and noisy losses are centered at different limits. The larger exponent only describes the large-width DE correction around the finite-time local optimum. Since $\mathfrak{f}=Bmt$,

$$t^\star_{0,\mathrm{const},\mathfrak{f}}\asymp\mathfrak{f}^{1-\xi_0},\qquad t^\star_{\sigma,\mathrm{src},\mathfrak{f}}\asymp\mathfrak{f}^{1-\bar\xi_\sigma}\quad(\alpha<1/2).$$

Thus the displayed noisy branch allocates a larger compute exponent to width than the clean benchmark. Above trace class the shift is complete, with $t^\star_{\sigma,\mathrm{const}}=\Theta(1)$ within U. Below trace class, both width and iteration count still diverge, but noise favors earlier stopping.

Fixed noise merges the eight propagation subregimes into four exponent regions. Within the source window below trace class, the balance is $\mathcal{F}_{pp}\asymp\sigma^2\mathcal{S}_t$; above it the U-restricted optimizer stops at finite time. The following proposition separates its plateau from the decaying centered correction.

Proposition E.5 (Source-window decay and trace-class plateau)**.** Under the hypotheses of Theorem E.4, if $\alpha < 1/2$, then

$$\mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \longrightarrow 0.$$

If $\alpha > 1/2$, then

$$P^\star_{\sigma,U} = \mathcal{R}_{\sigma,\infty}(t^\star_{\sigma,\mathrm{const}}) > 0,$$

and

$$\mathcal{R}^\star_{\sigma,U}(\mathfrak{f}) = P^\star_{\sigma,U} + \Theta(\mathfrak{f}^{-a_\sigma}), \qquad t^\star_{\sigma,\mathrm{const}} = \Theta(1).$$

E.4 Proof of Theorem E.4

Proof of Theorem E.4. Step 1: noisy Volterra reduction. Set $\eta_s \equiv \eta$, $B_s \equiv B$ in Proposition C.1. In the convolution form Eq. (A.13), the forcing and kernel from Eqs. (A.9) and (A.10) propagate initialization and one unit of clean-target DE risk, while labels contribute $\sigma^2\mathcal{S}$. Subtracting Eq. (A.12), with $\mathbf{1} = (1, 1, \ldots)$, gives

$$\mathcal{R}_\sigma - \mathcal{R}_0 = \sigma^2 \sum_{\ell\geq 1} \mathcal{K}^{*\ell} * \mathbf{1}.$$

Its first term is $\mathcal{S}$, so positivity and $\|\mathcal{K}\|_{\ell^1} \leq \kappa$ give

$$\mathcal{S}_t \leq \sum_{\ell\geq 1} (\mathcal{K}^{*\ell} * \mathbf{1})(t) \leq \mathcal{S}_t \sum_{\ell\geq 0} \kappa^\ell = \frac{\mathcal{S}_t}{1-\kappa}.$$

This proves Eq. (A.15).

Step 2: cumulative-kernel asymptotics. For $\alpha > 1/4$, Lemma C.12 gives, modewise and uniformly below the stability threshold,

$$\sum_{s=0}^{t-1} \frac{\eta^2}{B}\lambda^2 q_\eta(\lambda)^s = \frac{\eta^2}{B}\lambda^2 \frac{1-q_\eta(\lambda)^t}{1-q_\eta(\lambda)} \asymp \eta\lambda\left(1 - e^{-c_2\eta t\lambda}\right)$$

for some $c_2 > 0$. Inserting $\lambda_j \asymp j^{-2\alpha}$ proves Eq. (E.5); for $1/4 < \alpha < 1/2$, splitting at $j_t \asymp (\eta t)^{1/(2\alpha)}$ gives the same order on both sides:

$$\mathcal{S}_t \asymp \eta(\eta t)^{(1-2\alpha)/(2\alpha)} \asymp \left(\frac{t}{Bm}\right)^{(1-2\alpha)/(2\alpha)}.$$

For $0 < \alpha < 1/4$, the LM/IM filter fails at the lower edge, while the constant-schedule target-independent FB kernel estimate in Proposition C.17 gives

$$\mathcal{S}_t \asymp \frac{t}{Bm}.$$

For $\alpha > 1/2$, summability gives

$$\mathcal{S}_\infty(m) - \mathcal{S}_t(m) \asymp \sum_{j\geq 1} j^{-2\alpha} e^{-c_2\eta t j^{-2\alpha}} \asymp t^{-1+1/(2\alpha)}.$$

Together these three cases prove Eq. (E.6).

Step 3: optimization below the trace-class line. Suppose $\alpha < 1/2$. We identify the unique corner within $\mathcal{M}_{\mathrm{src}}(\mathfrak{f})$. With $t = \mathfrak{f}/(Bm)$, Eqs. (A.15) and (E.1) and, for $1/4 < \alpha < 1/2$, the noisy contribution give

$$\mathcal{F}_{pp} \asymp \mathfrak{f}^{-(2\alpha+2\beta-1)/(2\alpha)} m^{(2\alpha+2\beta-1)(1-\alpha)/\alpha}, \tag{E.13}$$

$$\sigma^2\mathcal{S}_t \asymp \sigma^2 \mathfrak{f}^{(1-2\alpha)/(2\alpha)} m^{-(1-2\alpha)/\alpha}. \tag{E.14}$$

Equating these terms and simplifying the two exponents gives

$$m^{[\alpha(1-2\alpha)+2\beta(1-\alpha)]/\alpha} \asymp \sigma^2 \mathfrak{f}^{\beta/\alpha},$$

$$p(1-\alpha) + 1 - 2\alpha = \alpha(1-2\alpha) + 2\beta(1-\alpha), \qquad p + 1 - 2\alpha = 2\beta,$$

$$m^\star_{\sigma,\mathrm{src}} \asymp (\sigma^2)^{\frac{\alpha}{\alpha(1-2\alpha)+2\beta(1-\alpha)}} \mathfrak{f}^{\frac{\beta}{\alpha(1-2\alpha)+2\beta(1-\alpha)}},$$

$$\mathcal{R}^\star_{\sigma,\mathrm{src}} \asymp (\sigma^2)^{\frac{(1-\alpha)p}{\alpha(1-2\alpha)+2\beta(1-\alpha)}} \mathfrak{f}^{-\frac{(1-2\alpha)p}{2[\alpha(1-2\alpha)+2\beta(1-\alpha)]}}.$$

The second line is the algebraic simplification; the first relation gives the third, and substitution into either balanced term gives the fourth.

For $0 < \alpha < 1/4$, Eq. (E.6) instead gives the three-step calculation

$$\sigma^2 \mathcal{S}_t \asymp \sigma^2 \mathfrak{f} m^{-2}, \qquad m^{[p(1-\alpha)+2\alpha]/\alpha} \asymp \sigma^2 \mathfrak{f}^{(p+2\alpha)/(2\alpha)},$$
$$m^\star_{\sigma,\mathrm{src}} \asymp (\sigma^2)^{\frac{\alpha}{p(1-\alpha)+2\alpha}} \mathfrak{f}^{\frac{p+2\alpha}{2[p(1-\alpha)+2\alpha]}}, \qquad \mathcal{R}^\star_{\sigma,\mathrm{src}} \asymp (\sigma^2)^{\frac{(1-\alpha)p}{p(1-\alpha)+2\alpha}} \mathfrak{f}^{-\frac{\alpha p}{p(1-\alpha)+2\alpha}}.$$

The second relation balances the first with Eq. (E.13); substitution gives the last line.

Step 4: validity of the sub-trace corners. For $1/4 < \alpha < 1/2$, the candidate has

$$m \asymp \mathfrak{f}^{\frac{\beta}{\alpha(1-2\alpha)+2\beta(1-\alpha)}}, \qquad \eta t \asymp \mathfrak{f}^{\frac{\alpha(1-2\alpha)}{\alpha(1-2\alpha)+2\beta(1-\alpha)}},$$
$$\frac{m^{2\alpha}}{\eta t} \asymp \mathfrak{f}^{\frac{\alpha p}{\alpha(1-2\alpha)+2\beta(1-\alpha)}}, \qquad t \asymp \mathfrak{f}^{\frac{(1-2\alpha)(\alpha+\beta)}{\alpha(1-2\alpha)+2\beta(1-\alpha)}}.$$

All exponents are positive, so the candidate lies polynomially inside $\mathcal{M}_{\mathrm{src}}(\mathfrak{f})$. The $\mathcal{F}_0$-exponent gap is $p^2/[2(\alpha(1-2\alpha)+2\beta(1-\alpha))] > 0$ in subregimes $\mathrm{LM}_{1,2}$ and $[2\alpha\beta - \frac{1}{2}(1-2\alpha)p]/[\alpha(1-2\alpha)+2\beta(1-\alpha)] > 0$ in subregime LM_3; its numerator equals $2\alpha^2$ at $\beta = 1/2$ and has derivative $4\alpha - 1 > 0$. In all three subregimes, Eq. (E.3) gives the positive kernel gap $\alpha(1-2\alpha)/[\alpha(1-2\alpha)+2\beta(1-\alpha)]$.

For $0 < \alpha < 1/4$, the corresponding candidate satisfies

$$m \asymp \mathfrak{f}^{\frac{p+2\alpha}{2[p(1-\alpha)+2\alpha]}}, \qquad t \asymp \mathfrak{f}^{\frac{p(1-2\alpha)+2\alpha}{2[p(1-\alpha)+2\alpha]}},$$
$$\eta t \asymp \mathfrak{f}^{\frac{2\alpha^2}{p(1-\alpha)+2\alpha}}, \qquad \frac{m^{2\alpha}}{\eta t} \asymp \mathfrak{f}^{\frac{\alpha p}{p(1-\alpha)+2\alpha}}.$$

All exponents are again positive, so this candidate also lies polynomially inside $\mathcal{M}_{\mathrm{src}}(\mathfrak{f})$. The width-floor gap is $p^2/[2(p(1-\alpha)+2\alpha)] > 0$ in subregime FB_1 and $2\alpha^2/[p(1-\alpha)+2\alpha] > 0$ in subregime FB_2. In the proved source range of Proposition C.18, these gaps make the displayed forcing–noise corner govern the finite-bulk DE formulas. On the high-source FB_2 branch, the same deduction additionally requires Assumption C.19 and Corollary C.20.

On the LM branches, Corollary C.15 applies for $\beta < 1 + 2\alpha$ off the critical lines. Positivity gives the matching in-window lower bound $\mathcal{F}_{pp} + \sigma^2 \mathcal{S}_t$; evaluation at the admissible candidate gives the upper bound. Thus the source-window-restricted optima in subregimes $\mathrm{LM}_{1,2,3}$ satisfy

$$\mathcal{F}_{pp} \asymp \sigma^2 \mathcal{S}_t,$$

whereas both FB subregimes in the proved source range use Proposition C.18. Their balanced values tend to zero, proving $\mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \to 0$.

Step 5: optimization above the trace-class line. Suppose $\alpha > 1/2$, and let $t^\star_{\sigma,U,\mathfrak{f}}$ minimize the compute-constrained diagonal family over U. Its associated width $m^\star_{\sigma,U,\mathfrak{f}} := \mathfrak{f}/(B t^\star_{\sigma,U,\mathfrak{f}})$ tends to infinity uniformly because $U \subset (1,\infty)$ is compact. Uniform convergence on U makes every subsequential limit minimize $\mathcal{R}_{\sigma,\infty}$; uniqueness gives $t^\star_{\sigma,U,\mathfrak{f}} \to t^\star_{\sigma,\mathrm{const}}$, with no optimizer-location rate claimed.

For the upper bound, evaluate Eq. (E.8) at $t^\star_{\sigma,\mathrm{const}}$; for the lower, use nonnegativity of the infinite-width value difference, convergence of the minimizers, and $G_\sigma(t^\star_{\sigma,\mathrm{const}}) > 0$. Thus

$$\Delta^\star_{\sigma,U}(\mathfrak{f}) \asymp \left(m^\star_{\sigma,U,\mathfrak{f}}\right)^{-a_\sigma} \asymp \mathfrak{f}^{-a_\sigma}.$$

By Eq. (E.7), the exponent is $2\alpha + 2\beta - 1$ in IM_1 and 1 in $\mathrm{IM}_{2,3}$, proving the finite-time conclusion over U.

The sub-trace argument gives the last two rows of Eq. (E.10) and Eq. (E.11). The trace-class argument gives the first two rows of Eq. (E.10) and the finite-time conclusion, completing the proof. □

Proof of Proposition E.5. If $\alpha < 1/2$, Steps 3–4 above show that the source-window optimum has a strictly positive decay exponent, and hence $\mathcal{R}^\star_{\sigma,\mathrm{src}}(\mathfrak{f}) \to 0$. If $\alpha > 1/2$, Step 5 identifies the finite limiting optimizer $t^\star_{\sigma,\mathrm{const}}$, and Eq. (E.8) gives

$$\mathcal{R}^\star_{\sigma,U}(\mathfrak{f}) = \mathcal{R}_{\sigma,\infty}(t^\star_{\sigma,\mathrm{const}}) + \Theta(\mathfrak{f}^{-a_\sigma}).$$

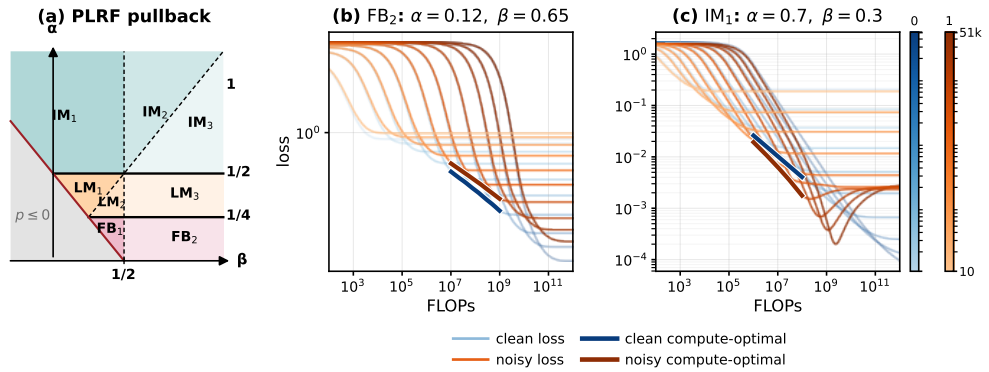


Figure 12: The PLRF propagation map and representative centered-risk compute frontiers in FB_2 and IM_1. Their continuing forcing–noise balance and finite-time stopping illustrate the phase-dependent shift of compute from iterations to width.

The positive stable noise resolvent makes the limiting value strictly positive. This is the claimed trace-class plateau. □

Figure 12 places representative constant-schedule Volterra trajectories from FB_2 and IM_1 on the PLRF propagation map. Their contrast supports the predicted continuing forcing–noise balance below trace class and finite-time noisy stopping within U above trace class.

E.5 Clean-to-noisy crossover scales

Weak-noise early-stopping prediction in the IM regime. For the crossover comparison below, the formal large-iteration matching convention compares the clean infinite-width transient with the unsaturated terminal-noise deficit as $\sigma^2 \downarrow 0$. With B and a stable η held fixed, it predicts

$$t^{\star}_{\sigma,\mathrm{const}} \asymp \begin{cases} (\sigma^2)^{-\alpha/\beta}, & \mathrm{IM}_1,\ \beta > 0, \\ (\sigma^2)^{-\alpha/\beta}, & \mathrm{IM}_2, \\ (\sigma^2)^{-1}, & \mathrm{IM}_3. \end{cases} \tag{E.15}$$

For subregime IM_1 with $\beta \le 0$, this balance yields no diverging weak-noise scale. These scales are formal order predictions from the weak-noise balance.

Scope of the weak-noise matching. For comparison only, first suppose $\beta > 0$ and set

$$q_{\mathcal{S}} := 1 - \frac{1}{2\alpha} > 0, \qquad q_{\mathrm{tail}} := q_{\mathcal{S}} + \min\left\{\frac{\beta}{\alpha}, 1\right\} > q_{\mathcal{S}}.$$

The clean infinite-width transient is $O(t^{-q_{\mathrm{tail}}})$, so the large-$t$ inputs and bounded positive Volterra resolvent give

$$\mathcal{R}_{\sigma,\infty}(t) = \mathcal{R}_{\sigma,\infty}(\infty) - c_\sigma \sigma^2 t^{-q_{\mathcal{S}}} + O(t^{-q_{\mathrm{tail}}}).$$

The big-O clean term has no specified leading coefficient or sign. Locating a weak-noise minimizer would require a uniform signed clean-tail expansion along a joint limit with $\sigma^2 \downarrow 0$.

For $\beta = 0$, the decay powers coincide; for $\beta < 0$, $q_{\mathrm{tail}} = q_{\mathcal{S}} + \beta/\alpha < q_{\mathcal{S}}$. Without a signed clean-tail expansion, neither identifies a diverging weak-noise minimizer. Hence the IM_1 row of Eq. (E.15) and the 4+3 Phase Ia crossover entry in Proposition E.6 are restricted to $\beta > 0$; the fixed-noise plateau conclusion above is unchanged.

Under the same numerical setting, Figure 13 varies noise and width to reveal how the finite early-stopping region lies between terminal training and no training in each IM subregime.

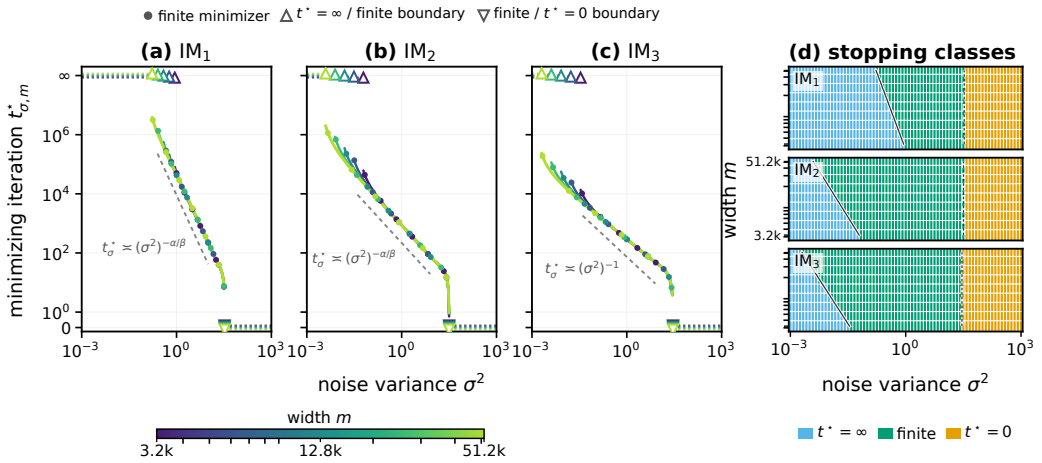


Figure 13: The minimizing iteration and stopping class across noise levels and widths in the three integrable-memory subregimes. The terminal–finite and finite–no-training transitions support the predicted subregime dependence of early stopping.

At small but fixed σ^2, when does the noisy optimum replace the clean optimum? Equivalently, at fixed compute, how large can the label-noise variance be before the fixed-noise optimum takes over? The next proposition retains the clean $4+3$ indexing because the clean $4+3$ crossover rates distinguish branches that merge under the propagation taxonomy. These are crossover scales, not sharp transition points. The proof is in Appendix E.5.

For any matched compute scale $\mathfrak{f}_{\mathrm{cross}}(\sigma)$, denote by $\sigma^2_{\mathrm{cross}}(\mathfrak{f})$ a fixed-compute inverse scale satisfying

$$\mathfrak{f}_{\mathrm{cross}}(\sigma_{\mathrm{cross}}) \asymp \mathfrak{f}.$$

Proposition E.6 (Matched clean-to-noisy compute scales)**.** Fix a small noise variance $\sigma^2 > 0$. Below trace class, match the noisy term at the clean-optimal width to the clean-optimal loss. Above trace class, use the weak-noise large-iteration prediction in Eq. (E.15), and match its noisy optimal iteration to the clean one. In 4+3 Phase Ia (our subregime IM$_1$), this convention applies only when $\beta > 0$; for $\beta \le 0$, it gives no diverging weak-noise minimizer. The resulting matched compute scales in the clean $4+3$ comparison coordinates are

$$\mathfrak{f}_{\mathrm{cross}}(\sigma) \asymp \begin{cases} (\sigma^2)^{-(2\alpha+1)/(2\beta)}, & \text{4+3 Phase Ia, } \beta > 0, \\ (\sigma^2)^{-(\alpha+\beta)/\beta}, & \text{4+3 Phase II,} \\ (\sigma^2)^{-2}, & \text{4+3 Phase III,} \\ (\sigma^2)^{-2/p}, & \text{4+3 Phase Ib,} \\ (\sigma^2)^{-\frac{2\beta-1+3\alpha-2\alpha\beta}{2\alpha^2}}, & \text{4+3 Phase Ic,} \\ (\sigma^2)^{-\frac{2\beta(1-\alpha)-\alpha}{\alpha(1-2\alpha)}}, & \text{4+3 Phase IVb.} \end{cases} \tag{E.16}$$

Equivalently, within the same source/weak-noise scope, any branch written in Eq. (E.16) as $\mathfrak{f}_{\mathrm{cross}}(\sigma) \asymp (\sigma^2)^{-\gamma}$ has $\sigma^2_{\mathrm{cross}}(\mathfrak{f}) \asymp \mathfrak{f}^{-1/\gamma}$, with the Phase Ia restriction $\beta > 0$. Within the source window, $1 \ll \mathfrak{f} \ll \mathfrak{f}_{\mathrm{cross}}(\sigma)$, equivalently $\sigma^2 \ll \sigma^2_{\mathrm{cross}}(\mathfrak{f})$, gives clean-optimal scaling; reversing both inequalities gives noisy-optimal scaling. The 4+3 Phase IVa branch within our subregime LM$_3$ has two matched compute scales,

$$\mathfrak{f}_{\mathrm{onset}}(\sigma) \asymp (\sigma^2)^{-1/\alpha}, \qquad \mathfrak{f}_{\mathrm{full}}(\sigma) \asymp (\sigma^2)^{-\frac{2\beta(1-\alpha)-\alpha}{\alpha(1-2\alpha)}}. \tag{E.17}$$

Their fixed-compute inverses are

$$\sigma^2_{\mathrm{onset}}(\mathfrak{f}) \asymp \mathfrak{f}^{-\alpha}, \qquad \sigma^2_{\mathrm{full}}(\mathfrak{f}) \asymp \mathfrak{f}^{-\alpha(1-2\alpha)/[2\beta(1-\alpha)-\alpha]}.$$

Because $\beta > 1/2$, $2\beta(1-\alpha) - \alpha > 1 - 2\alpha$, so $\sigma^2_{\text{onset}} \ll \sigma^2_{\text{full}}$. Hence

$$\begin{aligned} \sigma^2 \ll \sigma^2_{\text{onset}}(\mathfrak{f}) &\implies \text{clean-optimal scaling}, \\ \sigma^2_{\text{onset}}(\mathfrak{f}) \ll \sigma^2 \ll \sigma^2_{\text{full}}(\mathfrak{f}) &\implies \text{intermediate kernel/noise balance}, \\ \sigma^2 \gg \sigma^2_{\text{full}}(\mathfrak{f}) &\implies \text{full } \mathcal{F}_{pp}\text{–noise balance}. \end{aligned}$$

All locations are order predictions, with constants possibly depending on $\alpha, \beta, B, \eta, c$.

The displayed noise ranges invert Eq. (E.17).

Proof of Proposition E.6. For $\alpha > 1/2$, combine Eqs. (E.12) and (E.15) with $t^\star_{0,\text{const}} = \mathfrak{f}/(Bm^\star_{0,\text{const}})$. Equating clean and weak-noise iteration scales gives

4+3 phase	clean $t^\star$	weak-noise $t^\star$	σ^2_{match}
Ia	$\mathfrak{f}^{2\alpha/(2\alpha+1)}$	$(\sigma^2)^{-\alpha/\beta}$	$\mathfrak{f}^{-2\beta/(2\alpha+1)}$
II	$\mathfrak{f}^{\alpha/(\alpha+\beta)}$	$(\sigma^2)^{-\alpha/\beta}$	$\mathfrak{f}^{-\beta/(\alpha+\beta)}$
III	$\mathfrak{f}^{1/2}$	$(\sigma^2)^{-1}$	$\mathfrak{f}^{-1/2}$.

Inverting the last column gives the three trace-class branches of Eq. (E.16).

Below trace class, recall $p := 2\alpha + 2\beta - 1$ and write $\mathfrak{d}_{\text{Ic}} := 2\beta - 1 + 3\alpha - 2\alpha\beta$ and $\mathfrak{d}_{\text{IVb}} := 2\beta(1-\alpha) - \alpha$. For $1/4 < \alpha < 1/2$, evaluating Eq. (E.14) at the clean-optimal width gives

$$\sigma^2 \mathcal{S}_t \asymp \sigma^2 \mathfrak{f}^{(1-2\alpha)/(2\alpha)} \big(m^\star_{0,\text{const}}(\mathfrak{f})\big)^{-(1-2\alpha)/\alpha}.$$

For $0 < \alpha < 1/4$, use instead $\sigma^2 \mathcal{S}_t \asymp \sigma^2 \mathfrak{f} m^{-2}$. Substituting the clean branches from Eq. (E.12) gives, in Phase Ib for either range, $\sigma^2 \mathcal{S}_t \asymp \sigma^2$, $\mathcal{R}^\star_{0,\text{const}} \asymp \mathfrak{f}^{-p/2}$, and hence $\sigma^2 \asymp \mathfrak{f}^{-p/2}$. In Phase IVa, the same noise order matched with $\mathfrak{f}^{-\alpha}$ gives $\sigma^2 \asymp \mathfrak{f}^{-\alpha}$. The remaining substitutions are

$$\begin{aligned} \text{4+3 Phase IVb}: \quad & \sigma^2 \mathcal{S}_t \asymp \sigma^2 \mathfrak{f}^{-(1-2\alpha)(2\beta-1)/(2\mathfrak{d}_{\text{IVb}})}, && \mathcal{R}^\star_{0,\text{const}} \asymp \mathfrak{f}^{-(1-2\alpha)p/(2\mathfrak{d}_{\text{IVb}})}, \\ & && \sigma^2 \asymp \mathfrak{f}^{-\alpha(1-2\alpha)/\mathfrak{d}_{\text{IVb}}}, \\ \text{4+3 Phase Ic}: \quad & \sigma^2 \mathcal{S}_t \asymp \sigma^2 \mathfrak{f}^{1-p/\mathfrak{d}_{\text{Ic}}}, && \mathcal{R}^\star_{0,\text{const}} \asymp \mathfrak{f}^{-\alpha p/\mathfrak{d}_{\text{Ic}}}, \\ & && \sigma^2 \asymp \mathfrak{f}^{-2\alpha^2/\mathfrak{d}_{\text{Ic}}}. \end{aligned}$$

These are the Phase Ib, Ic, IVb, and IVa-onset variance thresholds; inversion gives their compute thresholds.

For the full-noise scale on the Phase IVa branch of our subregime LM_3, evaluate the $\mathcal{F}_{pp}$–noise corner in Eq. (E.11), direct substitution into Eq. (E.3) gives

$$\frac{\mathcal{K}_{pp}/\eta}{\mathcal{F}_{pp}} \asymp (\sigma^2)^{-\frac{2\beta(1-\alpha)-\alpha}{\alpha(1-2\alpha)+2\beta(1-\alpha)}} \mathfrak{f}^{-\frac{\alpha(1-2\alpha)}{\alpha(1-2\alpha)+2\beta(1-\alpha)}}.$$

This corner becomes self-consistent at

$$\sigma^2 \asymp \mathfrak{f}^{-\frac{\alpha(1-2\alpha)}{2\beta(1-\alpha)-\alpha}},$$

which is $\sigma^2_{\text{full}}(\mathfrak{f})$; inversion gives $\mathfrak{f}_{\text{full}}(\sigma)$. Since $\beta > 1/2$ implies $\mathfrak{d}_{\text{IVb}} > 1 - 2\alpha$, as $\sigma^2 \downarrow 0$, $\mathfrak{f}_{\text{onset}}(\sigma) \ll \mathfrak{f}_{\text{full}}(\sigma)$; equivalently, for large $\mathfrak{f}$, $\sigma^2_{\text{onset}}(\mathfrak{f}) \ll \sigma^2_{\text{full}}(\mathfrak{f})$, leaving the intermediate kernel/noise range. This proves both crossover parametrizations, with all source-window and order-level qualifications unchanged. □

Constant schedules therefore expose the same mechanism as the general theory: the spectrum fixes memory, memory accumulates label noise, and the resulting balance moves compute from iterations to width. The schedule sections ask how changing B_t/η_t can control this balance.

F END-TO-END LANGUAGE-MODEL EXPERIMENTS

We use three experiments to test increasingly stronger consequences of the schedule coordinates predicted by our theory.

Roadmap: First, we fix the intrinsic-time horizon and vary the ratio path $r(T) = B(T)/\eta(T)$, testing whether different ratio schedules produce different loss trajectories. Second, we fix both the intrinsic clock and the ratio path, but realize the same path either by varying learning rate at fixed batch size or by varying batch size at fixed learning rate. This tests whether the loss is approximately invariant to the particular learning-rate–batch-size factorization. Third, we adapt the practical seven-parameter functional scaling law of Li et al. (2025) but fit a forcing-memory surrogate on a single fixed-batch 8-1-1 trajectory and, without refitting, use it to predict the alternative factorization and unseen WSD schedules.

Together, the three experiments test *ratio-path separation*, *factorization collapse*, and *cross-schedule prediction*. The plain-SGD experiments directly test the coordinates $T = \sum_t \eta_t$ and $r = B/\eta$ derived by our theory; the Muon experiment uses separately calibrated coordinates and serves only as an optimizer-specific external-validity check. Code and configurations for reproducing the experiments in this section are included in the supplementary material.

F.1 TESTING RATIO-PATH EFFECTS AT MATCHED INTRINSIC TIME

Our theory assigns different roles to two schedule coordinates: intrinsic time $T_t = \sum_{s<t} \eta_s$ measures optimization progress, while $r_t = B_t/\eta_t$ controls the accumulation of stochastic error. This experiment tests whether the ratio path affects end-to-end LLM loss after the intrinsic-time horizon is fixed. If intrinsic time were the only relevant coordinate, different ratio schedules would produce similar loss curves. The theory instead predicts systematic differences, with diminishing improvement once the memory ceiling is reached.

We train a 30M-parameter nanoGPT model (Karpathy, 2022) on OpenWebText (Gokaslan & Cohen, 2019), with six layers, six attention heads, embedding width 384, and context length 256. Training uses plain SGD in FP32, without momentum, weight decay, or gradient clipping. We first train one model for 196,608 updates at a constant learning rate. We then copy this checkpoint into eleven runs and continue training them on the same ordered data stream. All runs use batch size $B = 8$; only the learning-rate schedule differs.

Let T_{fork} be the intrinsic time at the shared checkpoint. For each tail, the learning rate decays at

$$\eta(T) = \eta_0 X^{-\vartheta}, \qquad X = 1 + \frac{T - T_{\text{fork}}}{\tau_s}, \qquad \tau_s = 1024\eta_0,$$

and hence

$$\frac{B}{\eta(T)} = \frac{B}{\eta_0} X^{\vartheta}.$$

Larger ϑ therefore means faster learning-rate decay and faster growth of B/η. We use

$$\vartheta \in \{0, 0.125, 0.25, 0.375, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$$

and stop every run at $X = 15$ to ensure that all runs have the same intrinsic-time horizon, $T - T_{\text{fork}} = 14\tau_s$, but different numbers of updates and tokens. We evaluate each trajectory at 410 points on the same fixed set of 1,024 validation contexts. Note that, at fixed model, starting the same checkpoint, batch size, and intrinsic-time horizon, but not at fixed updates or tokens, the validation loss changes systematically with the ratio-growth exponent.

Figure 14 shows that the validation curves are systematically ordered by ϑ. At the same intrinsic time, larger ϑ gives lower validation loss, confirming that the ratio path B/η, rather than intrinsic time alone, organizes the schedule response. The improvement diminishes at large ϑ, as expected from a memory ceiling.

This experiment tests these predictions only qualitatively. It uses one seed, measures total validation cross-entropy rather than a noisy–clean gap, and does not independently estimate $q_{\mathcal{K}}$. It therefore does not identify the LM/IM boundary or verify the piecewise asymptotic exponent in Theorem 5.1.

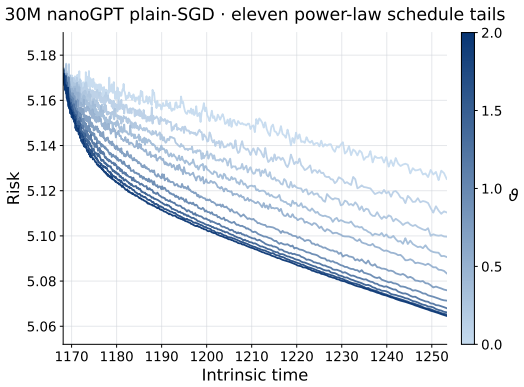


Figure 14: Effect of the ratio path at matched intrinsic time. All runs start from the same checkpoint and stop at the same intrinsic-time horizon, but use different numbers of updates and tokens. Larger ϑ means faster growth of B/η, and darker curves correspond to larger ϑ. Validation loss decreases with ϑ, with diminishing improvement at large ϑ.

F.2 Testing factorization invariance at matched ratio paths

This experiment tests whether an end-to-end validation-loss trajectory is governed mainly by intrinsic time and the ratio path, rather than by a particular factorization into learning rate and batch size. We compare two ratio-path shapes, named by their fixed-batch learning-rate realizations. WSD 80/20 is constant for the first 80% of training and decays exponentially over the final 20%, whereas 8-1-1 uses the same 80% prefix followed by two successive 10% decay stages. For each ratio path, one realization fixes batch size and varies learning rate, while the other fixes learning rate and varies batch size. Each pair starts from the same mature checkpoint, traverses the same future training tape, and is evaluated on the same fixed probe of 1024 validation contexts.

Both experiments use the same 300M-parameter GPT-2-tokenized nanoGPT architecture (Karpathy, 2022): 20 layers, 16 attention heads, width 1024, and context length 256. Each trajectory is trained on a frozen OpenWebText stream (Gokaslan & Cohen, 2019) for 6.5B tokens. The first 5.2B tokens form the shared prefix, and the remaining 1.3B tokens form one schedule tail. The left pair in Figure 15 uses plain SGD, while the right pair uses hybrid Muon.

For plain SGD, the matched clock and ratio are

$$T_t = \sum_{s<t} \eta_s, \qquad r(T_t) = \frac{B_t}{\eta_t}.$$

For a prescribed ratio path $r(T)$, the two factorizations are

$$\text{fixed batch:} \quad B_t = B_0, \quad \eta_t = \frac{B_0}{r(T_t)}, \qquad \text{fixed learning rate:} \quad \eta_t = \eta_0, \quad B_t = \eta_0 r(T_t).$$

The 300M plain-SGD run realizes this match exactly on an integer macro grid. At each grid cell, $\Delta T = 0.005$, and the integer schedule value is $g \in \{1, \dots, 10\}$ for WSD and $g \in \{3, 4, 9, 10\}$ for 8-1-1:

$$\begin{aligned} \text{fixed batch:} &\quad g \text{ updates with } B = 16, \quad \eta = 0.005/g, \\ \text{fixed learning rate:} &\quad 1 \text{ update with } B = 16g, \quad \eta = 0.005. \end{aligned}$$

Both realizations consume exactly $16g$ training contexts from the same contiguous segment of the frozen future tape, advance intrinsic time by the same ΔT, and realize $B/\eta = 3200g$ at every completed grid cell. Each source-schedule value is repeated for 16 such cells before the next value is used. Hence the match is exact at grid boundaries even though the optimizer-update counts differ.

For Muon (Jordan et al., 2024; Liu et al., 2025), a separate short calibration selects the empirical coordinates

$$\widetilde{T}_t = \sum_{s<t} \eta_s^2, \qquad \widetilde{r}(\widetilde{T}_t) = \frac{B_t}{\eta_t^2},$$

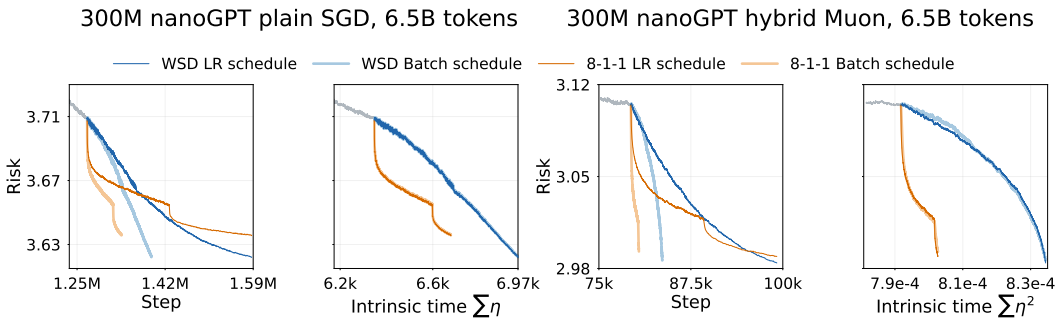


Figure 15: **Factorization collapse under matched ratio schedules.** Fixed-probe validation cross-entropy for the same 300M nanoGPT architecture trained for 6.5B OpenWebText tokens with plain SGD (left pair) and hybrid Muon (right pair), shown against optimizer step and the corresponding intrinsic clock. Within each WSD or 8-1-1 schedule family, the LR schedule varies η_t at fixed B_t, whereas the batch-size schedule varies B_t at fixed η_t. For plain SGD, these two realizations follow the same ratio path $r_t = B_t/\eta_t$ under $T_t = \sum_{s<t} \eta_s$; for Muon, they follow the same empirically calibrated path $\widetilde{r}_t = B_t/\eta_t^2$ under $\widetilde{T}_t = \sum_{s<t} \eta_s^2$. The paired trajectories differ in optimizer step but nearly coincide in the corresponding intrinsic time. Blue denotes WSD, orange denotes 8-1-1, and gray denotes the shared prefix.

with factorizations

$$\text{fixed batch:} \quad B_t = B_0, \quad \eta_t = \sqrt{\frac{B_0}{\widetilde{r}(\widetilde{T}_t)}}, \qquad \text{fixed learning rate:} \quad \eta_t = \eta_0, \quad B_t = \eta_0^2 \widetilde{r}(\widetilde{T}_t).$$

The Muon coordinate is empirical and is not identified with the plain-SGD ratio in our theory.

Figure 15 separates the ratio path from its factorization. On the optimizer-step axes, paired factorizations follow different trajectories because they require different numbers of updates to process the same data. On their respective intrinsic-time axes, the two factorizations of each WSD or 8-1-1 ratio path nearly coincide.

F.3 A FORCING-MEMORY SURROGATE PREDICTS UNSEEN SCHEDULES

We finally test whether the forcing–memory mechanism is expressive enough to describe the complete validation-risk trajectory, rather than only the factorization collapse. Our starting point is the practical seven-parameter FSL ansatz of Li et al. (2025). Their fixed-batch parameterization combines an intrinsic-time clean power law with a learning-rate-decay correction driven by the increments $\eta_{i-1} - \eta_i$; they fit its seven parameters on 8-1-1 and predict unseen WSD and cosine learning-rate schedules. We retain the intrinsic-time clean-plus-memory structure and the same fit-then-transfer protocol, but make the stochastic injection depend directly on the joint ratio $r(T) = B(T)/\eta(T)$. This gives the finite-window surrogate

$$\widehat{L}(T) = L_\infty + A_{\mathcal{F}}(1+T)^{-q_{\mathcal{F}}} + \int_0^T \frac{A_0 + A_1(1+u)^{-q_{\mathcal{F}}}}{r(u)} \big(1 + c_{\mathcal{K}}(T-u)\big)^{-q_{\mathcal{K}}} \, \mathrm{d}u \,.$$

Its seven parameters $(L_\infty, A_{\mathcal{F}}, A_0, A_1, c_{\mathcal{K}}, q_{\mathcal{F}}, q_{\mathcal{K}})$ replace the seven empirical FSL parameters of Li et al. (2025) with a forcing–kernel parameterization adapted to joint schedules. Their clean exponent s corresponds to our $q_{\mathcal{F}}$, whereas their γ parameterizes an integrated learning-rate-decay response rather than the memory kernel itself; for a pure power-law kernel, the corresponding exponents satisfy $\gamma = q_{\mathcal{K}} - 1$. The ratio form also applies when learning rate is fixed and batch size varies, a factorization not covered by the fixed-batch LRS ansatz.

For the OpenWebText run, the seven parameters are fitted only to the fixed-batch 8-1-1 validation trajectory and are then frozen. The same surrogate tracks the alternative 8-1-1 factorization and predicts both WSD trajectories without refitting (Figure 16). Thus it captures both the long-horizon loss decay and the change induced by a different schedule. The fitted exponents are $q_{\mathcal{K}} = 1.017$ and

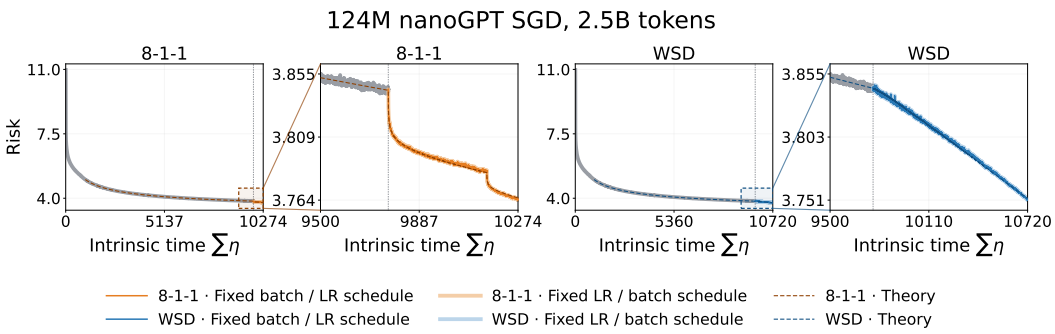


Figure 16: **A joint-schedule adaptation of the functional scaling law fits and transfers across LLM schedules.** Following the fit-then-transfer protocol of Li et al. (2025), the ratio-based surrogate is fitted only to the fixed-batch 8-1-1 validation trajectory, after which all parameters are frozen. It tracks the alternative 8-1-1 factorization and predicts both WSD trajectories without refitting, showing that the same forcing–memory parameterization captures long-horizon LLM loss decay across learning-rate and batch-size schedules. The fitted exponents are $q_{\mathcal{K}} = 1.017$ and $q_{\mathcal{F}} = 0.374$, placing the effective memory response close to the $q_{\mathcal{K}} = 1$ boundary. Risk denotes fixed-probe validation cross-entropy and $T = \sum_t \eta_t$.

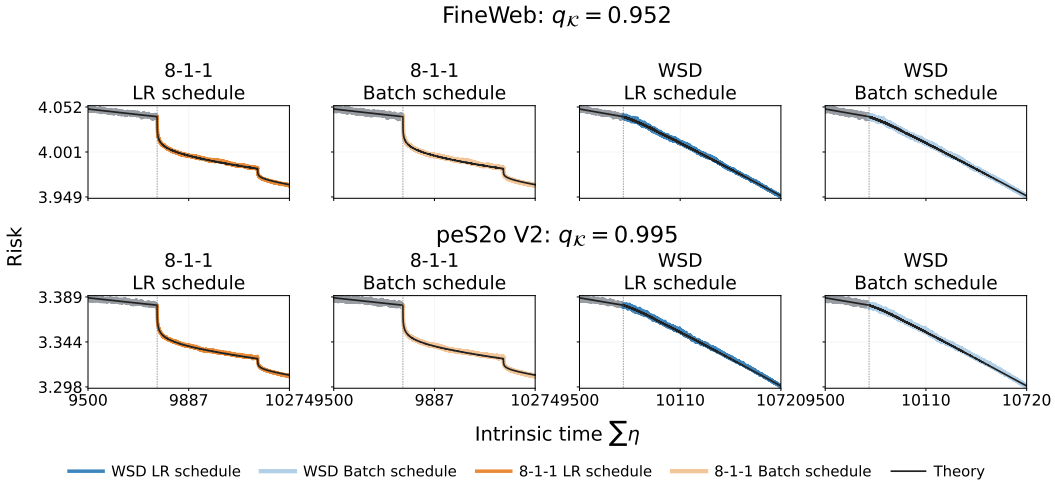


Figure 17: **Different datasets independently recover $q_{\mathcal{K}} \approx 1$.** Each surrogate is fitted only to fixed-batch 8-1-1 and transferred without refitting to the alternative factorization and both WSD trajectories. The fitted memory coordinates are 0.952 on FineWeb and 0.995 on peS2o V2; together with the OpenWebText value 1.017 in Figure 16, all lie within 0.05 of one. This supports dataset-level universality of the near-one effective response coordinate across the tested corpora. Risk is fixed-probe validation cross-entropy, $T = \sum_t \eta_t$, and absolute levels are dataset-specific.

$q_{\mathcal{F}} = 0.374$: the effective memory exponent lies close to the $q_{\mathcal{K}} = 1$ boundary, while the forcing exponent $q_{\mathcal{F}} < 1$.

To test whether this fitted coordinate is specific to one dataset, we repeat the identical fit-then-transfer analysis with separately trained 124M models on 2.5B-token subsets of the `sample-10BT` configuration of FineWeb (Penedo et al., 2024) and the peS2o V2 `s2orc` full-text corpus (Soldaini & Lo, 2023). The fits give $(q_{\mathcal{K}}, q_{\mathcal{F}}) = (0.952, 0.405)$ on FineWeb and $(0.995, 0.365)$ on peS2o, compared with $(1.017, 0.374)$ on OpenWebText above. Across web text and scientific full text, all three independently fitted values of $q_{\mathcal{K}}$ lie within 0.05 of one. This three-corpus agreement supports $q_{\mathcal{K}} \approx 1$ as a dataset-robust effective response coordinate, rather than an artifact of a particular pretraining corpus. In both new datasets, the frozen parameters also track the alternative 8-1-1 factorization and both WSD trajectories without refitting (Figure 17).

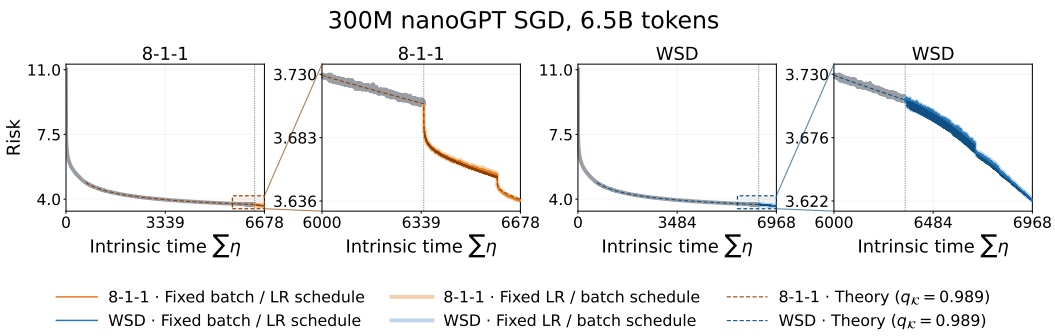


Figure 18: **The joint-schedule surrogate transfers at 300M scale.** The seven parameters are fitted only to the raw fixed-batch 8-1-1 validation trajectory. With those parameters frozen, the surrogate tracks the alternative 8-1-1 factorization and predicts both WSD trajectories without refitting. The fitted exponents are $q_\mathcal{K} = 0.989$ and $q_\mathcal{F} = 0.291$, again placing the effective surrogate coordinate close to $q_\mathcal{K} = 1$. Risk denotes fixed-probe validation cross-entropy and $T = \sum_t \eta_t$.

Protocol corresponding to the main-text experiment. This protocol uses a separately trained 300M plain-SGD nanoGPT model on a 6.5B-token OpenWebText corpus. For Figure 5, we fix $q_\mathcal{K} = 1$ and fit the remaining six parameters in Eq. (5.3) only to the raw fixed-batch 8-1-1 trajectory. These six parameters are then frozen when predicting WSD. For the profile in Figure 5(c), we repeat this procedure for each fixed $q_\mathcal{K}$: the remaining six parameters are fitted only on 8-1-1, after which the post-fork WSD prediction error is evaluated without refitting. This is a transfer-sensitivity profile rather than an independent estimate of the LLM memory exponent.

Unrestricted-fit robustness check. As a complementary check, we also allow all seven surrogate parameters to vary when fitting the same raw fixed-batch 8-1-1 trajectory. This unrestricted fit gives $q_\mathcal{K} = 0.989$ and $q_\mathcal{F} = 0.291$. With all seven parameters frozen, it also predicts the alternative 8-1-1 factorization and both WSD trajectories without refitting (Figure 18). The agreement between the restricted $q_\mathcal{K} = 1$ analysis and the unrestricted estimate $q_\mathcal{K} = 0.989$ supports using one as an effective finite-window coordinate, but does not independently identify an asymptotic memory exponent.

Summary. Together, the above three experiments in this section support a simple description of schedule-dependent LLM loss. At matched intrinsic time, changing the ratio path changes the validation trajectory; when the ratio path is fixed, different learning-rate–batch-size factorizations produce nearly the same trajectory after intrinsic-time alignment; and a forcing-memory surrogate fitted on one ratio path predicts unseen schedules without refitting. Thus, for the tested plain-SGD runs, the intrinsic clock $T = \sum_t \eta_t$ and the ratio path $r(T) = B(T)/\eta(T)$ organize the loss more consistently than optimizer step, learning rate, or batch size alone. The successful transfer across schedules, datasets, and model scales provides response-level evidence that the forcing–memory mechanism remains useful beyond the random-feature proxy. Across datasets and model scales, the fitted exponents consistently satisfy $q_\mathcal{K} \approx 1$ and $q_\mathcal{F} < 1$, placing the tested LLMs near the $\mathrm{LM}_1/\mathrm{IM}_1$ boundary of our phase map. The Muon results similarly suggest optimizer-specific schedule coordinates, but are not implied by our plain-SGD theory.