MODEL MERGING RESHAPES LEARNING DYNAMICS: A FUNCTIONAL SCALING LAW PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study model merging in power-law kernel regression, a tractable setting for understanding how checkpoint averaging reshapes learning dynamics. Specifically, we establish a functional scaling law (FSL) for fractional averaging in intrinsic time under general learning rate schedules, decomposing merged loss dynamics into signal learning and noise forgetting. Compared with the last iterate, merging increases signal error and can saturate its decay exponent. Its effect on noise depends on the injection time: merging retains more early noise but suppresses late noise, making its overall benefit depend on the learning rate schedule. The key mechanism is that averaging retains less-trained checkpoints, while late noise affects only the checkpoints following its injection. This characterization yields implications for merging design: tail and power averaging mitigate signal saturation, power averaging corresponds to a dynamic exponential moving average with a growing window, and sufficiently dense checkpoints preserve the noise-reduction rate of continuous averaging. Controlled SGD experiments in kernel regression empirically support the predicted data-scaling rates, optimal-window trends, and checkpoint-count criterion. Large language model (LLM) pretraining experiments show cross-schedule prediction without refitting and illustrate how the inferred signal–noise tradeoff explains observed merging gains.

1 INTRODUCTION

Model merging is increasingly explored for improving training efficiency and monitoring progress in large language model (LLM) pretraining. Here, merging means averaging checkpoints from a single training trajectory: if $\boldsymbol{\theta}_k$ denotes the parameters after k updates, the merged model is $\bar{\boldsymbol{\theta}}_k = \sum_{i=0}^{k} w_{k,i}\boldsymbol{\theta}_i$, with nonnegative weights summing to one.

Recent work uses checkpoint averaging to recover performance gains associated with terminal learning rate (LR) decay (Li et al., 2025b; Tian et al., 2025), estimate annealed performance during training (DeepSeek-AI, 2024), and combine averaging with horizon-free or moderately decaying LR schedules (Meterez et al., 2026; Luo et al., 2025a). These applications raise a common question:

Question. *How does model merging reshape learning dynamics, and how do its effects depend on the learning rate scheduling?*

Classical averaging theory gives convergence guarantees (Polyak & Juditsky, 1992; Dieuleveut & Bach, 2016; Mücke et al., 2019), and recent pretraining studies analyze its interaction with LR scheduling (Meterez et al., 2026; Ousherovitch & Wang, 2026). We seek a quantitative account of whole loss trajectories that distinguishes the cost of less-trained checkpoints from the effect of their shared noise. Final convergence rates alone do not determine when merging helps or select the lowest-risk rule within a rate-optimal family.

We address these questions within the functional scaling law (FSL) framework (Li et al., 2025a; 2026). FSL uses accumulated LR as *intrinsic time* and expresses loss dynamics through signal learning and noise forgetting. For fractional averaging, which allocates weights according to relative

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.

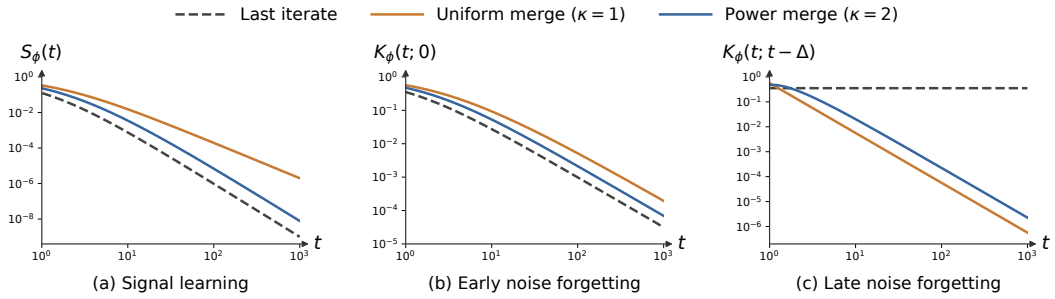


Figure 1: **Theoretical illustration of merging dynamics.** Last iterate, uniform merging ($\phi(u) = 1$), and power merging ($\phi(u) = 2u$), using $S(t) = (1+t)^{-3}$ and $K(u) = (1+u)^{-3/2}$. **(a)** Uniform merging saturates at exponent two; power merging preserves exponent three with a constant penalty. **(b)** Merging retains more initial noise. **(c)** At a fixed lag $\Delta = 1$, merged responses decay as t^{-2} while the last-iterate response is constant. These are evaluations of the theoretical functions, not independent training experiments.

intrinsic time, MM-FSL describes the expected excess risk of the merged model as

$$\overline{\mathcal{E}}_\phi(t) = \underbrace{S_\phi(t)}_{\text{signal learning}} + \underbrace{\int_0^t K_\phi(t;\tau)\varphi(\tau)\,\mathrm{d}\tau}_{\text{noise forgetting}}. \tag{1.1}$$

Here ϕ specifies the averaging profile, φ is the LR in intrinsic time, and $K_\phi(t;\tau)$ measures the remaining effect of noise injected at τ. The description applies throughout training and serves as a surrogate for LLM loss dynamics. Unlike the last-iterate kernel, the merged kernel generally depends separately on evaluation and injection times.

Our contributions follow three steps:

- **First, dynamics and mechanisms:** we derive the exact averaging transformations of the signal function and forgetting kernel. They reveal a signal cost, increased initial-noise retention, and late-noise suppression (Figure 1). The LR schedule determines how these effects combine.
- **Second, design beyond rate optimality:** we characterize risk-optimal tail fractions and power exponents within their respective families, after joint LR tuning. A common effective window law explains both choices. We also connect power averaging to dynamic EMA and quantify the checkpoint density needed to retain continuous averaging's benefit.
- **Third, empirical validation:** controlled SGD experiments test data efficiency, risk-minimizing profiles, and checkpoint density in kernel regression. In LLM pretraining, fitting on WSD predicts merged losses across LR schedules without refitting; a signal–noise decomposition and budget comparison further explain the observed merging gains.

2 Related Work

Merging in LLM pretraining. Checkpoint averaging has been explored as an alternative to terminal LR decay (Li et al., 2025b; Tian et al., 2025), estimates annealed performance (DeepSeek-AI, 2024), and complements LR schedules (Luo et al., 2025a; Meterez et al., 2026; Ousherovitch & Wang, 2026). We quantify how these choices jointly shape loss trajectories.

Averaging theory and profile design. Convergence and bias–variance analyses establish the benefits of averaging (Polyak & Juditsky, 1992; Dieuleveut & Bach, 2016; Zou et al., 2021), including removal of saturation by tail averaging (Mücke et al., 2019). Beyond rates, we optimize asymptotic risk within tail and power families after LR tuning, resolving constants and lower-order terms that distinguish their profiles. Further background, including FSL, appears in Appendix F.

3 PRELIMINARIES

Setting and notation. We write $f \eqsim g$ for equivalence up to positive constants independent of the time or budget under consideration, $f \lesssim g$ for a one-sided bound, and $f \sim g$ when $f/g \to 1$. Following Li et al. (2025a; 2026), consider a feature map $\boldsymbol{\phi} : \mathcal{X} \to \ell^2$ with covariance $\mathbf{H} = \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top]$. In its eigenbasis, samples obey $y = f^*(\mathbf{x}) + \epsilon$, where ϵ is independent, centered, and has unit variance, and $f^* = \sum_j \theta_j^* \phi_j$ in prediction norm. The excess risk of $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle$ is $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2} \sum_j \lambda_j (\theta_j - \theta_j^*)^2$, with

$$\lambda_j \eqsim j^{-q}, \qquad \lambda_j |\theta_j^*|^2 \eqsim j^{-1-sq}, \qquad q > 1, \quad s > 0. \tag{3.1}$$

Smaller capacity exponent q means slower spectral decay; smaller source exponent s means a harder target. We require finite prediction norm, not $\boldsymbol{\theta}^* \in \ell^2$.

Dynamics and intrinsic time. Starting from $\boldsymbol{\theta}_0 = 0$, one-pass SGD with batch size one uses LR η_k to update $\boldsymbol{\theta}_k$ to $\boldsymbol{\theta}_{k+1}$. Its accumulated LR after k updates is $t_k = \sum_{i=0}^{k-1} \eta_i$. For continuous training-step time z, define $t(z) = \int_0^z \eta(v)\, \mathrm{d}v$, its inverse $z(t)$, $\varphi(t) = \eta(z(t))$, and $T_N = t(N)$. Assume $\eta > 0$ almost everywhere during active training and $\int_0^t \varphi(\tau)\, \mathrm{d}\tau < \infty$ for finite t. Approximating the gradient-noise covariance by $\mathbf{H}$ in the label-noise-dominated regime gives

$$\mathrm{d}\vartheta_j(t) = -\lambda_j(\vartheta_j(t) - \theta_j^*)\, \mathrm{d}t + \sqrt{\lambda_j \varphi(t)}\, \mathrm{d}W_j(t), \qquad \vartheta_j(0) = 0, \tag{3.2}$$

with independent Brownian motions W_j. We analyze these continuous-time dynamics. For fixed batch size B and label-noise variance σ^2, its risk noise term is multiplied by σ^2/B; we normalize both to one.

Baseline FSL. The expected excess risk satisfies

$$\begin{aligned} \mathcal{E}(t) &= S(t) + \int_0^t K(t-\tau)\varphi(\tau)\, \mathrm{d}\tau, \\ S(t) &= \frac{1}{2} \sum_j \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t}, \qquad K(u) = \frac{1}{2} \sum_j \lambda_j^2 e^{-2\lambda_j u}. \end{aligned} \tag{3.3}$$

The signal function S is the gradient-flow excess risk; K describes how gradient noise is forgotten after injection. We use K for both the two-time kernel and its lag form: $K(t;\tau) = K(t-\tau)$ for the last iterate. Under (3.1), $S(t) \eqsim (c_S + t)^{-s}$ and $K(u) \eqsim (c_K + u)^{-\nu}$, with $\nu = 2 - 1/q \in (1, 2)$ and fixed $c_S, c_K > 0$. The factor $1/2$ is included in both S and K.

Fractional merging. An average with training-step density w_z has intrinsic-time density $a_t(u) = w_{z(t)}(z(u))/\varphi(u)$. We focus on the fractional family

$$a_t(u) = \frac{1}{t} \phi\Big(\frac{u}{t}\Big), \qquad \bar{\boldsymbol{\vartheta}}_\phi(t) = \int_0^1 \phi(v) \boldsymbol{\vartheta}(tv)\, \mathrm{d}v, \qquad \phi \geq 0, \quad \int_0^1 \phi(v)\, \mathrm{d}v = 1. \tag{3.4}$$

The profile allocates weights over relative learning progress, paralleling fractional LR parameterizations (Li et al., 2026), and is shared across evaluation times. We study $\overline{\mathcal{E}}_\phi(t) = \mathbb{E}[\mathcal{E}(\bar{\boldsymbol{\vartheta}}_\phi(t))]$ relative to the last iterate on the same trajectory. Fixing ϕ across LR schedules generally requires different training-step weights: intrinsic-time uniform merging is step-uniform only under constant LR. Throughout, φ denotes the LR and ϕ the merge profile.

4 FUNCTIONAL SCALING LAWS FOR MODEL MERGING

The following theorem transforms the baseline FSL into the merged-risk trajectory, accounting for shared noise across checkpoints. For a fixed intrinsic-time profile, S_ϕ and K_ϕ do not explicitly depend on the LR schedule: LR controls noise injection through φ and maps steps to intrinsic time through $t(z)$. This separation underlies cross-schedule prediction. See the full proof in Appendix A.

Theorem 1 (MM-FSL). *Under (3.1) and (3.2), every nonnegative normalized profile ϕ satisfies*

$$\overline{\mathcal{E}}_\phi(t) = S_\phi(t) + \int_0^t K_\phi(t;\tau)\varphi(\tau)\,\mathrm{d}\tau, \tag{4.1}$$

where, for $0 \le \tau \le t$,

$$\begin{aligned} S_\phi(t) &= \int_0^1 \int_0^1 \phi(u)\phi(v) S\left(\frac{t(u+v)}{2}\right) \mathrm{d}u\,\mathrm{d}v, \\ K_\phi(t;\tau) &= \int_{\tau/t}^1 \int_{\tau/t}^1 \phi(u)\phi(v) K\left(\frac{t(u+v)}{2} - \tau\right) \mathrm{d}u\,\mathrm{d}v. \end{aligned} \tag{4.2}$$

If additionally ϕ is positive and continuous on $(0,1)$ and $\phi(u) \sim c_0 u^{\kappa-1}$ as $u \downarrow 0$, with $c_0, \kappa > 0$, then for $t \ge 2$,

$$S_\phi(t) \asymp \begin{cases} t^{-s}, & s < 2\kappa, \\ t^{-s}\log t, & s = 2\kappa, \\ t^{-2\kappa}, & s > 2\kappa. \end{cases} \tag{4.3}$$

For each fixed $x \in (0,1)$ and $\nu = 2 - 1/q$,

$$K_\phi(t; xt) \asymp t^{-\nu} F_\phi(x), \qquad F_\phi(x) = \int_x^1 \int_x^1 \frac{\phi(u)\phi(v)}{(u+v-2x)^\nu}\,\mathrm{d}u\,\mathrm{d}v. \tag{4.4}$$

Constants are independent of t but may depend on the profile and, in (4.4), on x.

4.1 Signal Learning: Saturation and Constant Penalties

Characteristic times and saturation. Mode j has characteristic time $t_j^{\mathrm{char}} = \lambda_j^{-1} \asymp j^q$. At time t, the moving frontier $j_*(t) \asymp t^{1/q}$ separates learned modes ($\lambda_j t \gtrsim 1$) from slower modes that retain their initial error. Merging leaves these characteristic times unchanged but replaces the last iterate's exponential squared response with the spectral filter

$$h_\phi(z) = \left(\int_0^1 \phi(u) e^{-zu}\,\mathrm{d}u\right)^2 \asymp (1+z)^{-2\kappa}. \tag{4.5}$$

Indeed, checkpoints from a mode's initial learning period receive weight of order $(t_j^{\mathrm{char}}/t)^\kappa$. Their residual error creates a polynomial tail even after the current iterate has learned that mode. Splitting the spectral sum at $j_*(t)$ gives

$$S_\phi(t) \asymp \underbrace{t^{-2\kappa} \sum_{j \le j_*(t)} j^{q(2\kappa-s)-1}}_{\text{residual in learned modes}} + \underbrace{\sum_{j > j_*(t)} j^{-1-sq}}_{\text{unlearned tail}}. \tag{4.6}$$

The unlearned tail contributes t^{-s}; the learned modes dominate when $s > 2\kappa$, producing saturation at 2κ. At equality their harmonic sum adds the logarithm. Uniform merging therefore caps the exponent at two, whereas power profiles can raise this threshold.

The cost persists without saturation. Monotonicity of S in (4.2) gives $S_\phi(t) \ge S(t)$. If $S(t) \sim C_S t^{-s}$ and $s < 2\kappa$, then for independent U, V with density ϕ,

$$\lim_{t\to\infty} \frac{S_\phi(t)}{S(t)} = \mathbb{E}\left[\left(\frac{U+V}{2}\right)^{-s}\right] \ge (\mathbb{E}U)^{-s} \ge 1. \tag{4.7}$$

The penalty is strictly greater than one whenever positive weight precedes the evaluation time; equality corresponds to a point mass at the last iterate. For example, $\phi(u) = 2u$ preserves exponent $s = 3$ but has an eightfold asymptotic signal penalty (Figure 1(a)). Moving weight toward later checkpoints reduces this cost.

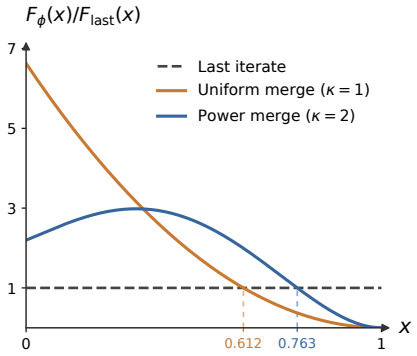
4.2 Noise Forgetting and the Net Benefit of Merging

Initial retention and late suppression. For noise injected at $\tau = 0$, the same monotonicity argument gives $K_\phi(t;0) \geq K(t)$. Its decay follows (4.3) with s replaced by ν: initial noise experiences the same saturation mechanism as signal error. Uniform merging and power merging with $\kappa = 2$ preserve the exponent $\nu \in (1,2)$ but increase the retained noise. By contrast, for a fixed lag $\Delta > 0$ and finite positive terminal density $\phi(u) \to c_1$ as $u \uparrow 1$,

$$K_\phi(t; t-\Delta) \eqsim t^{-2}, \qquad K_{\text{last}}(t; t-\Delta) = K(\Delta). \tag{4.8}$$

Only checkpoints after injection contain this late noise. Their total weight is of order Δ/t, whose square sets the risk suppression. The constant is proportional to c_1^2; power merging has $c_1 = \kappa$. Appendix A.1 gives the endpoint proofs and the full fixed-lag expression.

Noise at intermediate times. For fixed $x = \tau/t \in (0,1)$, compare profiles through $F_\phi(x)/F_{\text{last}}(x)$, where $F_{\text{last}}(x) = [2(1-x)]^{-\nu}$. If $K(u) \sim C_K u^{-\nu}$, this is the limiting ratio $K_\phi(t; xt)/K(t(1-x))$. The accompanying plot uses $\nu = 3/2$: values above one indicate increased noise retention, while values below one indicate suppression. Increasing κ from one to two reduces initial retention but delays the transition to suppression, from $x \approx 0.612$ to $x \approx 0.763$. The power-profile curve is nonmonotone: endpoint behavior alone does not determine the interior response. Appendix A.1 derives the ratio limit.



When does merging help? Subtracting merged risk from last-iterate risk makes the interaction with LR scheduling explicit:

$$\mathcal{E}(t) - \overline{\mathcal{E}}_\phi(t) = \underbrace{\int_0^t [K(t-\tau) - K_\phi(t;\tau)]\varphi(\tau)\,\mathrm{d}\tau}_{\text{net noise reduction}} - \underbrace{[S_\phi(t) - S(t)]}_{\text{signal cost}}. \tag{4.9}$$

Merging helps when net noise reduction exceeds signal cost. The LR schedule weights retained and suppressed noise through $\varphi(\tau)$. At the same intrinsic time and with the same profile, schedules share the signal contribution; at the same training step, they can also differ in accumulated intrinsic time and hence signal progress. LR decay changes noise injection, while merging changes how past injections contribute to the output. A profile's gain therefore depends on the LR schedule.

5 Implications for Designing Merging Strategies

We now turn the dynamics into design rules for window length, weighting, and checkpoint count. Consider N updates with constant LR within each run, so $T_N = \eta_0 N$, and define $s_c = \nu - 1 = 1 - 1/q$. We allow joint tuning of the averaging rule and $0 < \eta_0 \leq \eta_{\max}$ across budgets. Thus, constant LR does not mean the same LR for every N.

5.1 Risk-Optimal Tail and Power Averaging

Starting from uniform merging. The exact kernel transformation simplifies to

$$K_{\text{unif}}(t;\tau) = \frac{1}{2t^2}\sum_j \left(1 - e^{-\lambda_j(t-\tau)}\right)^2 \eqsim \frac{(t-\tau)^2}{t^2(1+t-\tau)^\nu}. \tag{5.1}$$

Its integral gives noise risk of order $\eta_0 t^{-s_c}$. Thus full uniform merging removes the constant-LR noise floor, but the signal exponent saturates: $\overline{\mathcal{E}}_{\text{unif}}(T_N) \eqsim T_N^{-\min\{s,2\}} + T_N^{1/q}/N$, with a logarithmic correction at $s = 2$. At a fixed LR the slower noise decay can mask this signal saturation; after tuning LR across budgets, the cap limits how much target regularity can improve data efficiency.

Two ways to control the signal cost. Tail and power profiles reduce the weight on early checkpoints:

$$\begin{aligned} \text{Tail: } \phi_r(u) &= \frac{\mathbf{1}_{[1-r,1]}(u)}{r}, \quad L_{\text{eff}} = rT_N, \quad 0 < r \le 1, \\ \text{Power: } \phi_\kappa(u) &= \kappa u^{\kappa-1}, \qquad\quad L_{\text{eff}} = T_N/\kappa, \quad \kappa > 0. \end{aligned} \tag{5.2}$$

Here r is the *merge ratio*, the terminal fraction of intrinsic time, analogous to the decay ratio in Li et al. (2026). Under constant LR it also equals the fraction of training updates; $r = 1$ gives full uniform merging. For power profiles, L_{eff} is a window scale. Every fixed $0 < r < 1$ eliminates signal saturation, recovering the established benefit of tail averaging (Mücke et al., 2019). Power merging preserves the signal exponent when $2\kappa > s$. For fixed $0 < r < 1$, or $2\kappa > \max\{s, s_c\}$, both families give risk of order $T_N^{-s} + T_N^{1/q}/N$. LR tuning therefore yields data-scaling exponents s for $s < s_c$ and $sq/(sq+1)$ for $s > s_c$. For $s > 2$, full uniform merging is limited to the exponent $2q/(2q+1)$, whereas an admissible tail or power profile attains $sq/(sq+1)$. Removing saturation therefore changes the data-scaling exponent for sufficiently regular targets. Within either nonsaturating family, however, the common rate does not select the lowest-risk profile.

To resolve that choice, strengthen the spectral assumptions to $\lambda_j \sim c_\lambda j^{-q}$ and $\lambda_j|\theta_j^*|^2 \sim c_* j^{-1-sq}$, with $c_\lambda, c_* > 0$. For fixed $0 < r < 1$, or $2\kappa > \max\{s, s_c\}$, write the leading risk as

$$\overline{\mathcal{E}}_\phi(T_N) = A_s(\phi) T_N^{-s}(1 + o(1)) + B_{s_c}(\phi)\frac{T_N^{1/q}}{N}(1 + o(1)). \tag{5.3}$$

The coefficients are computable from the profile. For independent U, V with density ϕ,

$$\begin{aligned} A_s(\phi) &= 2^s C_S\, \mathbb{E}[(U+V)^{-s}], \\ B_{s_c}(\phi) &= C_N\, \mathbb{E}\big[|U-V|^{-s_c} - (U+V)^{-s_c}\big], \end{aligned} \tag{5.4}$$

where $C_S = \lim_{t\to\infty} t^s S(t)$ and $C_N = \lim_{t\to\infty} \frac{1}{4} t^{s_c} \sum_j \lambda_j e^{-\lambda_j t}$ depend only on the spectrum and target. The first coefficient measures the cost of averaging earlier signal states. The second accounts for noise correlations between checkpoints, including the zero-initialization correction. It cannot be replaced by an independent-checkpoint variance estimate. Appendices B.1 and C.1 reduce these expectations to explicit functions of r and κ.

Theorem 2 (Risk-optimal averaging windows)**.** *Under the strengthened spectral assumptions, optimize the profile parameter and $0 < \eta_0 \le \eta_{\max}$ within each of the tail and power families. For hard tasks, $s < s_c$, the risk-minimizing choices satisfy $\eta_0^* \to \eta_{\max}$ and*

$$r_N^* \asymp (\kappa_N^*)^{-1} \asymp N^{-\frac{s_c - s}{1+s_c}}, \qquad L_{\text{eff}}^* \asymp N^{\frac{1+s}{1+s_c}}. \tag{5.5}$$

For easy tasks, $s > s_c$, an asymptotically risk-optimal fixed profile minimizes

$$\phi^* \in \arg\min_{\phi \in \mathcal{P}} A_s(\phi) B_{s_c}(\phi)^{sq}, \tag{5.6}$$

where $\mathcal{P}$ is either the tail family with finite A_s (including $r = 1$ when $s < 2$), or the power family with $\kappa > s/2$. The power-family objective has a finite interior minimizer.

One window law, two implementations. For hard tasks, optimal averaging retains a shrinking fraction of training but an increasing amount of intrinsic time. Its signal penalty above the exact gradient-flow baseline $S(T_N)$ has order $T_N^{-s} L_{\text{eff}}/T_N$, while its noise cost has order $\eta_0 L_{\text{eff}}^{-s_c}$. Balancing these terms gives (5.5). This refines the risk beyond its leading N^{-s} term. For easy tasks, optimizing T_N in (5.3) gives leading risk proportional to $[A_s(\phi)B_{s_c}(\phi)^{sq}]^{1/(sq+1)} N^{-sq/(sq+1)}$; profile selection minimizes its constant. The easy-regime criterion thus compares risk constants after LR tuning; the hard-regime result optimizes the correction above $S(\eta_{\max} N)$. Optimality is within the stated families. Optimal LRs, exact prefactors, proofs, and the case $s = s_c$ appear in Appendices B and C.

5.2 Power Merging as Dynamic EMA

Power merging has an online implementation. For a fixed exponent κ, differentiating the average gives

$$\frac{\mathrm{d}\bar{\boldsymbol{\vartheta}}_\kappa(t)}{\mathrm{d}t} = \frac{\kappa}{t}(\boldsymbol{\vartheta}(t) - \bar{\boldsymbol{\vartheta}}_\kappa(t)), \qquad \frac{\mathrm{d}\bar{\boldsymbol{\theta}}_\kappa(z)}{\mathrm{d}z} = \frac{\kappa\eta(z)}{t(z)}(\boldsymbol{\theta}(z) - \bar{\boldsymbol{\theta}}_\kappa(z)). \tag{5.7}$$

Its update rate decreases as the intrinsic-time window grows. With right-endpoint discretization, the exact recursion is

$$\bar{\boldsymbol{\theta}}_{\kappa,k} = \rho_k \bar{\boldsymbol{\theta}}_{\kappa,k-1} + (1-\rho_k)\boldsymbol{\theta}_k, \qquad \rho_k = \left(\frac{t_{k-1}}{t_k}\right)^{\kappa}. \tag{5.8}$$

Equivalently, checkpoint i receives weight $(t_i^\kappa - t_{i-1}^\kappa)/t_k^\kappa$. At $\kappa = 1$, this is η_{i-1}/t_k, so uniformity is in intrinsic time rather than optimizer steps. Decay makes the step-time update rate smaller, which matters when interpreting merged trajectories under nonconstant LR.

The retention rule $2^{-f/k}$ in anytime pretraining (Meterez et al., 2026) has update coefficient asymptotically $(f \log 2)/k$ for $f > 0$. It corresponds to $\kappa = f \log 2$ under constant LR and asymptotically $\kappa = 2f \log 2$ under inverse-square-root LR. This explains its growing-window structure, without predicting the empirically selected f. The budget-dependent κ_N^* in Theorem 2 is fixed within a run; substituting a changing κ_k online generally defines a different profile.

Why a fixed window behaves differently. With fixed nonzero LR and fixed intrinsic-time EMA rate $\gamma > 0$, the noise risk converges to a positive constant (Proposition 1 in Appendix C.6). Fixed EMA can lower this floor but cannot eliminate it, whereas growing-window averaging continues to reduce noise. This statement does not cover families that retune the LR or EMA rate as the budget increases.

5.3 How Many Checkpoints Are Needed?

Pretraining studies find that checkpoint spacing and window length both affect merging (Sanyal et al., 2024; Li et al., 2025b; Tian et al., 2025). To separate them, fix an intrinsic-time window $[t-L, t]$ and average its m equally spaced right endpoints: $\bar{\boldsymbol{\vartheta}}_m(t) = m^{-1}\sum_{i=1}^m \boldsymbol{\vartheta}(t - L + iL/m)$. Let $\mathcal{N}_m$ and $\mathcal{N}_{\mathrm{cont}}$ denote the noise risks of this average and the continuous uniform average over the same window.

Theorem 3 (Checkpoint count and noise reduction)**.** *Under the constant-LR SDE with deterministic initialization and $\lambda_j \eqsim j^{-q}$, for $t \geq L \geq 2$,*

$$0 \leq \mathcal{N}_m(t) - \mathcal{N}_{\mathrm{cont}}(t) \leq C\frac{\eta_0}{m}, \qquad \mathcal{N}_m(t) \eqsim \eta_0\left(L^{-s_c} + \frac{1}{m}\right). \tag{5.9}$$

Constants are independent of t, L, m. The continuous noise scale is preserved if and only if $m \gtrsim L^{s_c}$, equivalently $h = L/m \lesssim L^{1/q}$. If $m/L^{s_c} \to \infty$, then $\mathcal{N}_m/\mathcal{N}_{\mathrm{cont}} \to 1$.

Slow modes remain correlated across the window and contribute L^{-s_c}; finitely many observations leave a further $1/m$ noise scale. Thus averaging every update is unnecessary, but a fixed checkpoint count retains a positive noise floor under fixed LR. The sufficient count grows sublinearly with window length. For example, $q = 2$ requires only $m \gtrsim \sqrt{L}$ to preserve the noise rate; choosing $m/\sqrt{L} \to \infty$ also preserves its leading value. Thus the averaging window can expand while checkpoint collection becomes less frequent in intrinsic time. For a fixed fractional tail, $m \gg L^{s_c}$ also preserves the leading total risk. Under constant LR, $h = \eta_0 \Delta k$ for a step interval Δk; this correspondence is not uniform under varying LR. Appendix D provides the proof.

6 Experiments

We first test the design predictions of Section 5 with controlled SGD experiments, then assess MM-FSL in LLM pretraining through prediction, decomposition, and budget scaling. Full protocols appear in Appendix E.

6.1 Controlled Kernel Regression

We run sample-based SGD with Gaussian power-law features, unit label noise, and 512 trajectories per configuration, evaluating population excess risk after averaging. Figure 2 tests three consequences of MM-FSL, using constant LR within each run and tuning it across budgets for panels (a)–(c).

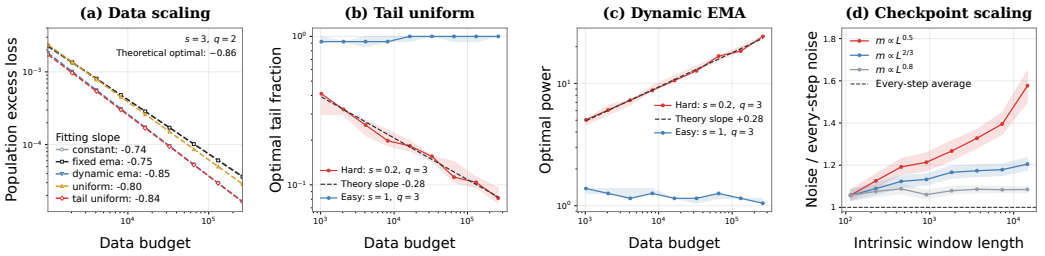


Figure 2: **Controlled SGD tests of merging design.** **(a)** Fitted data-scaling curves after LR selection. **(b,c)** Selected tail ratios and power exponents; dashed slopes follow Theorem 2. **(d)** Pure-noise risk relative to every-step averaging for $q = 3$. Panel (a) shows fits, not raw observations; grids, fitting ranges, and bootstrap bands are detailed in Appendix E.1.

Data efficiency and optimal profiles. For $s = 3, q = 2$, uniform merging has fitted decay exponent 0.80, matching the saturation-limited rate $4/5$. Tail and dynamic EMA reach 0.84 and 0.85, close to the nonsaturating rate $6/7$, compared with 0.74 for the last iterate and 0.75 for fixed EMA. Panels (b,c) resolve design choices within the tail and power families. With $q = 3$ and $s = 0.2 < s_c = 2/3$, selected merge ratios shrink and power exponents grow, following the predicted $N^{\mp 0.28}$ trends. For $s = 1 > s_c$, both remain of constant order, consistent with the regime-dependent window law.

Checkpoint density. Panel (d) compares $m \propto L^a$ for $a = 0.5, 2/3, 0.8$. Below the predicted threshold $s_c = 2/3$, the noise penalty relative to every-step averaging grows appreciably; the two denser rules keep it much smaller over the measured range. This supports the checkpoint-count criterion of Theorem 3 in discrete SGD.

6.2 Predicting Merged Loss across LR Schedules

We use 124M LLaMA-style (Touvron et al., 2023) dense model pretrained on FineWeb-Edu dataset (Penedo et al., 2024); full configurations are in Appendix E. We consider intrinsic-time uniform merging and dynamic EMA ($\kappa = 2$). For each averaging profile separately, we fit MM-FSL on WSD and predict cosine and constant-LR losses without refitting. Prediction changes only the LR-dependent intrinsic clock and noise-injection intensity, holding that profile and its fitted parameters fixed.

Figure 3 captures cosine flattening and continued constant-LR improvement for both profiles, supporting cross-schedule prediction with a fixed intrinsic-time profile.

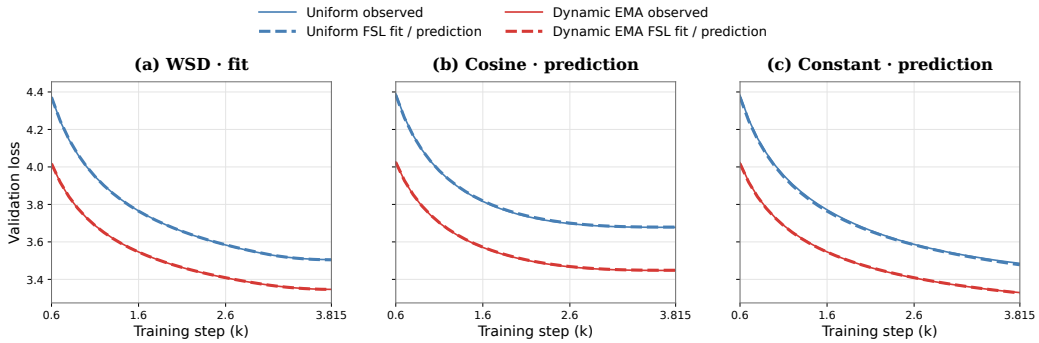


Figure 3: **Cross-schedule prediction without refitting.** WSD calibrates MM-FSL; cosine and constant LR are prediction targets. Blue denotes intrinsic-time uniform merging and red dynamic EMA. Solid curves show observed validation losses; dashed curves show fits or predictions.

6.3 Interpreting the Signal–Noise Tradeoff

We compare constant-LR dynamic EMA ($\kappa = 25$) with unaveraged constant, cosine, and WSD training in Figure 4.

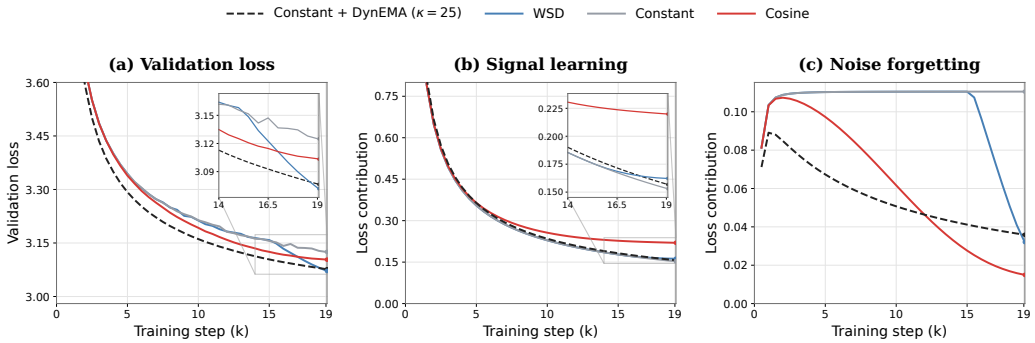


Figure 4: **Interpreting merging gains with MM-FSL.** Validation loss and inferred signal–noise contributions; insets highlight late training.

Merging on the same trajectory. Dynamic EMA slightly increases the inferred signal term by retaining less-trained checkpoints, but its noise term keeps decreasing as the constant-LR last-iterate term plateaus. The noise reduction outweighs the signal cost over the displayed range, consistent with (4.9). Merging thus improves loss without altering the LR schedule.

Comparison across LR schedules. Cosine has the smallest terminal noise term but higher total loss than dynamic EMA because its signal term is larger. WSD preserves signal learning during its stable phase, then sharply reduces noise, closing its loss gap to dynamic EMA. Their comparable final losses arise from different paths: terminal decay versus sustained noise reduction through a growing average. This explains why relative performance can change with the evaluation horizon. The components are MM-FSL estimates, not independently measured losses; they interpret the observed trajectories.

6.4 Dynamic EMA across Training Budgets

Figure 5 compares fixed EMA ($\beta = 0.99$) and dynamic EMA ($\kappa = 50$) over 5B–20B tokens. Dynamic EMA has lower final loss at every budget, with shared-floor fitted exponents 0.62 versus 0.51. LR and EMA selection protocols are recorded in the appendix. The comparison supports the displayed configurations over this range; it does not establish superiority over a separately tuned fixed-EMA family or test asymptotic noise floors. In particular, a shared fitted loss floor describes the finite-budget curves but does not verify the distinct noise floors in Proposition 1. Quantitative profile optimality in LLMs remains to be tested.

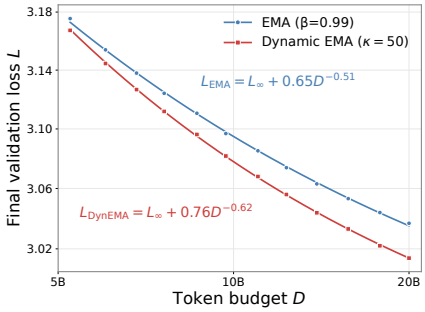


Figure 5: **EMA across budgets.**

7 Conclusion

MM-FSL quantifies the signal–noise tradeoff of averaging across LR schedules, yielding risk-optimal windows and a checkpoint-count criterion. Controlled SGD supports these design predictions; LLM experiments demonstrate cross-schedule prediction and interpret merging gains through signal–noise dynamics. Extending the theory to adaptive optimizers and testing quantitative profile optimality in LLMs are natural next steps.

AI USE STATEMENT

In this work, we used generative AI tools to assist with modifying scientific figures and tables, improving manuscript readability, checking grammar, editing software code, and searching for information and relevant literature. We reviewed all AI-assisted work, including verifying that the revised figures and tables accurately represented the underlying data, ensuring that language edits preserved the intended scientific meaning, reviewing AI-assisted code for correctness, and checking the cited literature and associated claims against the original sources. We take full responsibility for the final content of this work, including all text, claims, code, and figures produced with the assistance of generative AI.

REFERENCES

Giuseppe Curci, Stefano Simonazzi, Andrea Molinari, and Andrea Zugarini. Practical guidelines for model merging in LLMs pre-training. In Yunyao Li, Georg Rehm, and Mei Tu (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1519–1532, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-394-4. doi: 10.18653/v1/2026.acl-industry.105. URL https://aclanthology.org/2026.acl-industry.105/.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016. doi: 10.1214/15-AOS1391. URL https://doi.org/10.1214/15-AOS1391.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Jean Kaddour. Stop wasting my time! saving days of ImageNet and BERT training with latest weight averaging. *arXiv preprint arXiv:2209.14981*, 2022. URL https://arxiv.org/abs/2209.14981.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.10196.

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, 2025a. URL https://papers.nips.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. *arXiv preprint arXiv:2602.06797*, 2026. URL https://arxiv.org/abs/2602.06797.

Yunshui Li, Yiyuan Ma, Shen Yan, Chaoyi Zhang, Jing Liu, Jianqiao Lu, Ziwen Xu, Mengzhao Chen, Minrui Wang, Shiyi Zhan, et al. Model merging in pre-training of large language models. *Advances in Neural Information Processing Systems*, 38:148152–148175, 2025b.

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based LLM pretraining. *arXiv preprint arXiv:2511.18903*, 2025a. URL https://arxiv.org/abs/2511.18903.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules, 2025b. URL https://arxiv.org/abs/2503.12811.

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

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, and Sham Kakade. Anytime pretraining: Horizon-free learning-rate schedules with weight averaging. *arXiv preprint arXiv:2602.03702*, 2026. URL https://arxiv.org/abs/2602.03702.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4d0b954f0bef437c29dfa73fafdf3fa5-Abstract.html.

Adam Ousherovitch and Yixin Wang. Terminal shrinkage averaging reveals a schedule-estimator interaction in llm pretraining. *arXiv preprint arXiv:2609.25482*, 2026.

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, 2024. URL https://arxiv.org/abs/2406.17557.

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

Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Nolan Miller. Training trajectories, mini-batch losses and the curious role of the learning rate. *arXiv preprint arXiv:2301.02312*, 2023.

Sunny Sanyal, Atula Tejaswi Neerkaje, Jean Kaddour, Abhishek Kumar, and Sujay Sanghavi. Early weight averaging meets high learning rates for LLM pre-training. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IA8CWtNkUr.

Minhak Song, Beomhan Baek, Kwangjun Ahn, and Chulhee Yun. Through the river: Understanding the benefit of schedule-free methods for language model training. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.09846.

Changxin Tian, Jiapeng Wang, Qian Zhao, Kunlong Chen, Jia Liu, Ziqi Liu, Jiaxin Mao, Wayne Xin Zhao, Zhiqiang Zhang, and Jun Zhou. WSM: Decay-free learning rate schedule via checkpoint merging for LLM pre-training. *arXiv preprint arXiv:2507.17634*, 2025. URL https://arxiv.org/abs/2507.17634.

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *Advances in Neural Information Processing Systems*, 38:101068–101104, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1806.04498.

Wenjie Zhou, Bohan Wang, Hongtao Zhang, Chenxi Jia, Wei Chen, and Xueqi Cheng. Extra-merge: Tracing the rank-1 subspace of model merging in language model pre-training. *arXiv preprint arXiv:2605.26484*, 2026.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *arXiv preprint arXiv:2103.12692*, 2021.

A Proof of Theorem 1

We first derive the exact transformation of the signal function and forgetting kernel, and then establish their power-law scaling. Throughout, the label-noise variance is one, and the factor $1/2$ from the excess risk is included in the definitions of both S and K.

Step 1: Solving and averaging the trajectory. For each spectral direction j, the solution of (3.2) satisfies

$$\vartheta_j(v) - \theta_j^* = -\theta_j^* e^{-\lambda_j v} + \sqrt{\lambda_j} \int_0^v e^{-\lambda_j (v-\tau)} \sqrt{\varphi(\tau)}\, \mathrm{d}W_j(\tau). \tag{A.1}$$

Define the averaged signal response and noise response by

$$L_\phi(z) := \int_0^1 \phi(u) e^{-zu}\, \mathrm{d}u, \qquad R_{\phi,j}(t;\tau) := \int_{\tau/t}^1 \phi(u) e^{-\lambda_j (tu-\tau)}\, \mathrm{d}u. \tag{A.2}$$

Since ϕ is a probability density, $0 < L_\phi(z) \le 1$ and $0 \le R_{\phi,j}(t;\tau) \le 1$.

Averaging (A.1) over $v = tu$ and exchanging the averaging integral with the stochastic integral gives

$$\bar{\vartheta}_{\phi,j}(t) - \theta_j^* = -\theta_j^* L_\phi(\lambda_j t) + \sqrt{\lambda_j} \int_0^t R_{\phi,j}(t;\tau) \sqrt{\varphi(\tau)}\, \mathrm{d}W_j(\tau). \tag{A.3}$$

The lower integration limit τ/t in $R_{\phi,j}$ reflects that noise injected at τ affects only checkpoints with $tu \ge \tau$.

The exchange is justified by stochastic Fubini. Indeed, for each fixed j and t,

$$\begin{aligned} &\int_0^1 \phi(u) \left(\int_0^{tu} \lambda_j \varphi(\tau) e^{-2\lambda_j (tu-\tau)}\, \mathrm{d}\tau \right)^{1/2} \mathrm{d}u \\ &\qquad \le \left(\lambda_j \int_0^t \varphi(\tau)\, \mathrm{d}\tau \right)^{1/2} < \infty. \end{aligned} \tag{A.4}$$

All calculations are coordinatewise and require only finite prediction norm of the target, not $\boldsymbol{\theta}^* \in \ell^2$.

Step 2: Exact signal and noise transformations. The stochastic integral in (A.3) has mean zero. Its cross term with the deterministic component therefore vanishes in expectation. Itô isometry yields

$$\mathbb{E}\Big[\big(\bar{\vartheta}_{\phi,j}(t) - \theta_j^* \big)^2 \Big] = |\theta_j^*|^2 L_\phi(\lambda_j t)^2 + \lambda_j \int_0^t R_{\phi,j}(t;\tau)^2 \varphi(\tau)\, \mathrm{d}\tau. \tag{A.5}$$

Multiplying by $\lambda_j/2$, summing over j, and applying Tonelli's theorem gives

$$\overline{\mathcal{E}}_\phi(t) = S_\phi(t) + \int_0^t K_\phi(t;\tau) \varphi(\tau)\, \mathrm{d}\tau, \tag{A.6}$$

where

$$S_\phi(t) = \frac{1}{2} \sum_{j \ge 1} \lambda_j |\theta_j^*|^2 L_\phi(\lambda_j t)^2, \qquad K_\phi(t;\tau) = \frac{1}{2} \sum_{j \ge 1} \lambda_j^2 R_{\phi,j}(t;\tau)^2. \tag{A.7}$$

These quantities are finite, since

$$\overline{\mathcal{E}}_\phi(t) \le \frac{1}{2} \sum_{j \ge 1} \lambda_j |\theta_j^*|^2 + \frac{1}{2} \left(\sum_{j \ge 1} \lambda_j^2 \right) \int_0^t \varphi(\tau)\, \mathrm{d}\tau < \infty. \tag{A.8}$$

Here, both spectral sums converge under (3.1).

Expanding the square in the signal term and using Tonelli's theorem again, we obtain

$$\begin{aligned} S_\phi(t) &= \int_0^1 \int_0^1 \phi(u)\phi(v) \left[\frac{1}{2} \sum_{j \ge 1} \lambda_j |\theta_j^*|^2 e^{-\lambda_j t(u+v)} \right] \mathrm{d}u\, \mathrm{d}v \\ &= \int_0^1 \int_0^1 \phi(u)\phi(v) S\left(\frac{t(u+v)}{2} \right) \mathrm{d}u\, \mathrm{d}v. \end{aligned} \tag{A.9}$$

Similarly, expanding the squared noise response gives

$$K_\phi(t;\tau) = \int_{\tau/t}^1 \int_{\tau/t}^1 \phi(u)\phi(v) \left[\frac{1}{2} \sum_{j\geq 1} \lambda_j^2 e^{-\lambda_j (t(u+v)-2\tau)} \right] \mathrm{d}u\, \mathrm{d}v$$
$$= \int_{\tau/t}^1 \int_{\tau/t}^1 \phi(u)\phi(v) K\left(\frac{t(u+v)}{2} - \tau \right) \mathrm{d}u\, \mathrm{d}v. \tag{A.10}$$

This proves the exact decomposition and both transformations in the theorem.

Step 3: A spectral summation estimate. We next establish the spectral estimate used to determine the signal exponent. For $q > 0$, $\mu > 1$, and $b > 0$, consider

$$A_{\mu,b}(t) := \sum_{j\geq 1} j^{-\mu}(1 + tj^{-q})^{-b}.$$

For $t \geq 2$, let $J = \lfloor t^{1/q} \rfloor$. Splitting the sum at J gives

$$A_{\mu,b}(t) \eqsim t^{-b} \sum_{j\leq J} j^{qb-\mu} + \sum_{j>J} j^{-\mu}. \tag{A.11}$$

The second sum satisfies $\sum_{j>J} j^{-\mu} \eqsim J^{1-\mu} \eqsim t^{-(\mu-1)/q}$. For the first sum,

$$\sum_{j\leq J} j^{qb-\mu} \eqsim \begin{cases} J^{qb-\mu+1}, & qb-\mu > -1, \\ \log(J+1), & qb-\mu = -1, \\ 1, & qb-\mu < -1. \end{cases}$$

Substituting these estimates into (A.11) yields

$$A_{\mu,b}(t) \eqsim \begin{cases} t^{-(\mu-1)/q}, & (\mu-1)/q < b, \\ t^{-b}\log t, & (\mu-1)/q = b, \\ t^{-b}, & (\mu-1)/q > b. \end{cases} \tag{A.12}$$

Thus, the decay is determined by the slower of the spectral exponent $(\mu-1)/q$ and the filter exponent b, with a logarithmic correction when they coincide.

Step 4: Signal scaling and saturation. The assumption $\phi(u) \sim c_0 u^{\kappa-1}$ near zero implies that, for some $u_0 \in (0,1)$ and constants $c_-, c_+ > 0$,

$$c_- u^{\kappa-1} \leq \phi(u) \leq c_+ u^{\kappa-1}, \qquad 0 < u \leq u_0.$$

For sufficiently large z, integrating over $[0, 1/z]$ gives

$$L_\phi(z) \geq c_- e^{-1} \int_0^{1/z} u^{\kappa-1}\, \mathrm{d}u = \frac{c_- e^{-1}}{\kappa} z^{-\kappa}. \tag{A.13}$$

Conversely,

$$L_\phi(z) \leq c_+ \int_0^{u_0} u^{\kappa-1} e^{-zu}\, \mathrm{d}u + e^{-zu_0} \int_{u_0}^1 \phi(u)\, \mathrm{d}u$$
$$\leq c_+ \Gamma(\kappa) z^{-\kappa} + e^{-zu_0} \lesssim z^{-\kappa}. \tag{A.14}$$

Since L_ϕ is positive and continuous with $L_\phi(0) = 1$, these estimates extend to

$$h_\phi(z) := L_\phi(z)^2 \eqsim (1+z)^{-2\kappa}, \qquad z \geq 0. \tag{A.15}$$

Combining (A.7), (A.15), and the spectral assumptions gives

$$S_\phi(t) \eqsim \sum_{j\geq 1} j^{-(1+sq)}(1 + tj^{-q})^{-2\kappa}. \tag{A.16}$$

Applying (A.12) with $\mu = 1 + sq$ and $b = 2\kappa$ proves

$$S_\phi(t) \eqsim \begin{cases} t^{-s}, & s < 2\kappa, \\ t^{-s}\log t, & s = 2\kappa, \\ t^{-2\kappa}, & s > 2\kappa. \end{cases} \tag{A.17}$$

The saturation threshold is therefore determined by the polynomial tail of the averaging filter.

Step 5: Kernel scaling at fixed relative injection time. Fix $x \in (0,1)$. From (A.10) and $K(w) \asymp (c_K + w)^{-\nu}$, we obtain

$$t^{\nu} K_{\phi}(t; xt) \asymp J_t(x) := \int_x^1 \int_x^1 \frac{\phi(u)\phi(v)}{(c_K/t + (u+v-2x)/2)^{\nu}} \, \mathrm{d}u \, \mathrm{d}v. \tag{A.18}$$

We first verify that

$$F_{\phi}(x) = \int_x^1 \int_x^1 \frac{\phi(u)\phi(v)}{(u+v-2x)^{\nu}} \, \mathrm{d}u \, \mathrm{d}v$$

is finite and positive.

The only singularity of the denominator occurs at $(u,v) = (x,x)$. Choose $\delta = (1-x)/2$. Continuity of ϕ implies that it is bounded on $[x, x+\delta]$, so the contribution from this corner is bounded by a constant times

$$\int_0^{\delta} \int_0^{\delta} (a+b)^{-\nu} \, \mathrm{d}a \, \mathrm{d}b \le \int_0^{2\delta} r^{1-\nu} \, \mathrm{d}r < \infty, \tag{A.19}$$

where we used $\nu = 2 - 1/q < 2$. Outside this corner square, $u + v - 2x \ge \delta$, and integrability follows from $\phi \in L^1(0,1)$. Positivity of ϕ on $(0,1)$ ensures $F_{\phi}(x) > 0$.

For $t \ge 2$, the regularized integrand in (A.18) increases with t and is bounded above by $2^{\nu}\phi(u)\phi(v)(u+v-2x)^{-\nu}$. Consequently,

$$0 < J_2(x) \le J_t(x) \le 2^{\nu} F_{\phi}(x), \qquad J_t(x) \longrightarrow 2^{\nu} F_{\phi}(x). \tag{A.20}$$

In particular, $J_t(x) \asymp F_{\phi}(x)$ for $t \ge 2$, with comparison constants allowed to depend on the fixed x and profile ϕ. Substituting this into (A.18) proves

$$K_{\phi}(t; xt) \asymp t^{-\nu} F_{\phi}(x).$$

Together with the exact decomposition and signal scaling established above, this completes the proof. □

A.1 Signal Penalties and Noise Endpoints

The signal constant. Let U, V be independent with density ϕ and put $M = (U+V)/2$. The exact transformation gives $S_{\phi}(t)/S(t) = \mathbb{E}[S(tM)/S(t)]$. If $S(t) \sim C_S t^{-s}$ and $s < 2\kappa$, then $\mathbb{E}[M^{-s}] < \infty$: near $U = V = 0$, the integrand has radial order $r^{2\kappa - s - 1}$. Moreover, regularized power-law bounds give $S(tM)/S(t) \le CM^{-s}$ for $t \ge 2$ and $0 < M \le 1$. Dominated convergence therefore proves (4.7). Jensen's inequality yields $\mathbb{E}[M^{-s}] \ge (\mathbb{E}U)^{-s}$. For $\phi(u) = 2u$ and $s = 3$, the ratio is $32 \int_0^1 \int_0^1 uv(u+v)^{-3} \, \mathrm{d}u \, \mathrm{d}v = 8$.

Initial noise. At $\tau = 0$, the spectral representation gives

$$K_{\phi}(t;0) = \frac{1}{2} \sum_j \lambda_j^2 h_{\phi}(\lambda_j t) \asymp \begin{cases} t^{-\nu}, & \nu < 2\kappa, \\ t^{-\nu} \log t, & \nu = 2\kappa, \\ t^{-2\kappa}, & \nu > 2\kappa. \end{cases} \tag{A.21}$$

Indeed, $\lambda_j^2 \asymp j^{-2q} = j^{-1-\nu q}$, so the spectral splitting argument above applies with s replaced by ν. Monotonicity of K gives $K_{\phi}(t;0) \ge K(t)$ directly from the transformation.

Noise a fixed lag before evaluation. Let $\Delta > 0$ be fixed and assume $\phi(u) \to c_1 \in (0, \infty)$ as $u \uparrow 1$. Set the two checkpoint times to $t - \Delta + a$ and $t - \Delta + b$, with $a, b \in [0, \Delta]$. Then

$$\begin{aligned} K_{\phi}(t; t-\Delta) &= \frac{1}{t^2} \int_0^{\Delta} \int_0^{\Delta} \phi\left(1 - \frac{\Delta - a}{t}\right) \phi\left(1 - \frac{\Delta - b}{t}\right) K\left(\frac{a+b}{2}\right) \, \mathrm{d}a \, \mathrm{d}b \\ &\sim \frac{c_1^2}{t^2} \int_0^{\Delta} \int_0^{\Delta} K\left(\frac{a+b}{2}\right) \, \mathrm{d}a \, \mathrm{d}b. \end{aligned} \tag{A.22}$$

Dominated convergence applies because ϕ is bounded near one and K is bounded on $[0, \Delta]$. The final integral is a strictly positive finite constant. For uniform merging the formula is exact with $c_1 = 1$; for power merging $c_1 = \kappa$.

Relative retention away from the endpoints. If $K(u) \sim C_K u^{-\nu}$, the corner bound used in the main proof also gives $t^\nu K_\phi(t; xt) \to C_K 2^\nu F_\phi(x)$ for fixed $x \in (0,1)$. The last-iterate response satisfies $t^\nu K(t(1-x)) \to C_K(1-x)^{-\nu}$. Their ratio is therefore $F_\phi(x)/F_{\text{last}}(x)$. The plot in Section 4.2 illustrates that this ratio need not decrease monotonically. Its vertical guides mark the crossings for uniform and power merging; the value at $x = 1$ is the limiting ratio. The fixed-x limit is not uniform at the endpoints and does not replace the fixed-lag calculation above.

B Tail Averaging: Risk Coefficients and Optimal Windows

This appendix derives the profile-dependent coefficients and proves the tail-averaging part of Theorem 2. Throughout, we assume

$$\lambda_j \sim c_\lambda j^{-q}, \qquad \lambda_j |\theta_j^*|^2 \sim c_* j^{-(1+sq)}, \qquad q > 1, \quad s > 0,$$

with $c_\lambda, c_* > 0$, and write $s_c = 1 - 1/q$. We use the normalization $\sigma = B = 1$ from Section 3. Under constant LR η_0, the total intrinsic time is $T = \eta_0 N$. The merge ratio $r \in (0,1]$ denotes the fraction of intrinsic-time progress included in the final averaging window: $\phi_r(u) = r^{-1}\mathbf{1}_{[1-r,1]}(u)$. Thus, its length is $L = rT$, and $r = 1$ gives full uniform averaging.

B.1 Exact Risk Representation and Leading Coefficients

The family notation below is related to the main text by $A_s(\phi_r) = A_s(r)$ and $B_{s_c}(\phi_r) = B_{s_c}(r)$.

Spectral asymptotics. The strengthened spectral assumptions give

$$S(t) \sim C_S t^{-s}, \qquad G(t) := \frac{1}{4}\sum_{j \geq 1} \lambda_j e^{-\lambda_j t} \sim C_N t^{-s_c}, \tag{B.1}$$

where

$$C_S = \frac{c_* \Gamma(s)}{2q(2c_\lambda)^s}, \qquad C_N = \frac{c_\lambda^{1/q}\Gamma(s_c)}{4q}. \tag{B.2}$$

These follow by rescaling the corresponding spectral sums. The same calculation applied to their derivatives yields

$$-tS'(t) \sim sC_S t^{-s}, \qquad -tG'(t) \sim s_c C_N t^{-s_c}. \tag{B.3}$$

In particular, the derivative asymptotics follow from the spectral representation; no separate second-order regularity assumption is needed.

Exact signal and noise terms. The signal transformation gives

$$S_r(T) = \frac{1}{r^2}\int_{1-r}^1 \int_{1-r}^1 S\left(\frac{T(u+v)}{2}\right) \mathrm{d}u\, \mathrm{d}v. \tag{B.4}$$

To compute the noise contribution, let $\xi_j(t)$ denote the stochastic component of $\vartheta_j(t)$. Under constant LR,

$$\mathbb{E}[\xi_j(a)\xi_j(b)] = \frac{\eta_0}{2}\left(e^{-\lambda_j|a-b|} - e^{-\lambda_j(a+b)}\right).$$

Consequently,

$$\begin{aligned} \overline{\mathcal{E}}_r(T) &= S_r(T) + \eta_0 V_r(T), \\ V_r(T) &= \frac{1}{r^2}\int_{1-r}^1 \int_{1-r}^1 \left[G(T|u-v|) - G(T(u+v))\right] \mathrm{d}u\, \mathrm{d}v. \end{aligned} \tag{B.5}$$

The second term inside the integral accounts for deterministic initialization. All sums and integrals are finite because $\sum_j \lambda_j < \infty$ and the initial prediction error is finite.

Fixed-fraction asymptotics. For each fixed $r \in (0,1)$,

$$S_r(T) \sim A_s(r) T^{-s}, \qquad V_r(T) \sim B_{s_c}(r) T^{-s_c}, \tag{B.6}$$

where

$$\begin{aligned} A_s(r) &= \frac{2^s C_S}{r^2} \int_{1-r}^{1} \int_{1-r}^{1} (u+v)^{-s} \, \mathrm{d}u \, \mathrm{d}v, \\ B_{s_c}(r) &= \frac{C_N}{r^2} \int_{1-r}^{1} \int_{1-r}^{1} \left[|u-v|^{-s_c} - (u+v)^{-s_c} \right] \mathrm{d}u \, \mathrm{d}v. \end{aligned} \tag{B.7}$$

The signal limit follows directly from (B.4). For the noise term, the only singularity in the limiting integrand is along $u = v$. It is integrable because $s_c < 1$. The bound $G(t) \lesssim (1+t)^{-s_c}$ therefore permits dominated convergence. The same conclusions extend to $r = 1$ when $s < 2$, which ensures integrability of the signal coefficient at $u = v = 0$.

Since $\eta_0 = T/N$, we obtain

$$\overline{\mathcal{E}}_r(T) = A_s(r) T^{-s}(1 + o(1)) + B_{s_c}(r) \frac{T^{1/q}}{N} (1 + o(1)). \tag{B.8}$$

The two remainders refer separately to the signal and noise contributions.

Explicit coefficient formulas. Define

$$J_p(r) = \frac{1}{r^2} \int_{1-r}^{1} \int_{1-r}^{1} (u+v)^{-p} \, \mathrm{d}u \, \mathrm{d}v.$$

For $p \neq 1, 2$ and $0 < r < 1$,

$$J_p(r) = \frac{2^{2-p} - 2(2-r)^{2-p} + [2(1-r)]^{2-p}}{(1-p)(2-p) r^2}. \tag{B.9}$$

For $r < 1$, the cases $p = 1, 2$ are obtained by continuous extension in p. At $r = 1$, $J_p(1)$ is finite precisely when $p < 2$. Thus,

$$A_s(r) = 2^s C_S J_s(r), \qquad B_{s_c}(r) = C_N \left[\frac{2 r^{-s_c}}{(1-s_c)(2-s_c)} - J_{s_c}(r) \right]. \tag{B.10}$$

Both coefficients are positive. As $r \downarrow 0$,

$$A_s(r) = C_S \left(1 + \frac{s}{2} r + O(r^2) \right), \qquad B_{s_c}(r) \sim b \, r^{-s_c}, \tag{B.11}$$

where

$$b = \frac{2 C_N}{(1-s_c)(2-s_c)}. \tag{B.12}$$

B.2 Easy Tasks: Minimizing the Leading Risk Constant

Suppose $s > s_c$. For a fixed admissible merge ratio r, minimizing the leading risk in (B.8) over T gives

$$T_r^* = \left(\frac{sq \, A_s(r)}{B_{s_c}(r)} N \right)^{\frac{q}{sq+1}}. \tag{B.13}$$

Because $s > s_c$, we have $T_r^*/N \to 0$, so the LR constraint is asymptotically inactive. Substitution gives

$$\inf_{0 < \eta_0 \le \eta_{\max}} \overline{\mathcal{E}}_r(\eta_0 N) \sim D_{s,q} \left[A_s(r) B_{s_c}(r)^{sq} \right]^{\frac{1}{sq+1}} N^{-\frac{sq}{sq+1}}, \tag{B.14}$$

where $D_{s,q} = (1+sq)(sq)^{-sq/(sq+1)}$ is independent of r. This proves the optimization criterion in (5.6).

Existence and admissible endpoints. As $r \downarrow 0$, $A_s(r) \to C_S > 0$ while $B_{s_c}(r) \to \infty$. The objective therefore diverges at this endpoint. At $r = 1$, it is finite when $s < 2$ and diverges when $s \geq 2$. Continuity on the remaining domain guarantees at least one minimizer. An interior minimizer satisfies

$$\frac{A_s'(r^*)}{A_s(r^*)} + sq\,\frac{B_{s_c}'(r^*)}{B_{s_c}(r^*)} = 0.$$

Neither uniqueness nor an interior minimizer is required for the theorem.

Budget-dependent fractions. Allowing the merge ratio to vary with N does not improve the leading optimum above. Indeed, uniformly over merge ratios, $S_r(T) \geq S(T)$ and $V_r(T) \gtrsim T^{-s_c}$ for sufficiently large T. The latter bound follows by restricting to spectral modes with $\lambda_j T$ of order one. Comparison with any fixed-fraction strategy then forces a risk-minimizing sequence to satisfy $T \asymp N^{q/(sq+1)}$.

Such a sequence cannot have $r \to 0$: if $rT \to \infty$, its noise coefficient diverges as r^{-s_c}; if rT remains bounded, $V_r(T)$ stays bounded away from zero. When $s \geq 2$, the sequence also cannot have $r \to 1$, since Fatou's lemma applied to (B.4) gives $T^s S_r(T) \to \infty$. On the remaining compact sets, the asymptotics in (B.6) are uniform. For $s < 2$, this uniformity includes $r = 1$ by dominated convergence. Hence the leading optimal risk is obtained by minimizing (B.14).

B.3 Hard Tasks: Optimizing a Shrinking Relative Window

Suppose $s < s_c$. Here, a fixed merge ratio already attains the rate N^{-s}. To determine the risk-minimizing merge ratio, we must resolve the additional cost above the last-iterate gradient-flow risk. Fixed-fraction asymptotics alone are insufficient because the optimal merge ratio depends on N.

Uniform shrinking-window expansion. Consider $T \to \infty$ and $r \to 0$, with $L = rT \to \infty$. Changing variables in the signal integral gives

$$S_r(T) = \int_0^1 \int_0^1 S\left(T\left[1 - \frac{r(a+b)}{2}\right]\right) \mathrm{d}a\,\mathrm{d}b.$$

Using (B.3) uniformly on $[T(1-r), T]$ yields

$$S_r(T) - S(T) \sim a\,T^{-s} r, \qquad a = \frac{sC_S}{2}. \tag{B.15}$$

For the noise term,

$$V_r(T) = \int_0^1 \int_0^1 \left[G(L|a-b|) - G(T[2 - r(a+b)])\right] \mathrm{d}a\,\mathrm{d}b. \tag{B.16}$$

The first term is asymptotic to bL^{-s_c}, where b is defined in (B.12). The second is of order $T^{-s_c} = o(L^{-s_c})$. Consequently,

$$V_r(T) \sim b(Tr)^{-s_c}. \tag{B.17}$$

Together,

$$\overline{\mathcal{E}}_r(T) - S(T) = aT^{-s} r(1 + o(1)) + b\eta_0 T^{-s_c} r^{-s_c}(1 + o(1)). \tag{B.18}$$

These expansions hold along any sequence satisfying $T \to \infty$, $r \to 0$, and $Tr \to \infty$.

Localization of the optimum. Taking $\eta_0 = \eta_{\max}$ and, for example, $r = N^{-(s_c - s)/(1+s_c)}$ gives

$$\overline{\mathcal{E}}_r(\eta_{\max} N) = S(\eta_{\max} N)(1 + o(1)).$$

For every admissible strategy, $\overline{\mathcal{E}}_r(T) \geq S(T) \geq S(\eta_{\max} N)$. Regular variation of S therefore forces any risk-minimizing sequence to satisfy $T/(\eta_{\max} N) \to 1$, or $\eta_0^* \to \eta_{\max}$.

The same comparison implies $r_N^* \to 0$: a retained fraction bounded away from zero incurs a strictly positive asymptotic signal penalty relative to $S(T)$. Moreover, $Tr_N^* \to \infty$, since otherwise the noise contribution stays bounded away from zero while $\eta_0^* \to \eta_{\max} > 0$. Thus, the optimal sequence lies in the regime of (B.18).

The LR constraint is active. For completeness, the optimum actually reaches the LR bound for all sufficiently large budgets. In the localized regime, differentiation of the exact risk at fixed r gives

$$\frac{\partial}{\partial \eta_0}\overline{\mathcal{E}}_r(\eta_0 N) = NS_r'(T) + V_r(T) + TV_r'(T).$$

The spectral derivative asymptotics imply $TS_r'(T) \sim -sS(T)$ and $TV_r'(T) \sim -s_c V_r(T)$. At a minimizing sequence, $\eta_0 V_r(T) = o(S(T))$, so this derivative is negative for sufficiently large N. An interior LR cannot therefore minimize the risk, and $\eta_0^* = \eta_{\max}$ eventually.

Optimal window and risk refinement. For fixed $T = \eta_{\max} N$, the leading excess risk in (B.18) is minimized at

$$\widehat{r}_N = \left(\frac{s_c b \eta_{\max}}{a} T^{s-s_c}\right)^{\frac{1}{1+s_c}}. \tag{B.19}$$

To see that this also identifies the true asymptotic optimizer, set $r = y\widehat{r}_N$. After division by $aT^{-s}\widehat{r}_N$, the leading excess becomes

$$y + \frac{1}{s_c} y^{-s_c}.$$

This function has a unique minimum at $y = 1$ and diverges at both endpoints. The uniform expansions above therefore give

$$r_N^* \sim \widehat{r}_N \eqsim N^{-\frac{s_c - s}{1+s_c}}, \qquad T r_N^* \eqsim N^{\frac{1+s}{1+s_c}}, \tag{B.20}$$

proving (5.5).

More precisely, the optimal risk satisfies

$$\inf_{\substack{0<\eta_0\le\eta_{\max}\\ 0<r\le 1}} \overline{\mathcal{E}}_r(\eta_0 N) = S(T) + \left(1 + \frac{1}{s_c}\right) aT^{-s}\widehat{r}_N(1 + o(1)), \qquad T = \eta_{\max} N. \tag{B.21}$$

We retain the exact baseline $S(T)$ here: approximating it only by $C_S T^{-s}$ need not resolve the smaller window-dependent correction. Thus, optimal tail selection refines the risk beyond its leading N^{-s} rate.

B.4 The Critical Regime

When $s = s_c$, the signal and noise terms have the same dependence on N for fixed η_0:

$$\overline{\mathcal{E}}_r(\eta_0 N) \sim N^{-s}\left[A_s(r)\eta_0^{-s} + B_s(r)\eta_0^{1-s}\right]. \tag{B.22}$$

For a fixed merge ratio, the optimal LR is

$$\eta_r^* = \min\left\{\eta_{\max}, \frac{sA_s(r)}{(1-s)B_s(r)}\right\}. \tag{B.23}$$

The optimal merge ratio minimizes the bracket in (B.22) after this substitution. Whether the LR constraint is active depends on the signal and noise amplitudes, so s and q alone do not determine the critical-regime optimum.

B.5 Data-Scaling Rates

For any fixed $r \in (0, 1)$, the weaker two-sided power-law assumptions already imply

$$\overline{\mathcal{E}}_r(T) \eqsim T^{-s} + \frac{T^{1/q}}{N}.$$

The unconstrained balance occurs at $T \eqsim N^{q/(sq+1)}$. It is feasible when $s \ge s_c$; otherwise, the constraint $T \le \eta_{\max} N$ is active. Hence,

$$\inf_{0<\eta_0\le\eta_{\max}} \overline{\mathcal{E}}_r(\eta_0 N) \eqsim \begin{cases} N^{-s}, & s < s_c, \\ N^{-\frac{sq}{sq+1}}, & s \ge s_c. \end{cases} \tag{B.24}$$

These rate statements do not select a particular merge ratio. The preceding analysis goes further by minimizing the leading risk constant in the easy regime and the window-dependent excess risk in the hard regime.

C Power Merging: Optimal Weights and EMA

This appendix proves the power-averaging part of Theorem 2 and Proposition 1. We retain the notation and strengthened spectral assumptions of Appendix B. In particular,

$$S(t) \sim C_S t^{-s}, \qquad G(t) := \frac{1}{4}\sum_{j\geq 1} \lambda_j e^{-\lambda_j t} \sim C_N t^{-s_c}, \qquad s_c = 1 - \frac{1}{q},$$

where C_S, C_N are given in (B.2). Throughout, $\sigma = B = 1$. Under constant LR η_0, write $T = \eta_0 N$ and consider $\phi_\kappa(u) = \kappa u^{\kappa-1}$, $\kappa > 0$.

C.1 Risk Representation and Profile Coefficients

Here $A_s(\phi_\kappa) = A_s^{\mathrm{pow}}(\kappa)$ and $B_{s_c}(\phi_\kappa) = B_{s_c}^{\mathrm{pow}}(\kappa)$ connect the family coefficients to the main-text notation.

Exact signal and noise terms. Let U, V be independent random variables with density ϕ_κ on $[0,1]$. The signal transformation and the covariance calculation in (B.5) give

$$\begin{aligned} S_\kappa(T) &= \mathbb{E}\left[S\left(\frac{T(U+V)}{2}\right)\right], \\ V_\kappa(T) &= \mathbb{E}\left[G(T|U-V|) - G(T(U+V))\right], \\ \overline{\mathcal{E}}_\kappa(T) &= S_\kappa(T) + \eta_0 V_\kappa(T). \end{aligned} \tag{C.1}$$

The expectations here are over the averaging locations U, V; the stochastic training noise has already been integrated out.

For fixed κ with $2\kappa > \max\{s, s_c\}$, dominated convergence yields

$$S_\kappa(T) \sim A_s^{\mathrm{pow}}(\kappa) T^{-s}, \qquad V_\kappa(T) \sim B_{s_c}^{\mathrm{pow}}(\kappa) T^{-s_c}, \tag{C.2}$$

where

$$\begin{aligned} A_s^{\mathrm{pow}}(\kappa) &= 2^s C_S\, \mathbb{E}[(U+V)^{-s}], \\ B_{s_c}^{\mathrm{pow}}(\kappa) &= C_N\, \mathbb{E}\left[|U-V|^{-s_c} - (U+V)^{-s_c}\right]. \end{aligned} \tag{C.3}$$

The signal integral is finite precisely when $2\kappa > s$. For the noise coefficient, integrability near $U = V = 0$ requires $2\kappa > s_c$, while integrability along the diagonal requires $s_c < 1$. Both conditions hold in the stated regime.

Consequently,

$$\overline{\mathcal{E}}_\kappa(T) = A_s^{\mathrm{pow}}(\kappa) T^{-s}(1 + o(1)) + B_{s_c}^{\mathrm{pow}}(\kappa) \frac{T^{1/q}}{N}(1 + o(1)). \tag{C.4}$$

Explicit coefficient formulas. Define

$$I_p(\kappa) = \int_0^1 z^{\kappa-1}(1+z)^{-p}\, \mathrm{d}z.$$

Splitting the square into the regions $v \leq u$ and $u \leq v$, and substituting $v = uz$, gives

$$\mathbb{E}[(U+V)^{-p}] = \frac{2\kappa^2}{2\kappa - p} I_p(\kappa), \qquad 2\kappa > p.$$

Similarly, for $0 < p < 1$ and $2\kappa > p$,

$$\mathbb{E}[|U-V|^{-p}] = \frac{2\kappa^2}{2\kappa - p} \mathrm{B}(\kappa, 1-p),$$

where B denotes the beta function. Thus,

$$\begin{aligned} A_s^{\mathrm{pow}}(\kappa) &= 2^s C_S \frac{2\kappa^2}{2\kappa - s} I_s(\kappa), \\ B_{s_c}^{\mathrm{pow}}(\kappa) &= C_N \frac{2\kappa^2}{2\kappa - s_c}\left[\mathrm{B}(\kappa, 1-s_c) - I_{s_c}(\kappa)\right]. \end{aligned} \tag{C.5}$$

These are the coefficients used in (5.6).

Concentration near the current iterate. As $\kappa \to \infty$,

$$A_s^{\mathrm{pow}}(\kappa) = C_S\left(1 + \frac{s}{\kappa} + O(\kappa^{-2})\right), \qquad B_{s_c}^{\mathrm{pow}}(\kappa) \sim b_{\mathrm{pow}}\kappa^{s_c}, \tag{C.6}$$

where

$$b_{\mathrm{pow}} = C_N\Gamma(1 - s_c). \tag{C.7}$$

The first expansion follows from concentration of U, V near one. For the second, use $\mathrm{B}(\kappa, 1-s_c) \sim \Gamma(1-s_c)\kappa^{s_c-1}$ and $I_{s_c}(\kappa) \sim 2^{-s_c}/\kappa$. Thus, increasingly concentrated weights reduce the signal penalty but increase the noise coefficient.

C.2 Easy Tasks: Optimal Fixed Exponents

Suppose $s > s_c$. For a fixed $\kappa > s/2$, optimizing the leading risk in (C.4) over intrinsic time gives

$$T_\kappa^* = \left(\frac{sq\, A_s^{\mathrm{pow}}(\kappa)}{B_{s_c}^{\mathrm{pow}}(\kappa)} N\right)^{\frac{q}{sq+1}}. \tag{C.8}$$

Since $s > s_c$, the associated LR tends to zero, so the upper LR constraint is asymptotically inactive. Substituting (C.8) yields

$$\inf_{0<\eta_0\le\eta_{\max}} \overline{\mathcal{E}}_\kappa(\eta_0 N) \sim D_{s,q}\left[A_s^{\mathrm{pow}}(\kappa)\big(B_{s_c}^{\mathrm{pow}}(\kappa)\big)^{sq}\right]^{\frac{1}{sq+1}} N^{-\frac{sq}{sq+1}}, \tag{C.9}$$

where $D_{s,q} = (1+sq)(sq)^{-sq/(sq+1)}$. This proves the objective in (5.6).

Existence of an interior optimum. As $\kappa \downarrow s/2$, the signal coefficient diverges, while the noise coefficient remains finite and positive because $s > s_c$. As $\kappa \to \infty$, (C.6) implies that the objective diverges through the noise coefficient. Continuity therefore guarantees at least one finite interior minimizer. Every such minimizer satisfies

$$\partial_\kappa \log A_s^{\mathrm{pow}}(\kappa^*) + sq\,\partial_\kappa \log B_{s_c}^{\mathrm{pow}}(\kappa^*) = 0.$$

The argument does not require uniqueness.

Budget-dependent choices. Allowing κ to vary with N does not improve the leading optimum above. A fixed spectral direction gives a signal lower bound of order $T^{-2\kappa}$ when κ is small. Comparison with a fixed admissible exponent therefore bounds a risk-minimizing sequence away from $\kappa = 0$. The weight on the latter half of training is then uniformly positive, so modes with $\lambda_j T$ of order one imply $V_\kappa(T) \gtrsim T^{-s_c}$. Together with $S_\kappa(T) \ge S(T)$, this forces $T \asymp N^{q/(sq+1)}$.

Fatou's lemma applied to the signal representation excludes limit points $\kappa \le s/2$. The growing-κ estimates proved below exclude $\kappa \to \infty$, since the noise coefficient would diverge at this intrinsic-time scale. The fixed-profile asymptotics are uniform on the remaining compact subsets of $(s/2, \infty)$, which establishes the leading optimality of (C.9).

C.3 Hard Tasks: Optimal Concentration of the Weights

Suppose $s < s_c$. As in tail averaging, the optimal weighting refines the risk above its leading N^{-s} decay. We first establish expansions that remain valid when κ grows with the training budget.

Signal penalty for growing κ. Let U, V have density ϕ_κ and put $M = (U+V)/2$. Then

$$\mathbb{E}[1-M] = \frac{1}{\kappa+1}, \qquad \mathbb{E}[(1-M)^2] = O(\kappa^{-2}).$$

Using the derivative asymptotics in (B.3), a Taylor expansion near $M = 1$ gives, jointly as $T \to \infty$ and $\kappa \to \infty$,

$$S_\kappa(T) - S(T) \sim a_{\mathrm{pow}}\frac{T^{-s}}{\kappa}, \qquad a_{\mathrm{pow}} = sC_S. \tag{C.10}$$

The contribution from $M < 1/2$ is negligible: the density of U, V suppresses this region exponentially in κ, and its negative moments are uniformly controlled for sufficiently large κ. Thus, this expansion concerns the signal difference directly and does not require substituting a growing κ into a fixed-profile asymptotic formula.

Noise over the effective window. Write $L = T/\kappa$ and suppose $L \to \infty$. For $X = \kappa(1-U)$, the density is

$$p_\kappa(x) = \left(1 - \frac{x}{\kappa}\right)^{\kappa-1} \mathbf{1}_{[0,\kappa]}(x).$$

It converges to e^{-x} and, for $\kappa \geq 2$, is bounded by $e^{-x/2}$. For independent X, Y with this density,

$$T|U - V| = L|X - Y|.$$

Dominated convergence therefore gives

$$L^{s_c}\mathbb{E}[G(T|U-V|)] \longrightarrow C_N \int_0^\infty \int_0^\infty e^{-x-y}|x-y|^{-s_c}\,\mathrm{d}x\,\mathrm{d}y = C_N\Gamma(1-s_c).$$

The diagonal singularity is integrable because $s_c < 1$. Meanwhile, $\mathbb{E}[G(T(U+V))] = O(T^{-s_c})$, which is negligible relative to L^{-s_c}. Hence,

$$V_\kappa(T) \sim b_{\mathrm{pow}} \left(\frac{T}{\kappa}\right)^{-s_c}. \tag{C.11}$$

Combining the signal and noise estimates gives

$$\overline{\mathcal{E}}_\kappa(T) - S(T) = a_{\mathrm{pow}} T^{-s}\kappa^{-1}(1+o(1)) + b_{\mathrm{pow}}\eta_0 T^{-s_c}\kappa^{s_c}(1+o(1)), \tag{C.12}$$

whenever $T \to \infty$, $\kappa \to \infty$, and $T/\kappa \to \infty$.

Localization of the optimum. Choosing $\eta_0 = \eta_{\max}$ and $\kappa = N^{(s_c-s)/(1+s_c)}$ gives risk $S(\eta_{\max}N)(1+o(1))$. Since every averaged model satisfies $\overline{\mathcal{E}}_\kappa(T) \geq S(T)$, a risk-minimizing sequence must have $T/(\eta_{\max}N) \to 1$.

It must also satisfy $\kappa \to \infty$. Indeed, convexity of S gives

$$S_\kappa(T) \geq S\left(T\frac{\kappa}{\kappa+1}\right),$$

so bounded κ incurs a nonvanishing relative signal penalty. Finally, $T/\kappa \to \infty$ is necessary for the noise risk to vanish. If T/κ remained bounded, the covariance representation above would retain a positive noise contribution.

The same derivative argument used in Appendix B.3 shows that the LR constraint is eventually active. In this localized regime, $TS'_\kappa(T) \sim -sS(T)$ and $TV'_\kappa(T) \sim -s_cV_\kappa(T)$, while $\eta_0 V_\kappa(T) = o(S(T))$. Thus, increasing η_0 decreases the risk until $\eta_0 = \eta_{\max}$.

Optimal exponent and risk refinement. Set $T = \eta_{\max}N$. The leading excess in (C.12) is minimized at

$$\widehat{\kappa}_N = \left(\frac{a_{\mathrm{pow}}}{s_c b_{\mathrm{pow}}\eta_{\max}} T^{s_c-s}\right)^{\frac{1}{1+s_c}}. \tag{C.13}$$

With $\kappa = y\widehat{\kappa}_N$, the leading excess divided by $a_{\mathrm{pow}}T^{-s}/\widehat{\kappa}_N$ becomes

$$y^{-1} + \frac{1}{s_c}y^{s_c}.$$

This has a unique minimum at $y = 1$ and diverges at both endpoints. The uniform expansions therefore imply

$$\kappa_N^* \sim \widehat{\kappa}_N \asymp N^{\frac{s_c-s}{1+s_c}}, \qquad \frac{T}{\kappa_N^*} \asymp N^{\frac{1+s}{1+s_c}}, \tag{C.14}$$

proving (5.5).

More precisely,

$$\inf_{\substack{0<\eta_0\leq\eta_{\max}\\ \kappa>0}} \overline{\mathcal{E}}_\kappa(\eta_0 N) = S(T) + \left(1+\frac{1}{s_c}\right)\frac{a_{\mathrm{pow}}T^{-s}}{\widehat{\kappa}_N}(1+o(1)), \qquad T = \eta_{\max}N. \tag{C.15}$$

As in the tail analysis, keeping the exact baseline $S(T)$ makes the smaller weight-dependent correction explicit.

C.4 The Critical Regime and Data-Scaling Rates

When $s = s_c$, any fixed $\kappa > s/2$ satisfies

$$\overline{\mathcal{E}}_\kappa(\eta_0 N) \sim N^{-s}\left[A_s^{\mathrm{pow}}(\kappa)\eta_0^{-s} + B_s^{\mathrm{pow}}(\kappa)\eta_0^{1-s}\right].$$

For fixed κ, the optimal LR is

$$\eta_\kappa^* = \min\left\{\eta_{\max}, \frac{sA_s^{\mathrm{pow}}(\kappa)}{(1-s)B_s^{\mathrm{pow}}(\kappa)}\right\}. \tag{C.16}$$

The optimal exponent minimizes the bracket after this substitution. Its value can therefore depend on the signal and noise amplitudes and the LR bound.

More generally, for any fixed $2\kappa > \max\{s, s_c\}$, the two-sided scaling assumptions imply $\overline{\mathcal{E}}_\kappa(T) \eqsim T^{-s} + T^{1/q}/N$. The optimization in Appendix B.5 consequently gives

$$\inf_{0<\eta_0\le\eta_{\max}} \overline{\mathcal{E}}_\kappa(\eta_0 N) \eqsim \begin{cases} N^{-s}, & s < s_c, \\ N^{-\frac{sq}{sq+1}}, & s \ge s_c. \end{cases} \tag{C.17}$$

Thus, a suitable fixed exponent already attains the optimal data-scaling rate. Optimizing κ further selects the leading risk constant or the excess above gradient-flow risk, as characterized above.

C.5 Power Merging as Dynamic EMA

Exact continuous-time identity. For fixed $\kappa > 0$,

$$\bar{\boldsymbol{\vartheta}}_\kappa(t) = \kappa t^{-\kappa}\int_0^t u^{\kappa-1}\boldsymbol{\vartheta}(u)\,\mathrm{d}u.$$

Differentiating for $t > 0$ gives

$$\frac{\mathrm{d}\bar{\boldsymbol{\vartheta}}_\kappa(t)}{\mathrm{d}t} = \frac{\kappa}{t}\left(\boldsymbol{\vartheta}(t) - \bar{\boldsymbol{\vartheta}}_\kappa(t)\right). \tag{C.18}$$

Conversely, multiplying this ODE by t^κ and integrating recovers the power average, with the initial condition inherited from $\boldsymbol{\vartheta}(0)$. The identity holds coordinatewise and in prediction norm.

Returning to training-step time z gives

$$\frac{\mathrm{d}\bar{\boldsymbol{\theta}}_\kappa(z)}{\mathrm{d}z} = \frac{\kappa\eta(z)}{t(z)}\left(\boldsymbol{\theta}(z) - \bar{\boldsymbol{\theta}}_\kappa(z)\right). \tag{C.19}$$

Thus, maintaining a fixed intrinsic-time profile requires an EMA update rate that depends on the LR schedule.

A discrete implementation. Let $t_k = \sum_{i=0}^{k-1}\eta_i$. Approximating the trajectory by its right endpoint on each intrinsic-time interval gives

$$\bar{\boldsymbol{\theta}}_{\kappa,k} = \sum_{i=1}^{k}\frac{t_i^\kappa - t_{i-1}^\kappa}{t_k^\kappa}\boldsymbol{\theta}_i. \tag{C.20}$$

This discrete average admits the exact recursion

$$\bar{\boldsymbol{\theta}}_{\kappa,k} = \rho_k\bar{\boldsymbol{\theta}}_{\kappa,k-1} + (1-\rho_k)\boldsymbol{\theta}_k, \qquad \rho_k = \left(\frac{t_{k-1}}{t_k}\right)^\kappa. \tag{C.21}$$

For fixed κ and $\eta_{k-1}/t_k \to 0$, $1-\rho_k \sim \kappa\eta_{k-1}/t_k$. Under constant LR, $\rho_k = ((k-1)/k)^\kappa$ and $1-\rho_k \sim \kappa/k$. No past checkpoints need to be stored.

Connection to anytime pretraining. For $f > 0$, the EMA retention coefficient $2^{-f/k}$ used by Meterez et al. (2026) has update coefficient

$$1 - 2^{-f/k} \sim \frac{f \log 2}{k}.$$

Write $b = f \log 2$. The induced weights have asymptotic density bu^{b-1} in relative training-step time. Indeed, away from the initial endpoint,

$$(1 - e^{-b/i}) \prod_{\ell=i+1}^{k} e^{-b/\ell} \sim \frac{b}{k} \left(\frac{i}{k}\right)^{b-1}.$$

If $\eta(z) \sim cz^{-p}$ with $0 \le p < 1$, then $t(z) \sim cz^{1-p}/(1-p)$. Transforming from relative training-step time to relative intrinsic time therefore gives

$$\kappa = \frac{b}{1-p}. \tag{C.22}$$

This yields $\kappa = f \log 2$ for constant LR and $\kappa = 2f \log 2$ for asymptotic inverse-square-root decay. The correspondence concerns the asymptotic profile; warmup and finite-time offsets can change the early weights.

All identities above use a fixed κ within a training run. Selecting κ_N separately for each budget is compatible with these identities, whereas replacing it online by a time-varying exponent generally produces a different averaging rule.

C.6 Proof of the Fixed-EMA Noise Floor

Proposition 1 (Noise floor of fixed EMA)**.** *Under* (3.2) *with fixed* $\eta_0 > 0$ *and deterministic initialization, the noise contribution* $\mathcal{N}_\gamma(t)$ *of fixed EMA satisfies*

$$\lim_{t\to\infty} \mathcal{N}_\gamma(t) = \frac{\eta_0}{4} \sum_{j\ge 1} \frac{\gamma\lambda_j}{\gamma + \lambda_j} \in (0, \infty). \tag{C.23}$$

Response to a noise injection. Consider fixed EMA with rate $\gamma > 0$, initialized at the deterministic initial model:

$$\frac{\mathrm{d}\bar{\boldsymbol{\vartheta}}_\gamma(t)}{\mathrm{d}t} = \gamma \left(\boldsymbol{\vartheta}(t) - \bar{\boldsymbol{\vartheta}}_\gamma(t)\right).$$

For a spectral direction with eigenvalue λ, averaging the stochastic convolution gives the response to noise of age u,

$$R_\gamma(\lambda, u) = \gamma \int_0^u e^{-\gamma(u-v)} e^{-\lambda v}\, \mathrm{d}v = \begin{cases} \dfrac{\gamma}{\gamma - \lambda}\left(e^{-\lambda u} - e^{-\gamma u}\right), & \lambda \ne \gamma, \\ \gamma u e^{-\gamma u}, & \lambda = \gamma. \end{cases} \tag{C.24}$$

Consequently, the stochastic component in mode j is

$$\bar{\xi}_{\gamma,j}(t) = \sqrt{\eta_0 \lambda_j} \int_0^t R_\gamma(\lambda_j, t-\tau)\, \mathrm{d}W_j(\tau).$$

Integrated noise risk. Itô isometry gives

$$\mathcal{N}_\gamma(t) = \frac{\eta_0}{2} \sum_{j\ge 1} \lambda_j^2 \int_0^t R_\gamma(\lambda_j, u)^2\, \mathrm{d}u. \tag{C.25}$$

Direct integration yields

$$\int_0^\infty R_\gamma(\lambda, u)^2\, \mathrm{d}u = \frac{\gamma}{2\lambda(\gamma + \lambda)}, \tag{C.26}$$

including $\lambda = \gamma$ by continuity. Since the integrands are nonnegative, monotone convergence implies

$$\mathcal{N}_\gamma(t) \uparrow \frac{\eta_0}{4} \sum_{j\ge 1} \frac{\gamma\lambda_j}{\gamma + \lambda_j}. \tag{C.27}$$

The limit is strictly positive and finite because

$$0 < \frac{\gamma\lambda_j}{\gamma + \lambda_j} \le \lambda_j, \qquad \sum_{j\ge 1} \lambda_j < \infty.$$

This proves Proposition 1.

In particular, for all sufficiently large t, $\mathcal{N}_\gamma(t)$ is at least half its positive limit. The limit is smaller than the last-iterate noise floor $\eta_0 \operatorname{Tr}(\mathbf{H})/4$, but does not vanish. This conclusion assumes a fixed nonzero LR; it does not assert a noise floor under a decaying LR schedule.

D Checkpoint Spacing and Discrete Averaging

This appendix proves Theorem 3 and the corresponding signal approximation. We use (3.2) with constant LR, deterministic initialization, and $\sigma = B = 1$. Assume $\lambda_j \eqsim j^{-q}$, $q > 1$, and write $s_c = 1 - 1/q$.

Let $a = t - L \ge 0$, $h = L/m$, and $u_i = a + ih$ for $i = 1, \dots, m$. We compare

$$\bar{\boldsymbol{\vartheta}}_m(t) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\vartheta}(u_i) \quad \text{and} \quad \bar{\boldsymbol{\vartheta}}_{\mathrm{cont}}(t) = \frac{1}{L}\int_a^t \boldsymbol{\vartheta}(u)\,\mathrm{d}u.$$

All comparison constants below are independent of a, L, and m, unless otherwise stated.

D.1 Covariance Representation and Discretization Error

Exact noise covariance. Let $\xi_j(u)$ denote the stochastic component of $\vartheta_j(u)$. Under constant LR,

$$\mathbb{E}[\xi_j(u)\xi_j(v)] = \frac{\eta_0}{2}\left(e^{-\lambda_j|u-v|} - e^{-\lambda_j(u+v)}\right).$$

Define

$$G(w) = \frac{1}{4}\sum_{j\ge 1} \lambda_j e^{-\lambda_j w}. \tag{D.1}$$

This function is decreasing and convex, with $G(0) = \operatorname{Tr}(\mathbf{H})/4 < \infty$. The spectral assumption implies $G(w) \eqsim (1+w)^{-s_c}$. The discrete and continuous noise contributions are therefore

$$\begin{aligned} \mathcal{N}_m(t) &= \frac{\eta_0}{m^2}\sum_{i,j=1}^{m} \left[G(|u_i - u_j|) - G(u_i + u_j)\right], \\ \mathcal{N}_{\mathrm{cont}}(t) &= \frac{\eta_0}{L^2}\int_a^t\int_a^t \left[G(|u-v|) - G(u+v)\right]\,\mathrm{d}u\,\mathrm{d}v. \end{aligned} \tag{D.2}$$

Separate the stationary covariance and the initialization correction by writing

$$\frac{\mathcal{N}_m(t)}{\eta_0} = A_m - B_m, \qquad \frac{\mathcal{N}_{\mathrm{cont}}(t)}{\eta_0} = A_{\mathrm{cont}} - B_{\mathrm{cont}}.$$

Explicitly,

$$\begin{aligned} A_m &= \frac{G(0)}{m} + \frac{2}{m^2}\sum_{\ell=1}^{m-1}(m-\ell)G(\ell h), \\ A_{\mathrm{cont}} &= \frac{2}{L^2}\int_0^L (L-v)G(v)\,\mathrm{d}v, \\ B_m &= \frac{1}{m^2}\sum_{i,j=1}^{m} G(2a + ih + jh), \\ B_{\mathrm{cont}} &= \frac{1}{L^2}\int_0^L\int_0^L G(2a + x + y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned} \tag{D.3}$$

Stationary covariance error. Set $F(v) = (L - v)G(v)$ on $[0, L]$. It is decreasing and convex because

$$F'(v) = -G(v) + (L - v)G'(v) \le 0, \qquad F''(v) = -2G'(v) + (L - v)G''(v) \ge 0.$$

Moreover, A_m is exactly $2/L^2$ times the composite trapezoidal approximation to $\int_0^L F(v)\,\mathrm{d}v$. Convexity implies that this approximation overestimates the integral.

For a decreasing function, the trapezoidal sum is the mean of its left- and right-endpoint sums. Its excess over the integral is therefore at most $h(F(0) - F(L))/2$. Since $F(0) = LG(0)$ and $F(L) = 0$, we obtain

$$0 \le A_m - A_{\mathrm{cont}} \le \frac{G(0)}{m}. \tag{D.4}$$

Initialization correction. The function $G(2a + x + y)$ decreases in both coordinates. Thus, its right-endpoint sum underestimates the integral: $B_m \le B_{\mathrm{cont}}$. Bounding the error successively in each coordinate gives

$$0 \le B_{\mathrm{cont}} - B_m \le \frac{2G(2a)}{m} \le \frac{2G(0)}{m}. \tag{D.5}$$

Combining (D.4) and (D.5) yields the explicit bound

$$0 \le \mathcal{N}_m(t) - \mathcal{N}_{\mathrm{cont}}(t) \le \frac{3\eta_0 G(0)}{m} = \frac{3\eta_0}{4m}\,\mathrm{Tr}(\mathbf{H}). \tag{D.6}$$

This proves the first assertion of (5.9).

D.2 Noise Scaling and the Required Checkpoint Count

Continuous averaging. Dropping the nonnegative initialization correction and using $G(v) \lesssim (1+v)^{-s_c}$ gives, for $L \ge 2$,

$$\mathcal{N}_{\mathrm{cont}}(t) \le \frac{2\eta_0}{L}\int_0^L G(v)\,\mathrm{d}v \lesssim \eta_0 L^{-s_c}.$$

For the lower bound, noise injected before the window contributes a nonnegative risk. The contribution from noise injected within the window is the same as for averaging a trajectory initialized deterministically at the window's beginning. Consequently,

$$\mathcal{N}_{\mathrm{cont}}(t) \ge \frac{\eta_0}{2L^2}\sum_{j\ge 1}\int_0^L \left(1 - e^{-\lambda_j v}\right)^2\,\mathrm{d}v. \tag{D.7}$$

Restricting to modes with $\lambda_j L \le 1$ and using $1 - e^{-x} \gtrsim x$ for $x \in [0, 1]$ gives

$$\mathcal{N}_{\mathrm{cont}}(t) \gtrsim \eta_0 L \sum_{\lambda_j \le L^{-1}} \lambda_j^2 \eqsim \eta_0 L^{-s_c}.$$

Hence, uniformly over $a \ge 0$,

$$\mathcal{N}_{\mathrm{cont}}(t) \eqsim \eta_0 L^{-s_c}, \qquad L \ge 2. \tag{D.8}$$

The finite-count contribution. Each summand in (D.2) is nonnegative. Keeping only the diagonal terms gives

$$\mathcal{N}_m(t) \ge \frac{\eta_0}{m^2}\sum_{i=1}^{m}\left[G(0) - G(2u_i)\right].$$

At least half of the sampled endpoints satisfy $u_i \ge L/2 \ge 1$. Since G is strictly decreasing,

$$\mathcal{N}_m(t) \ge \frac{\eta_0}{2m}\left[G(0) - G(2)\right] \gtrsim \frac{\eta_0}{m}. \tag{D.9}$$

Together with $\mathcal{N}_m(t) \ge \mathcal{N}_{\mathrm{cont}}(t)$, this gives the lower bound

$$\mathcal{N}_m(t) \gtrsim \eta_0\left(L^{-s_c} + \frac{1}{m}\right).$$

The matching upper bound follows from (D.6) and (D.8). Therefore,

$$\mathcal{N}_m(t) \eqsim \eta_0\left(L^{-s_c} + \frac{1}{m}\right). \tag{D.10}$$

Preserving the continuous-averaging benefit. Equation (D.10) shows that $\mathcal{N}_m(t) \eqsim \eta_0 L^{-s_c}$ holds precisely when $m \gtrsim L^{s_c}$. Since $h = L/m$, this is equivalent to $h \lesssim L^{1/q}$.

The sharper error bound gives

$$0 \le \frac{\mathcal{N}_m(t)}{\mathcal{N}_{\mathrm{cont}}(t)} - 1 \lesssim \frac{L^{s_c}}{m}. \tag{D.11}$$

Thus, $m/L^{s_c} \to \infty$ ensures relative convergence to continuous averaging, completing the proof of Theorem 3.

A fixed count leaves a noise floor. Fix m and $\eta_0 > 0$, and let $L \to \infty$. Then $h = L/m \to \infty$. In (D.3), all off-diagonal terms in A_m vanish, whereas its diagonal contribution remains $G(0)/m$. Also, $B_m \to 0$ because $u_i + u_j \ge 2h \to \infty$. Hence, uniformly over choices of $a \ge 0$,

$$\mathcal{N}_m(t) \longrightarrow \frac{\eta_0 G(0)}{m} = \frac{\eta_0}{4m} \operatorname{Tr}(\mathbf{H}) > 0. \tag{D.12}$$

A growing window alone therefore cannot produce vanishing noise risk with a fixed checkpoint count and fixed nonzero LR.

D.3 Signal Approximation on a Fractional Tail

Right-endpoint sampling reduces signal error. The exact signal contributions are

$$\begin{aligned} S_m(t) &= \frac{1}{m^2} \sum_{i,j=1}^{m} S\left(\frac{u_i + u_j}{2}\right), \\ S_{\mathrm{cont}}(t) &= \frac{1}{L^2} \int_a^t \int_a^t S\left(\frac{u+v}{2}\right) \mathrm{d}u \, \mathrm{d}v. \end{aligned} \tag{D.13}$$

Since S is decreasing, the right-endpoint sum underestimates the integral. All sampled times are at most t, so

$$S(t) \le S_m(t) \le S_{\mathrm{cont}}(t). \tag{D.14}$$

Thus, the discretization effects on signal and noise have opposite signs: later representatives reduce signal error, while finite sampling increases noise.

Relative convergence for a fixed tail. Now fix a merge ratio $r \in (0,1)$ and set $a = (1-r)t$, $L = rt$. The source and capacity assumptions imply $S(t) \eqsim t^{-s}$ and $|S'(v)| \lesssim_r t^{-s-1}$ uniformly for $v \in [(1-r)t, t]$ and sufficiently large t. Within each integration cell, replacing (u, v) by its right endpoints changes $(u+v)/2$ by at most h. Therefore,

$$0 \le S_{\mathrm{cont}}(t) - S_m(t) \lesssim_r h\, t^{-s-1} \lesssim_r \frac{t^{-s}}{m}. \tag{D.15}$$

Because $S(t) \le S_{\mathrm{cont}}(t) \le S((1-r)t)$, both signal contributions have order t^{-s}, and

$$0 \le 1 - \frac{S_m(t)}{S_{\mathrm{cont}}(t)} \lesssim_r \frac{1}{m}. \tag{D.16}$$

Let $\overline{\mathcal{E}}_m(t) = S_m(t) + \mathcal{N}_m(t)$ and $\overline{\mathcal{E}}_{\mathrm{cont}}(t) = S_{\mathrm{cont}}(t) + \mathcal{N}_{\mathrm{cont}}(t)$. Combining (D.11) and (D.16) gives

$$\left| \frac{\overline{\mathcal{E}}_m(t)}{\overline{\mathcal{E}}_{\mathrm{cont}}(t)} - 1 \right| \lesssim_r \frac{1}{m} + \frac{L^{s_c}}{m}. \tag{D.17}$$

Consequently, $m \gg L^{s_c}$ preserves the leading risk of continuous averaging over a fixed fractional tail. These relative bounds remain valid when the constant LR is retuned across training budgets, since the noise terms scale linearly with η_0.

D.4 Translation to Training-Step Spacing

Under constant LR, checkpoints separated by Δk updates have intrinsic-time spacing $h = \eta_0 \Delta k$. The count condition can therefore be written as

$$\eta_0 \Delta k \lesssim L^{1/q}.$$

For a general LR schedule, checkpoints at update indices k_{i-1} and k_i instead have spacing

$$h_i = t(k_i) - t(k_{i-1}) = \int_{k_{i-1}}^{k_i} \eta(z)\,\mathrm{d}z.$$

Equal training-step spacing need not give an equal intrinsic-time grid. Moreover, the noise intensity is then time dependent, so extending the theorem requires retaining φ in the covariance calculation.

E Experimental Protocol and Reporting Details

This section records the training, averaging, and fitting protocols for Section 6. The controlled kernel experiments and three LLM experiments are documented separately. The LLM studies distinguish cross-schedule prediction, model-inferred decomposition, and the comparison across training budgets.

E.1 Controlled Kernel Regression Protocol

Data generation and evaluation. The experiments in Figure 2 use fresh independent samples $x_k \sim \mathcal{N}(0, \mathbf{H})$ and $\epsilon_k \sim \mathcal{N}(0,1)$ at each SGD update, with $\mathbf{H} = \mathrm{diag}(j^{-q})_{j=1}^d$ and $y_k = x_k^\top \boldsymbol{\theta}^* + \epsilon_k$. For nonzero targets, we set $\lambda_j |\theta_j^*|^2 = j^{-1-sq} / \sum_{\ell=1}^d \ell^{-1-sq}$ and choose positive coordinates of $\boldsymbol{\theta}^*$. All runs use zero initialization, batch size one, and a constant LR within a trajectory. These are sample-based SGD updates, including sample-dependent gradient noise. For each averaged parameter vector, we evaluate the exact finite-dimensional population excess risk $\frac{1}{2}\sum_{j=1}^d \lambda_j(\bar{\theta}_j - \theta_j^*)^2$ and then average it over 512 independent trajectories. Methods share training samples within each comparison. The data budget N equals the number of updates. All merge rules use post-update iterates.

Data-scaling comparison, panel (a). We use $d = 2048$, $q = 2$, $s = 3$, and budgets $N = 2^{10}, 2^{11}, \ldots, 2^{18}$. For each method and budget, we select the lowest mean risk among 33 geometrically spaced LRs from 2×10^{-6} to 0.1. The five estimators are the last iterate, fixed EMA with $\beta = 0.99$, full uniform averaging, tail averaging with $r = 1/2$, and power averaging with $\kappa = 4$. Both EMA states start at zero; fixed EMA uses $\bar{\theta}_k = \beta\bar{\theta}_{k-1} + (1-\beta)\theta_k$ without bias correction, and power averaging uses (5.8). Under constant LR its retention is $((k-1)/k)^\kappa$. Only the LR is selected in this panel; the averaging parameters remain fixed across budgets.

Panel (a) displays fitted power laws, including markers placed on those fitted curves. We fit $\log \mathcal{E} = \log A - \alpha \log N$ by unweighted least squares, with zero excess-risk floor. The last iterate and uniform averaging use all nine budgets; fixed EMA, dynamic EMA, and tail averaging use the last five, $N \geq 2^{14}$. These fitted trends are displayed over the entire budget range, so the early portions of the latter three curves extrapolate their fitting intervals. The fitted exponents are 0.7381, 0.7474, 0.7956, 0.8497, and 0.8396 for the last iterate, fixed EMA, uniform, dynamic EMA, and tail averaging, respectively. Because LR is retuned across budgets, this panel does not test the fixed-LR noise floor of fixed EMA.

Joint LR/profile selection, panels (b,c). We use $d = 4096$, $q = 3$, and $s \in \{0.2, 1\}$ with the same nine budgets. For each budget and family, mean risk is minimized jointly over eight geometric LRs in $[0.0125, 0.2]$ and 61 geometric profile values: $r \in [1/128, 1]$ for tail averaging or $\kappa \in [0.5, 128]$ for power averaging. The retained tail length is rounded to the nearest integer number of updates; the plotted fraction is this actual length divided by N. Each power exponent is fixed within a run. The searches use 512 paired trajectories for each LR and scenario, and reported optima are minima over the tested grids. Shaded bands are 95% paired-bootstrap percentile intervals from 400 resamples, repeating joint LR/profile selection in each resample.

For the hard task, $s_c = 2/3$ and $(s_c - s)/(1 + s_c) = 0.28$. The selected LR reaches the allowed maximum 0.2 at every budget. The dashed lines have the prescribed theoretical slopes -0.28 and $+0.28$; their vertical offsets are fitted to the last five budgets. Thus they compare scaling trends rather than absolute asymptotic constants. For the easy task, the selected tail ratio reaches $r = 1$ at larger budgets, while selected power exponents remain of order one.

Checkpoint growth, panel (d). To measure the noise contribution directly, we set $\boldsymbol{\theta}^* = 0$, $d = 4096$, $q = 3$, and $\eta_0 = 0.09057236642639067$. For $W = 1260 \cdot 2^i$, $i = 0, \ldots, 7$, we train for $N = 2W$ updates and average the last W updates. The intrinsic window length is $L = \eta_0 W$. For each positive divisor m of W, we collect exactly m equally spaced right endpoints within this window. The baseline averages all W post-update iterates. We combine four independent groups of 128 trajectories and report the ratio of mean sparse-average risk to mean every-step risk.

The three growth rules target $m = 60(L/L_{\min})^a$ for $a \in \{0.5, 2/3, 0.8\}$. For each window we choose the tested divisor closest to this target in logarithmic distance. The common starting count 60 was calibrated using the smallest-window results to lie within 10% of the baseline; it was not fixed before inspecting those results. Bands are 95% paired-bootstrap percentile intervals for the ratio of mean risks, using 600 resamples. The reference is discrete every-step averaging, not the continuous-SDE average. At the largest window, the ratios are approximately 1.58, 1.20, and 1.08, respectively. The plot tests finite-window behavior consistent with the checkpoint-count prediction; it does not establish relative convergence or the exact continuous-time error bound.

E.2 LLM Training Configurations

We report the prediction, decomposition, and budget experiments separately. All three studies use a 124M LLaMA-style dense model, sequence length 1024, and a global batch of 512 sequences (524,288 tokens per optimizer update). The prediction study uses FineWeb-Edu. The decomposition and budget studies use a LLaMA-style model with 12 layers, 10 attention heads, 10 KV heads, hidden width 640, and feed-forward width 1792, with untied input/output embeddings, zero dropout, and sequence length 1024.

Table 1: Configurations of the three LLM studies. Token counts include the small overshoot from completing the final update.

Field	Prediction	Decomposition	Budget comparison
Model	124M LLaMA-style dense		
Training corpus	FineWeb-Edu		
Optimizer	MuonW		
Sequence length	1024		
Global batch	512 sequences; 524,288 tokens/update		
Training updates	3,815	19,074	38,147
Training tokens	2,000,158,720	10,000,269,312	20,000,014,336

Throughout these experiments, MuonW denotes a hybrid optimizer that applies Muon to hidden-layer weight matrices and AdamW to the remaining parameters, including embeddings, the output head, and normalization parameters.

The prediction LR history, verified against the recorded cumulative LR and per-update LR, has peak 0.005, 76 warmup updates, and 3815 updates in total. WSD uses a 763-update linear decay, with boundary at update 3052. Constant LR remains at its peak after warmup. The reported peak is the MuonW LR.

For decomposition, the Muon and AdamW peak LRs are 0.001 and 0.0001. Warmup lasts 381 updates. WSD linearly decays over the last 20% of training (3815 updates). The constant-LR trajectory retains its peak after warmup. For the budget study, warmup lasts 763 updates and all compared runs use constant LR after warmup. Muon peak LR is selected from $\{0.0005, 0.001, 0.005\}$, with AdamW peak LR always one tenth of the corresponding Muon LR.

Both latter studies use Muon momentum 0.95, Nesterov updates, five Newton–Schulz iterations, and update RMS 0.2; AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$. Gradient clipping is 1. A microbatch

of 16 sequences and global accumulation count 32 give the reported global batch. On eight GPUs this corresponds to four local accumulation steps per update. The saved configurations agree on architecture, dataset path, batch size, and sequence length within each schedule comparison.

The saved precision setting for decomposition and budget runs is BF16. These two studies evaluate every 500 updates and at the final update, using the configured corpus validation file. The implementation evaluates with `model.eval()`, no gradients, and BF16 autocast, averaging 100 equal-size batch cross-entropies on deterministic evaluation batches while preserving the training RNG state.

E.3 Averaging Rule and Clock Convention

The clock must be specified separately for each study. The prediction model uses intrinsic time $t_k = \sum_{i=0}^{k-1} \eta_i$, including warmup, with full-history uniform and power averaging. Its finite-update weights on post-update states θ_j, $j = 1, \dots, k$, are

$$w_{k,j}^{\mathrm{unif}} = \frac{t_j - t_{j-1}}{t_k}, \qquad w_{k,j}^{\mathrm{power}} = \frac{t_j^{\kappa} - t_{j-1}^{\kappa}}{t_k^{\kappa}}.$$

There is no initial-state atom in these fitted weights. Power averaging has retention $(t_{k-1}/t_k)^{\kappa}$. The prediction experiment uses $\kappa = 2$.

The decomposition and budget runs instead record *physical-time* averaging. Their dynamic EMA retentions are $((k-1)/k)^{\kappa}$, with $\kappa = 25$ and $\kappa = 50$, respectively. Physical and intrinsic clocks are not interchangeable during warmup, even for a subsequently constant LR. The implementation updates averages after each optimizer update and initializes a nonempty average from its first contributing post-update state. Its fixed EMA normalizes accumulated mass: $Z_k = \beta Z_{k-1} + (1-\beta)$ and incoming normalized weight $(1-\beta)/Z_k$, with $Z_0 = 0$. All averages start with the first contributing post-update state, include warmup, and use the training-origin clock without a later reset. This startup convention differs from the uncorrected, zero-initialized fixed EMA used in the controlled SGD experiment.

Online averaging is distinct from checkpoint export: both later LLM studies configure a checkpoint interval of 1000 updates; the budget runs additionally enable raw-checkpoint export every 10 updates, while that additional export is disabled for decomposition. These intervals do not specify the online merge update frequency. The prediction averages are likewise maintained online; their weights do not depend on the checkpoint-export interval.

E.4 Cross-Schedule Calibration and Prediction

For each profile separately, WSD calibration uses updates $600, 700, \dots, 3800, 3815$, giving 34 observations. Uniform and dynamic EMA each have their own fitted floor, amplitudes, and exponents; they do not share a fitted base kernel. Cosine and constant losses are excluded from this fitting and cutoff selection. Parameters are frozen before target evaluation. This is retrospective source-only calibration: target curves had been explored earlier in the research process, so it is not a prospectively blind test.

We use the phenomenological spectral proxy

$$L_k = L_\infty + A G_s(k) + C H_\nu(k).$$

Writing $w_{k,j}$ for the finite-update merge weights, its features are

$$G_s(k) = \frac{1}{\Gamma(s)} \int_0^\infty \lambda^{s-1} e^{-\delta_S \lambda} \left(\sum_{j=1}^{k} w_{k,j} e^{-\lambda t_j} \right)^2 \mathrm{d}\lambda,$$

$$H_\nu(k) = \frac{1}{\Gamma(\nu)} \int_0^\infty \lambda^{\nu-1} e^{-\delta_K \lambda} \sum_{i=1}^{k} \eta_{i-1}^2 \left(\sum_{j=i}^{k} w_{k,j} e^{-\lambda (t_j - t_i)} \right)^2 \mathrm{d}\lambda.$$

The implementation names the noise exponent `gamma`; it corresponds to ν here. For a last iterate and an untruncated integral, these conventions give $S(t) = A(\delta_S + 2t)^{-s}$ and $K(u) = C(\delta_K + 2u)^{-\nu}$.

Thus the shifts in the main-text convention are $c_S = \delta_S/2$ and $c_K = \delta_K/2$, with corresponding amplitudes $A2^{-s}$ and $C2^{-\nu}$.

We fix $\delta_S = 0$ and choose δ_K from $\{0.0005, 0.005, 0.05, 0.5\}$ by each profile's WSD RMSE; both displayed profiles select 0.5, i.e., $c_K = 0.25$. The fitted parameters satisfy $s \in [0.05, 2]$, $\nu \in [0.1, 3]$, $A, C \geq 0$, and $0 \leq L_\infty \leq \min L_{\mathrm{WSD}}$. These empirical bounds and the zero signal shift are not the strict assumptions $c_S > 0$, $\nu \in (1, 2)$ of the kernel-regression theorems. No capacity exponent q is imposed on LLM fitting.

The objective is equally weighted squared error in loss space. A 17×17 exponent grid supplies nine starts for bounded Nelder–Mead refinement; amplitudes and floor are solved by constrained linear least squares for each exponent pair. The refinement limit is 800 iterations, with parameter tolerance 10^{-8} and objective tolerance 10^{-12}. Spectral integration uses a trapezoidal rule in $\log \lambda$ on 512 nodes spanning $[10^{-7}, 2 \times 10^5]$, checked against 1024 nodes with fixed parameters. The spectral integrals are therefore numerically truncated. The discrete mode propagator is $e^{-\lambda \eta_{i-1}}$ and innovation variance is η_{i-1}^2; checkpoint covariances are retained. This is a linear spectral proxy, not an exact LLM/Adam/Muon noise model or an exact continuous-SDE discretization.

The uniform fit yields $(L_\infty, A, C, s, \nu) = (3.087814, 2.092771, 34.818720, 0.662386, 1.498722)$; the dynamic fit yields $(3.022923, 2.144309, 97.321529, 0.810008, 1.595182)$. Table 2 reports trajectory RMSE, MAE, and signed terminal errors for WSD calibration and cross-schedule prediction.

Table 2: WSD calibration and cross-schedule prediction errors. RMSE and MAE are computed over 34 evaluation points from step 600 to 3815. Terminal error is the fitted or predicted loss minus the observed loss at step 3815. All values are in nats. Target-schedule losses are not used for refitting.

Merge profile	Schedule	Role	RMSE	MAE	Terminal error
Uniform	WSD	Fit	0.001019	0.000879	−0.001024
Uniform	Cosine	Prediction	0.002551	0.002216	+0.001272
Uniform	Constant	Prediction	0.005699	0.005317	−0.007639
Dynamic EMA	WSD	Fit	0.000502	0.000412	−0.000123
Dynamic EMA	Cosine	Prediction	0.001445	0.001402	+0.001008
Dynamic EMA	Constant	Prediction	0.001464	0.001162	−0.001921

E.5 Decomposition and Budget Fits

For Figure 4, WSD, cosine, constant, and constant with physical-time dynamic EMA ($\kappa = 25$) are jointly fitted at 39 observations per curve, updates 500–19074. All four share L_∞, A, C, s, ν, rather than just the floor. The cutoffs are fixed at $\delta_S = 0, \delta_K = 0.5$. The fitted values are

$$(L_\infty, A, C, s, \nu) = (2.873146, 1.427441, 110.095823, 0.614853, 3.000000).$$

The fitted noise exponent reaches the upper bound $\nu = 3$, outside the kernel-theory range $(1, 2)$. Expanding the bound to $\nu \leq 8$ yields $(L_\infty, A, C, s, \nu) = (2.87704, 1.42424, 96.48236, 0.61277, 3.37434)$, with pooled RMSE decreasing from 0.009213 to 0.009146 nats (approximately 0.73%). At all 39 evaluation points, dynamic EMA still has a larger signal contribution and a smaller noise contribution than the unaveraged constant-LR trajectory. The terminal ordering across all four curves is also preserved for both components. Component magnitudes do change: the terminal dynamic-EMA noise contribution decreases from 0.03587 to 0.02975 nats.

The displayed components use the original bounded fit. Both fits are phenomenological descriptions outside the theoretical noise-exponent range. The total-loss panel displays observed validation losses; the other panels display fitted signal and noise contributions. Neither component is measured independently, and total-loss agreement does not identify a unique decomposition.

For Figure 5, fixed EMA uses $\beta = 0.99$ and physical-time dynamic EMA uses $\kappa = 50$. At each evaluation budget, each method independently selects its best validation loss over the three tested peak LRs $\{0.0005, 0.001, 0.005\}$, keeping its averaging parameter fixed. These are pointwise envelopes of intermediate evaluations from 20B-token runs, not separate runs trained to each smaller

budget, and not a single trajectory that changes LR between evaluations. They also do not optimize β or κ separately per budget.

We jointly fit all 58 observations per method in the actual range 5.24288B–20.000014336B tokens, using

$$L_j(D) = L_\infty + A_j(D/10^9)^{-\alpha_j}.$$

The fit minimizes equally weighted squared residuals in original loss space, not log loss. Both methods share $L_\infty = 2.8941126145$, while

$$(A_{\mathrm{EMA}}, \alpha_{\mathrm{EMA}}) = (0.6476886125, 0.5086709992),$$

$$(A_{\mathrm{DynEMA}}, \alpha_{\mathrm{DynEMA}}) = (0.7616640495, 0.6174135902).$$

Amplitudes are nonnegative, the floor lies between zero and the minimum observed loss, and exponents lie in $[0.01, 3]$. The optimization uses a 21×21 exponent grid and 12 bounded Nelder–Mead starts with constrained linear solves for the floor and amplitudes. The pooled RMSE is 0.00077839 nats. Sparse log-spaced markers are used only for display; they do not subsample the fitting data. These fits describe a finite budget range; neither a shared asymptotic floor nor per-budget optimality of the EMA parameters follows from them.

F Additional Related Work

This section expands the background in Section 2, focusing on empirical averaging practices, analyses of averaged estimators, and approaches to modeling training dynamics.

F.1 Empirical Averaging and Pretraining Practice

Averaging along a training trajectory predates its recent use in LLM pretraining. Stochastic weight averaging (SWA) improves generalization by combining iterates from an optimization trajectory (Izmailov et al., 2018); EMA has also been widely used in image generation with GANs and diffusion models (Karras et al., 2018; Yazıcı et al., 2019; Ho et al., 2020). These uses motivate averaging as an output estimator, while our focus is its effect on the evolving signal and noise terms.

For language-model training, Latest Weight Averaging (LAWA) averages a moving window of checkpoints to reach a target performance earlier (Kaddour, 2022). Sanyal et al. (2024) report greater benefits at higher LRs and with larger checkpoint spacing. Hägele et al. (2024) find that fixed-window SWA improves constant-LR trajectories without fully matching terminal cooldown; their constant-LR-plus-cooldown recipe supports multiple training horizons from one trajectory. DeepSeek-V3 uses EMA to estimate performance after LR decay (DeepSeek-AI, 2024).

At larger scales, checkpoint merging improves constant-LR training and predicts annealing behavior (Li et al., 2025b). WSM relates checkpoint weights to prescribed monotone decay coefficients and identifies merge duration as an important practical factor (Tian et al., 2025). Averaging has also been combined with data curricula (Luo et al., 2025a). Curci et al. (2026) find larger benefits during the stable-LR phase than during decay, relating this difference to parameter variation across checkpoints. Together, these observations motivate studying the merge window, LR schedule, and checkpoint spacing jointly. Our spacing result concerns preservation of a continuous-average noise rate as training grows; it does not imply that arbitrarily increasing the spacing of a fixed number of checkpoints helps.

F.2 Risk Bounds and Connections to LR Scheduling

Classical stochastic approximation explains averaging through its statistical effect on the output estimator (Polyak & Juditsky, 1992). Subsequent work provides non-asymptotic and nonparametric analyses (Moulines & Bach, 2011; Dieuleveut & Bach, 2016). In least-squares learning, tail averaging and minibatching can remove saturation and attain optimal statistical rates (Mücke et al., 2019). For overparameterized linear regression, Zou et al. (2021) derive sharp excess-risk bounds for constant-step-size SGD with iterate or tail averaging; Meterez et al. (2025) provide a simpler derivation using linear-algebraic tools. Under power-law source and capacity conditions, Meterez et al. (2026) show that averaging enables horizon-free schedules to attain minimax rates, with LLM

experiments supporting comparable performance to horizon-tuned cosine schedules. These results provide the statistical context for our design analysis. Our dynamical framework resolves profile-dependent leading constants and risk refinements to select averaging parameters within families that share a convergence rate.

A complementary line connects averaging to LR schedules through the optimization trajectory. Sandler et al. (2023) relate SWA and EMA to particular schedules under a simplified trajectory model. WSM makes the update-reweighting connection constructive for LLM pretraining under simplifying trajectory assumptions (Tian et al., 2025). The river–valley perspective explains terminal decay through suppression of sharp-direction oscillations while preserving progress along flatter directions (Wen et al., 2024). Related geometric analysis connects schedule-free optimization to implicit iterate averaging (Song et al., 2025). Extra-Merge describes an approximately one-dimensional trajectory of late-stage merged checkpoints and interprets averaging as a geometric low-pass filter (Zhou et al., 2026). Schedule–estimator interactions are also studied directly by Ousherovitch & Wang (2026). MM-FSL complements these perspectives by quantifying the signal cost and injection-time-dependent noise response for fractional averaging under general LR schedules.

F.3 Models of Loss Dynamics

Empirical extensions of scaling laws model entire training curves rather than only terminal loss. Accumulated-LR and annealing-dependent descriptions can fit and predict curves across schedules and horizons (Tissue et al., 2025; Luo et al., 2025b). FSL provides a mechanistic description in power-law kernel regression, separating gradient-flow signal learning from LR-dependent noise injection and subsequent forgetting (Li et al., 2025a). It also characterizes optimal power-decay schedules for easy tasks and a warmup-stable-decay structure for hard tasks (Li et al., 2026). FSL has further guided batch-size scheduling (Wang et al., 2026). Our extension concerns the output formed by averaging the training trajectory. Its forgetting kernel generally depends on evaluation and injection times separately, whereas the baseline last-iterate kernel depends on their difference. This distinction allows us to model how averaging and LR scheduling jointly determine merged-loss dynamics.