OPTIMAL LEARNING-RATE SCHEDULES UNDER FUNCTIONAL SCALING LAWS: POWER DECAY AND WARMUP-STABLE-DECAY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study optimal learning-rate schedules (LRSs) under the functional scaling law (FSL) framework introduced in Li et al. (2025), which accurately models the loss dynamics of both linear regression and large language model (LLM) pre-training. Within FSL, loss dynamics are governed by two exponents: a source exponent $s > 0$ controlling the rate of signal learning, and a capacity exponent $\beta > 1$ determining the rate of noise forgetting. Focusing on a fixed training horizon N, we derive the optimal LRSs and reveal a sharp phase transition. In the *easy-task regime* $s \geqslant 1 - 1/\beta$, the optimal schedule follows a **power decay** to zero, $\eta^*(z) = \eta_{\text{peak}}(1 - z/N)^{2\beta-1}$, where the peak learning rate scales as $\eta_{\text{peak}} \eqsim N^{-\nu}$ for an explicit exponent $\nu = \nu(s, \beta)$. In contrast, in the *hard-task regime* $s < 1 - 1/\beta$, the optimal LRS exhibits a **warmup–stable–decay (WSD)** (Hu et al., 2024) structure: it maintains the largest admissible learning rate for most of training and decays only near the end, with the decay phase occupying a vanishing fraction of the horizon. We further analyze optimal *shape-fixed* schedules, where only the peak learning rate is tuned–a strategy widely adopted in practice–and characterize their strengths and intrinsic limitations. This yields a principled evaluation of commonly used schedules such as cosine and linear decay. Finally, we apply the power-decay LRS to one-pass stochastic gradient descent (SGD) for kernel regression and show the last iterate attains the *exact* minimax-optimal rate, eliminating the logarithmic suboptimality present in prior analyses. Numerical experiments corroborate our theoretical predictions.

1 INTRODUCTION

Learning-rate schedules (LRSs) are a fundamental component of stochastic optimization, governing the trade-off between optimization progress and statistical noise during training, and thereby shaping convergence behavior. They play a central role in both theoretical analyses and practical algorithm design in modern machine learning.

The formal study of learning-rate schedules dates back to the seminal work of Robbins & Monro (1951), which established sufficient conditions for convergence of stochastic approximation, namely $\sum_{k=1}^{\infty} \eta_k = \infty$ and $\sum_{k=1}^{\infty} \eta_k^2 < \infty$. A canonical realization satisfying these conditions is the *polynomial-decay* schedule $\eta_k \propto k^{-\kappa}$ with $\kappa \in (1/2, 1)$, which has been extensively adopted in the analysis of both convex and non-convex optimization (Lacoste-Julien et al., 2012; Bubeck, 2014). In linear regression, such schedules—when combined with iterate averaging (Ruppert, 1988)—can achieve minimax-optimal convergence rates in a statistical sense (Bach & Moulines, 2013; Dieuleveut & Bach, 2015; Mücke et al., 2019). More recently, a line of work has focused on the performance of the last iterate, showing that *exponential-decay* schedules can attain nearly optimal rates without averaging (Ge et al., 2019; Wu et al., 2022a; Zhang et al., 2024).

Despite the rich theoretical literature, two notable gaps remain. First, the learning-rate schedules that are theoretically optimal or near-optimal in classical analyses—such as polynomial or exponential decay—are rarely used in modern large-scale training. Second, practical training overwhelmingly relies on alternative schedules, most notably cosine decay (Loshchilov & Hutter, 2016; Hoffmann et al., 2022; Touvron et al., 2023) and warmup–stable–decay (WSD) (Zhai et al., 2022; Hu et al.,

2024; Liu et al., 2024; Team et al., 2025). WSD schedules typically keep the learning rate constant for the majority of the training horizon (often up to 80%) and defer decay to a short final annealing phase. This stark contrast raises a fundamental theoretical question: *why can schedules that delay decay until the very end of training still perform well?*

A key reason for these gaps is that most existing analyses do not derive learning-rate schedules from a principled approach. Instead, they follow a "propose-and-verify" paradigm: a specific schedule—typically motivated by classical sufficient conditions or heuristics—is posited a priori and then shown to achieve a desired convergence rate under certain conditions. While this approach provides useful guarantees, it decouples performance analysis from schedule design and offers limited guidance on important questions such as which decay shapes are preferable and how optimal schedules depend on problem characteristics like model capacity and task difficulty.

Recent work by Li et al. (2025) offers a complementary perspective in a controlled yet expressive setting. Under feature-space linear regression with power-law structure (Bahri et al., 2024; Bordelon et al., 2024; Lin et al., 2024; Paquette et al., 2024; Li et al., 2025), they derive a **functional scaling law (FSL)** that expresses the loss as an explicit and analytically tractable functional of the learning-rate schedule. Moreover, empirical evidence in Li et al. (2025) shows that this functional characterization remains accurate in large language model (LLM) pre-training, despite the substantial gap between the theoretical setting and practical LLM training. While the FSL provides a precise description of how a *given* learning-rate schedule shapes the loss, it does not address the problem of learning-rate schedule design, nor the characterization of optimal schedules.

In this paper, we focus on the problem of **optimal learning-rate schedules (LRSs)**. Specifically, we characterize the LRS that minimizes the final-step loss over a fixed training horizon N (equivalently, a fixed data budget). Under the FSL framework, this problem can be formulated as a constrained variational optimization problem, in which the loss is governed by a competition between *signal learning*, controlled by the source exponent $s > 0$, and *noise forgetting*, determined by the capacity exponent $\beta > 1$. Our main contributions are summarized as follows.

Optimal learning-rate schedules (LRSs). We derive that the optimal LRS depends critically on the task difficulty. In the *easy-task regime* ($s \geqslant 1 - 1/\beta$), the optimal LRS follows a **power decay** to zero:

$$\eta_*(z) = \eta_{\text{peak}} \left(1 - z/N\right)^{2\beta - 1},$$

where peak learning rate $\eta_{\text{peak}} \eqsim N^{-\frac{1+s\beta-\beta}{1+s\beta}}$. In contrast, in the *hard-task regime* ($s < 1 - 1/\beta$), the optimal LRS exhibits a **WSD-like** structure: the decay phase occupies only a $o_N(1)$ fraction of the training horizon, while retaining the same power-decay profile as the easy-task regime. Figure 1(left) provides an illustration of these optimal LRSs.

Shape-fixed optimality and capacity saturation. To isolate the essential structure underlying optimal LRSs, we consider a class of *fractional* schedules of the form $\eta(z) = \eta_0\, \zeta(z/N)$, which depend on training steps only through the relative progress z/N and exhibit a power-decay tail near the end of training: $\zeta(x) \propto (1 - x)^\gamma$ as $x \to 1$. Within this shape-fixed setting—where the decay shape is fixed and only the peak learning rate is tuned—we uncover a *capacity saturation* phenomenon: such schedules adapt to model capacity only up to $\beta \leqslant \gamma + 1$, beyond which the achievable convergence rate saturates regardless of peak-rate tuning. Figure 1(right) summarizes this behavior via a phase diagram of convergence rates over the (β, s) plane. This characterization clarifies the strengths and limitations of widely used practical schedules; in particular, Figure 1(middle) shows that cosine LRS ($\gamma = 2$; see Section 4) indeed exhibits the predicted saturation.

Improved convergence rates for kernel regression. Finally, to substantiate the predictions of the continuous-time FSL analysis, we provide a rigorous discrete-time analysis showing that one-pass SGD with a power-decay LRS attains the *exact* optimal convergence rate at the last iterate. To the best of our knowledge, this is the first such result without logarithmic factors, improving upon prior analyses based on exponential-decay LRS (Wu et al., 2022a; Lin et al., 2024; Li et al., 2025).

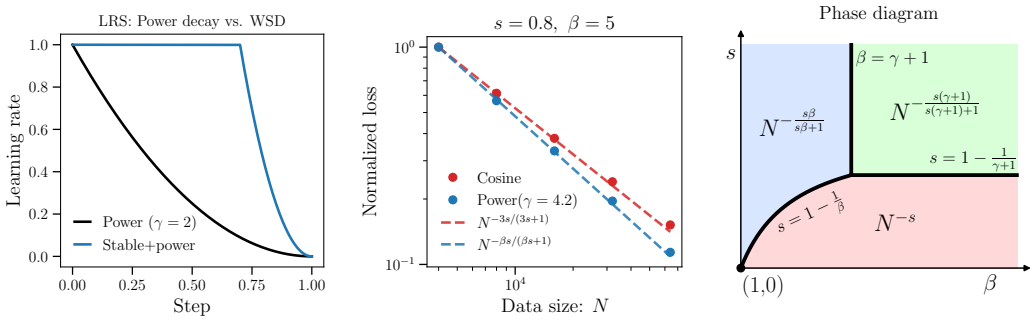

Figure 1: (left) Illustration of optimal learning-rate schedules (LRSs): power decay in the easy-task regime and WSD with power decay in the hard-task regime. **(middle)** Performance comparison of cosine ($\gamma = 2$) and power-decay ($\gamma = 4.2$) LRSs for feature-space linear regression with source exponent $s = 0.8$ and capacity exponent $\beta = 5$. Power decay achieves the minimax-optimal rate $N^{-\beta s/(\beta s+1)}$, whereas cosine decay suffers from capacity saturation and exhibits the suboptimal rate predicted by our theory (corresponding to the green region in the right phase diagram). For each data size, we perform 500 independent runs of SGD, and tune the peak learning rate to minimize the average final-step loss. **(right)** Phase diagram of convergence rates under shape-fixed fractional LRSs (Theorem 4.3). Each region in the (β, s) plane corresponds to distinct convergence rates. The vertical boundary $\beta = \gamma + 1$ marks a capacity-saturation threshold induced by fixing the decay shape; in the green region, this restriction leads to suboptimal convergence rates.

2 Preliminaries

Notation. We write $\eqsim$ to denote equivalence up to a multiplicative constant factor, and $\lesssim$ (resp. $\gtrsim$) to denote an inequality up to a multiplicative constant factor. For two nonnegative functions $f, g : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, we write $f(t) \eqsim g(t)$ if there exist constants $C_1, C_2 > 0$ (independent of t) such that $C_1 f(t) \leqslant g(t) \leqslant C_2 f(t),\ \forall\, t \geqslant 0$. We write a sequence $a_N = o_N(1)$ if $\lim_{N\to\infty} a_N = 0$.

2.1 Feature-Space Linear Regression

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathbb{R}$ and denote by $\mathcal{D}_{\mathcal{X}}$ the marginal on input domain $\mathcal{X}$. Samples $(\mathbf{x}, y) \sim \mathcal{D}$ satisfy $y = f^*(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{x}$. We assume $f^*(\mathbf{x}) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(\mathbf{x}) \rangle = \sum_{j=1}^{\infty} \theta_j^* \phi_j(\mathbf{x})$ for some feature map $\boldsymbol{\phi} : \mathcal{X} \to \ell^2$ and $\boldsymbol{\theta}^* \in \ell^2$.

Assumption 2.1 (Hypercontractive features)**.** For any $\mathbf{u}, \mathbf{v} \in \ell^2$, it holds that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\langle \mathbf{u}, \boldsymbol{\phi}(\mathbf{x})\rangle^2 \langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x})\rangle^2] \eqsim \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\langle \mathbf{u}, \boldsymbol{\phi}(\mathbf{x})\rangle^2] \mathbb{E}_{\mathbf{x}}[\langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x})\rangle^2].$$

Define the feature covariance operator by $\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\right]$. Without loss of generality, we assume that $\mathbf{H}$ is diagonalized. Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$ denote its eigenvalues, ordered non-increasingly. This assumption is without loss of generality due to the rotation equivariance of SGD.

Let $\{(\mathbf{x}_k, y_k)\}_{k=0}^{N-1}$ be N samples drawn independently from $\mathcal{D}$. To learn the target function, we consider the linear model $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle$ trained by **one-pass SGD**:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \nabla_{\boldsymbol{\theta}} \left(\tfrac{1}{2} (f(\mathbf{x}_k; \boldsymbol{\theta}_k) - y_k)^2 \right), \qquad \boldsymbol{\theta}_0 = \mathbf{0}, \tag{1}$$

where $(\eta_0, \eta_1, \ldots, \eta_{N-1})$ is the *learning-rate schedule (LRS)*. We measure the performance using the excess risk $\mathcal{E}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\left[(f(\mathbf{x}; \boldsymbol{\theta}) - f^*(\mathbf{x}))^2\right]$. The choice of the LRS plays a crucial role in determining the final-step performance. Our goal is to characterize the *optimal* LRS for a fixed N.

2.2 Functional Scaling Laws

We adopt the FSL framework of Li et al. (2025) to characterize the impact of LRS on the *final-step loss* and to identify optimal schedules. This framework operates under a power-law data assumption:

Assumption 2.2 (Power-law structures)**.** The following conditions hold for each $j \in \mathbb{N}_+$:

$$\lambda_j \eqsim j^{-\beta}, \qquad \lambda_j |\theta_j^*|^2 \eqsim j^{-(1+s\beta)}. \tag{2}$$

The condition $\lambda_j \eqsim j^{-\beta}$ is referred to as the **capacity condition**, where the exponent β controls the decay rate of the feature spectrum. A smaller β yields slower eigenvalue decay, corresponding to a higher capacity. The condition $\lambda_j |\theta_j^*|^2 \eqsim j^{-(1+s\beta)}$ is referred to as the **source condition**, which quantifies the alignment between the target function and the feature space. The capacity exponent s measures the *relative difficulty*: smaller values of s correspond to more challenging problems, in which a larger fraction of the signal energy is concentrated in high-frequency components.

In the FSL framework, rather than working directly with the discrete SGD, one adopts a continuous-time modeling perspective. Specifically, the discrete SGD is modeled by an Itô stochastic differential equation (SDE) (Li et al., 2019; Orvieto & Lucchi, 2019). A central concept in FSL is the **intrinsic time**, which encodes the effect of the LRS. At iteration k, the intrinsic time is given by $t_k := \sum_{j=0}^{k-1} \eta_j$. At the continuous level, $t = \Gamma(z) := \int_0^z \eta(z)\,\mathrm{d}z$, where z and $\eta(z)$ denote the continuous step and learning rate z step, respectively. Under this intrinsic-time parametrization, the SGD dynamics are modeled by the Itô SDE

$$\mathrm{d}\bar{\boldsymbol{\theta}}_t = -\nabla \mathcal{E}(\bar{\boldsymbol{\theta}}_t)\ \mathrm{d}t + \sqrt{\varphi(t)}\,\boldsymbol{\Sigma}(\bar{\boldsymbol{\theta}}_t)^{1/2}\ \mathrm{d}\mathbf{W}_t, \tag{3}$$

where $\bar{\boldsymbol{\theta}}_t$ denotes the parameter at intrinsic time t, $\varphi(t) := \eta(\Gamma^{-1}(t))$ is the learning rate in intrinsic time, $\mathbf{W}_t$ is a standard Wiener process, and $\boldsymbol{\Sigma}(\bar{\boldsymbol{\theta}}_t)$ denotes the covariance of gradient noise.

Theorem 4.4 of Li et al. (2025) shows that, in the label-noise–dominated regime $\sigma \gtrsim 1$ and under the stability condition $\sup_{t\geqslant 0} \varphi(t) \leqslant C$ for a sufficiently small constant $C > 0$, the solution to the SDE equation 3 satisfies, for all $t \gtrsim 1$,

$$\mathbb{E}[\mathcal{E}(\bar{\boldsymbol{\theta}}_t)] \eqsim \mathcal{F}[t] := \underbrace{(1+t)^{-s}}_{\text{signal learning}} + \int_0^t \underbrace{\mathcal{K}(t-\tau)}_{\text{forgetting kernel}}\ \varphi(\tau)\,\mathrm{d}\tau, \tag{4}$$

where $\mathcal{K}(t) := (1+t)^{-(2-1/\beta)}$. This FSL establishes a functional-level map from the LRS function to the loss at intrinsic time t and notably, the two terms exhibits a clean interpretation:

- The **signal-learning** term corresponds to learning under full-batch gradient descent, capturing the rate at which SGD extracts signal f^*. This rate is determined by the relative difficulty s.
- The **noise-accumulation** term characterizes how the LRS changes the accumulation and dissipation of gradient noise. The forgetting kernel $\mathcal{K}(t-\tau)$ characterizes how the noise injected at time τ still affects the loss at time t. Due to $\mathcal{K}(t) = (t+1)^{-(2-1/\beta)}$, a higher-capacity model (smaller β) tends to forget noise more slowly.

Remark 2.3. The theoretical and empirical evidence in Li et al. (2025) demonstrates that the FSL equation 4 can accurately model the effect of LRSs in feature-space linear regression, and even in large-scale LLM pre-training. In this work, we take this FSL as a starting point for studying optimal LRS—a direction that lies outside the scope of Li et al. (2025).

3 Optimal Learning Rate Schedules

We begin by asking the following natural question:

Given a data budget N, what is the optimal LRS when the loss dynamics follows the FSL equation 4?

To formalize the resource constraint, we move to domain of training steps. Let $\eta(z)$ denote the learning rate at (continuous) step z and define the intrinsic time $t(z) = \int_0^t \eta(z)\,\mathrm{d}z$. Then, $\mathrm{d}\tau = t'(z)\,\mathrm{d}z$ and $\eta(\tau) = t'(z)$. For a fixed N, by a change of variable, the noise term becomes

$$\int_0^{t(N)} \mathcal{K}(t(N)-\tau)\varphi(\tau)\,\mathrm{d}\tau = \int_0^N \mathcal{K}(t(N)-t(z))\left(t'(z)\right)^2\,\mathrm{d}z. \tag{5}$$

Hence, seeking the optimal LRS can be formulated as the resource-constrained *variational problem*:

$$\begin{aligned} \min_{t\in \mathrm{AC}([0,N])} \quad & \bar{\mathcal{F}}[t] \coloneqq (1+t(N))^{-s} \;+\; \int_0^N \mathcal{K}(t(N)-t(z))\left(t'(z)\right)^2 \mathrm{d}z \\ \text{s.t.} \quad & t(0)=0, \\ & 0 \leqslant t'(z) \leqslant \eta_{\text{stability}} \quad \text{for a.e. } z\in[0,N]. \end{aligned} \tag{6}$$

where $\mathrm{AC}([0,N])$ denotes the set of absolutely continuous function over $[0,N]$. The constraint $0 \leqslant t'(z) \leqslant \eta_{\text{stability}}$ enforces the nonnegativity of the learning rate and an upper bound required for training stability (Wu et al., 2018; 2022b).

Before stating our characterization of the optimal LRS, we first clarify what constitutes a good LRS. From equation 6, the signal-learning term favors a large intrinsic time $t(N)$, which corresponds to using a larger learning rate, whereas the noise-accumulation term favors a smaller learning rate. A good LRS must balance these two competing effects.

Theorem 3.1 (Optimal learning-rate schedules)**.** *Let t_* be a minimizer of equation 6 and define $\eta^*(z) \coloneqq t_*'(z)$ and $\mathcal{E}_N^* = \bar{\mathcal{F}}[t_*]$ be the final-step loss. Then the following holds.*

- ***Easy-task regime*** *($s \geqslant 1-\frac{1}{\beta}$): the optimal LRS follows a power decay to zero:*

$$\eta^*(z) = \eta_{\text{peak}}\left(1+o_N(1)-\frac{z}{N}\right)^{2\beta-1}, \qquad \eta_{\text{peak}} \eqsim N^{-\frac{1+s\beta-\beta}{1+s\beta}}, \tag{7}$$

  under which, the final-step loss satisfies

$$\mathcal{E}_N^* \eqsim N^{-\frac{s\beta}{s\beta+1}}.$$

- ***Hard-task regime*** *($s < 1-\frac{1}{\beta}$): the optimal LRS is a WSD form:*

$$\eta^*(z) = \begin{cases} \eta_{\text{stability}}, & 0 \leqslant z \leqslant N_1, \\ \eta_{\text{stability}}\left(1+o_N(1)-\frac{z-N_1}{N-N_1}\right)^{2\beta-1}, & N_1 < z \leqslant N, \end{cases} \tag{8}$$

  where the ratio of the decay phase satisfies

$$r_N^* \coloneqq \frac{N-N_1}{N} \eqsim N^{-\frac{(1-1/\beta)-s}{2-1/\beta}} = o_N(1). \tag{9}$$

  Under this optimal LRS, the final-step loss scales as

$$\mathcal{E}_N^* \eqsim N^{-s}.$$

This theorem provides a characterization of the optimal LRS under the assumption that FSL correctly describes the loss dynamics. The proof is deferred to Appendix C.

We see that the optimal LRS depends critically on the task difficulty. In the **easy-task regime**, the optimal LRS decreases monotonically from the beginning of training and follows a power decay to zero. Moreover, the peak learning rate scales with the training horizon according to a power law. In contrast, in the **hard-task regime**, the optimal LRS exhibits a WSD structure: it maintains the largest admissible learning rate $\eta_{\text{stability}}$ for most of training and decays only near the end. Notably, the decay phase occupies only a vanishing fraction of the training horizon as $N \to \infty$.

We next provide intuition for explaining these results through the trade-off between signal learning and noise forgetting:

- **The optimal decay shape.** In both regimes, the decay phase follows a power-decay profile whose exponent depends only on the capacity exponent β. Intuitively, higher-capacity models forget noise more slowly and therefore require a faster decay of the learning rate to ensure that injected noise dissipates sufficiently. More precisely, the optimal decay profile satisfies $\varphi^*(t) \propto \mathcal{K}(T-t)^{-1/2}$ in intrinsic time; see equation 19 and the accompanying derivation. The power-decay form arises from the power-law behavior of the forgetting kernel $\mathcal{K}(\cdot)$.

- **The emergence of a prolonged stable phase for hard tasks.** When s is small, the learning problem is hard: signal learning progresses very slowly and requires a large intrinsic time. In contrast, the rate of noise forgetting is governed by the capacity exponent β and is independent of the task difficulty s. As a result, the optimal strategy is to maintain the largest stable learning rate $\eta_{\text{stability}}$ for an extended period, and to allocate only a short decay phase at the end to eliminate accumulated noise. As predicted by equation 9, harder tasks (smaller s) indeed need longer stable phases in the optimal LRS.

4 Shape-Fixed Learning-Rate Schedules: Optimality and Limitations

The analysis in Section 3 provides a complete characterization of the optimal learning-rate schedule (LRS) and the best achievable performance. Beyond optimality, these schedules share two salient structural features: a **fractional** form, in which the schedule depends on training progress only through the normalized step z/N, and a **power-decay to zero** in the terminal phase. To disentangle the contributions of these two structures, we consider the following family of LRSs.

Definition 4.1 (Fractional LRS with power-decay tail)**.** Given a training horizon N, a LRS $\eta_N : [0, N] \to \mathbb{R}_{\geqslant 0}$ is called *fractional* if it can be written as

$$\eta_N(z) = \eta_0 \, \zeta\Big(\frac{z}{N}\Big), \qquad z \in [0, N], \tag{10}$$

where $\eta_0 > 0$ is the initial learning rate and the profile function $\zeta : [0, 1] \to [0, 1]$ satisfies $\zeta(0) = 1$. It is said to have a *power-decay tail* if there exist constants $\gamma > 0$ and $\delta \in (0, 1)$ such that

$$\zeta(x) \eqsim (1 - x)^\gamma, \qquad \forall\, x \in [\delta, 1]. \tag{11}$$

Examples. This class of schedules includes nearly all commonly used in practice:

- **Constant LRS** $\eta_N(z) \equiv \eta_0$: $\zeta(x) \equiv 1$, corresponding to $\gamma = 0$.
- **Cosine decay** $\eta_N(z) = \frac{1}{2}\eta_0(1 + \cos(\pi z/N))$: here $\zeta(x) = \frac{1}{2}(1 + \cos(\pi x))$. A Taylor expansion at $x = 1$ yields $\zeta(x) = \frac{\pi^2}{4}(1 - x)^2 + O\big(|1 - x|^4\big)$, and hence $\gamma = 2$.
- $1-$**sqrt decay** $\eta_N(z) = \eta_0\big(1 - \sqrt{z/N}\big)$: $\zeta(x) = 1 - \sqrt{x}$ admits the expansion $\zeta(x) = \frac{1}{2}(1 - x) + O\big(|1 - x|^2\big)$, implying $\gamma = 1$.
- **Power decay** $\eta_N(z) = \eta_0(1 - z/N)^\gamma$: $\zeta(x) = (1 - x)^\gamma$, which satisfies Definition 4.1 for any $\gamma > 0$. The case $\gamma = 1$ is commonly referred to as *linear decay* in the literature.

First, for a fractional LRS, the total intrinsic time is given by $t(N) = \int_0^N \eta(u)\, \mathrm{d}u = \eta_0 N \int_0^1 \zeta(x)\, \mathrm{d}x$. This reveals a key property of fractional schedules: when η_0 is fixed, the total intrinsic time grows *linearly* with the training horizon N. As a result, fractional LRSs can provide sufficient intrinsic time for *signal learning*. Second, the power-decay tail ensures that the learning rate vanishes smoothly in the terminal phase, allowing SGD to forget accumulated noise at a controlled rate.

Theorem 4.2 (Scaling law for fractional LRS)**.** *Let η_N be a fractional LRS, $\alpha = \min\{\beta,\ \gamma + 1\}$ and $T_N = \eta_0 N \int_0^1 \zeta(x)\, \mathrm{d}x$. Then the final-step loss satisfies*

$$\mathcal{F}[\eta_N] \ \eqsim\ T_N^{-s} + \eta_0 T_N^{-(1-\frac{1}{\alpha})} \big(\log T_N\big)^{\mathbf{1}\{\beta=\gamma+1\}}. \tag{12}$$

The quantity T_N represents the total intrinsic training time. The second term in equation 12 exhibits a **capacity saturation** effect: when the LRS decays more slowly than the forgetting kernel, noise forgetting is governed by the decay rate of the LRS rather than the model's intrinsic capacity. As a result, the exponent $\alpha = \min\{\gamma + 1, \beta\}$ acts as an **effective capacity exponent** governing noise forgetting. In particular, choosing a fast-decay tail $\gamma > \beta - 1$ recovers the intrinsic capacity $\alpha = \beta$. At the boundary case $\gamma + 1 = \beta$, an additional logarithmic factor appears.

Fix the tail exponent γ and optimize only over the peak learning rate η_0 on the right-hand side of equation 12. Let $\eta_0^\star$ denote the minimizer, and let $\mathcal{E}_N^\star$ be the resulting final-step excess risk.

Theorem 4.3 (Optimal fractional LRS). *Consider a fractional LRS with fixed tail exponent γ, and let $\alpha = \min\{\beta, \gamma + 1\}$. Then:*

- ***Easy-task regime*** *($s \geqslant 1 - \frac{1}{\alpha}$):*
  - *If $\beta = \gamma + 1$, then* $\eta_0^* \eqsim N^{-\frac{s-1+1/\alpha}{s+1/\alpha}} (\log N)^{-\frac{1}{s+1/\alpha}}$, $\mathcal{E}_N^* \eqsim N^{-\frac{s\alpha}{s\alpha+1}} (\log N)^{\frac{s\alpha}{s\alpha+1}}$.
  - *If $\beta \neq \gamma + 1$, then*

$$\eta_0^* \eqsim N^{-\frac{s-1+1/\alpha}{s+1/\alpha}}, \qquad \mathcal{E}_N^* \eqsim N^{-\frac{s\alpha}{s\alpha+1}}.$$

- ***Hard-task regime*** *($s < 1 - \frac{1}{\alpha}$):*

$$\eta_0^* \eqsim 1, \qquad \mathcal{E}_N^* \eqsim N^{-s}.$$

Comparing Theorem 4.3 with Theorem 3.1, we observe that fixing the decay shape and optimizing only the peak learning rate shifts the boundary between the easy- and hard-task regimes from $s = 1 - \frac{1}{\beta}$ to $s = 1 - \frac{1}{\alpha}$. Figure 1(right) provides a diagrammatic illustration of how the convergence rate under fractional LRSs varies with (s, β) for different γ's.

Remark 4.4. We observe that, in both regimes, a fractional LRS with decay exponent $\gamma > \beta - 1$ is sufficient to attain the same convergence rates as those achieved by the exact optimal LRS. By contrast, the latter employs the sharper decay exponent $\gamma = 2\beta - 1$ and, in the hard-task regime, requires a prolonged stable phase. This comparison highlights that an exact optimal LRS can differ substantially from schedules that merely achieve optimal convergence rates.

Practical implications. The above analysis mirrors a common practice in large-scale training, where one fixes the shape of LRS and tunes only the peak learning rate. Theorem 4.3 therefore provides a principled understanding of both the strengths and limitations of this approach. In particular, for a fractional LRS with power-decay exponent γ, tuning only the peak learning rate achieves the *optimal* convergence rate whenever the model capacity satisfies $\beta \leqslant \gamma + 1$. In contrast, when $\beta > \gamma + 1$, corresponding to lower-capacity models, this strategy can lead to suboptimal rates.

As concrete examples, cosine decay ($\gamma = 2$) achieves optimal rates for $\beta \leqslant 3$, while linear decay and 1−sqrt decay ($\gamma = 1$) achieve optimal rates for $\beta \leqslant 2$. Since modern deep learning models typically operate in high-capacity regimes, these results may provide an explanation of why shape-fixed LRSs perform well in practice.

5 SGD for Kernel Regression

In this section, we show that the optimal LRS derived within the continuous-time FSL framework can be indeed transferred to discrete-time SGD for kernel regression.

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a kernel if there exists a feature map $\phi : \mathcal{X} \mapsto \mathbb{H}$ such that $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathbb{H}}$, where $\mathbb{H}$ is a separable Hilbert space. Consider the hypothesis class $\mathcal{H} := \{f_{\boldsymbol{\theta}} = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\cdot) \rangle_{\mathbb{H}} : \boldsymbol{\theta} \in \mathbb{H}\}$. Then $\mathcal{H}$ is the reproducing kernel Hilbert space (Aronszajn, 1950) corresponding to the kernel K. Suppose $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[K(\mathbf{x}, \mathbf{x})] < \infty$. Then, there exist nonnegative eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and an orthonormal system $\{\mathbf{e}_j\}_{j=1}^{\infty}$ in $L^2(\mathcal{D}_{\mathcal{X}})$ such that $K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \mathbf{e}_j(\mathbf{x}) \mathbf{e}_j(\mathbf{x}')$, where the convergence is in $L^2(\mathcal{D}_{\mathcal{X}} \times \mathcal{D}_{\mathcal{X}})$.

Assumption 5.1 (Capacity condition). There exists a $\beta > 1$ such that $\lambda_j \lesssim j^{-\beta}$ for all $j \in \mathbb{N}_+$.

Assumption 5.2 (Source condition). $f^* = \sum_j a_j \lambda_j^{s/2} \mathbf{e}_j$ with $\sum_j a_j^2 \leqslant 1$ for some $s > 0$.

These are classic capacity and source conditions, widely used in the study of kernel methods (Ying & Pontil, 2008; Dieuleveut & Bach, 2015; Dieuleveut et al., 2017; Pillaud-Vivien et al., 2018; Guo & Shi, 2019; Guo et al., 2024; Mao & Guo, 2024). Assumption 2.2 can be viewed as a power-law variant of these conditions. Analogously, the capacity condition controls the effectively size of the hypothesis space through the eigenvalue decay. The source condition captures the regularity of the target function f^*. The parameter s measures how well the target aligns with the spectral geometry of the kernel and a larger s implies a smoother target function.

The following result from Caponnetto & De Vito (2007); Steinwart et al. (2009) established a lower bound on the best achievable convergence rate, independent of the choice of estimator. Suppose Assumption 5.1 and 5.2 hold. Let $\mathcal{D}_N = \{(\mathbf{x}_k, y_k)\}_{k=0}^{N-1}$ be a dataset of N samples drawn *i.i.d.* from $\mathcal{D}$. The minimax risk satisfies

$$\inf_{\widehat{\boldsymbol{\theta}}_N} \sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}_N} \left[\mathcal{E}(\widehat{\boldsymbol{\theta}}_N) \right] \gtrsim N^{-\frac{s\beta}{s\beta+1}}, \tag{13}$$

where the infimum is taken over all estimators $\widehat{\boldsymbol{\theta}}_N$ (i.e., measurable functions of $\mathcal{D}_N$) and the supremum is taken over all data distributions $\mathcal{D}$ satisfying the stated assumptions. This minimax lower bound serves as a fundamental criterion for assessing the optimality of learning algorithms.

We now present our main result. Through a detailed analysis of the dynamics of SGD with the power decay LRS, we derive the convergence rate of SGD with power decay.

Theorem 5.3 (Convergence rate of SGD with power decay)**.** *Suppose Assumptions 5.1 and 5.2 hold. Consider LRS $\eta_k = \eta_0(1 - k/N)^\gamma$. The following statements hold.*

- ***Easy-task regime*** *($s \geqslant 1 - \frac{1}{\beta}$). Choosing $\eta_0 \eqsim N^{-\frac{s\beta-\beta+1}{s\beta+1}}$ and $\gamma > \beta - 1$ yields*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim N^{-\frac{s\beta}{s\beta+1}}.$$

- ***Hard-task regime*** *($s < 1 - \frac{1}{\beta}$). Choosing $\eta_0 \eqsim 1$ and $\gamma > \frac{s}{1-s}$ yields*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim N^{-s}.$$

The proof is deferred to Appendix E.4 due to its technical complexity. The rate for the easy-task regime matches the minimax optimal rate equation 13. In the hard-task regime, the obtained rate is also optimal, in the sense that it coincides with the best rate achievable by one-pass SGD:

Proposition 5.4. *Suppose that $\max_{i\in[N]} \eta_i \lesssim 1$. If $s < 1 - \frac{1}{\beta}$, then $\sup_{\mathcal{D}} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim N^{-s}$.*

The $\sup_{\mathcal{D}}$ denotes the supremum over all data distributions satisfying Assumption 5.1 and 5.2. The proof of this proposition is deferred to Appendix E.5.

To the best of our knowledge, Theorem 5.3 provides the first theoretical guarantee that last-iterate SGD attains the exact minimax-optimal convergence rate in the easy-task regime. Existing guarantees achieving minimax rates rely on iterate averaging. Notably, this improvement is achieved by employing a power-decay LRS, which is inspired by our optimal LRS analysis under the FSL framework. By contrast, the analyses in Lin et al. (2024) and Zhang et al. (2024) rely on exponential-decay LRS and consequently incur additional logarithmic factors. Conceptually, power-decay LRSs precisely match the power noise-forgetting dynamics, striking the optimal balance between signal learning and noise dissipation; in contrast, exponential decay over-regularizes late iterations, leading to logarithmic suboptimality for last-iterate SGD.

6 Conclusion

In this paper, we study optimal learning rate schedules (LRSs) under a fixed training horizon within the functional scaling law framework. We derive optimal LRSs, identify their essential structural properties, and use these insights to explain the empirical success of cosine decay and warmup–stable–decay (WSD) schedules as well as their limitations. This characterization explains why different schedules can behave similarly in some regimes, yet diverge sharply in others, as a function of task difficulty and model capacity. Finally, we further leverage this structural understanding to improve the last-iterate convergence rate of SGD for kernel regression.

The core insight underlying our analysis is a fundamental trade-off between *signal learning* and *noise forgetting*. This perspective provides a unified lens for many existing LRSs and offers principled guidance for designing new ones, thereby bridging theoretical and practical training regimes. A promising direction for future work is to extend this analysis to more realistic optimization settings, incorporating adaptivity, momentum, and batch-size effects.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023. 13

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. 7

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

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. 2, 13

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs. *arXiv preprint arXiv:2502.15938*, 2025. 13

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024. 2, 13

Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15, 2014. 1

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007. 8

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023. 13

Aymeric Dieuleveut and Francis Bach. Non-parametric stochastic approximation with large step sizes. *Annals of Statistics*, 44(4), 2015. 1, 7, 13

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017. 7

Matthew Fahrbach, Adel Javanmard, Vahab Mirrokni, and Pratik Worah. Learning rate schedules in the presence of distribution shift. In *International Conference on Machine Learning*, pp. 9523–9546. PMLR, 2023. 13

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32, 2019. 1

Zheng-Chu Guo and Lei Shi. Fast and strong convergence of online learning algorithms. *Advances in Computational Mathematics*, 45(5):2745–2770, 2019. 7

Zheng-Chu Guo, Andreas Christmann, and Lei Shi. Optimality of robust online learning. *Foundations of Computational Mathematics*, 24(5):1455–1483, 2024. 7

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. 13

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, and Others. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 13

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 13

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and Others. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1, 13

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 1

Arlind Kadra, Maciej Janowski, Martin Wistuba, and Josif Grabocka. Power laws for hyperparameter optimization. *arXiv preprint arXiv:2302.00441*, 2023. 13

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 13

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024. 13

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012. 1

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *arXiv preprint arXiv:2509.19189*, 2025. 1, 2, 3, 4, 13

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017. 13

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40): 1–47, 2019. 4

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pp. 6553–6564. PMLR, 2021. 13

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024. 2, 8, 13

Licong Lin, Jingfeng Wu, and Peter L Bartlett. Improved scaling laws in linear regression via data reuse. *arXiv preprint arXiv:2506.08415*, 2025. 13

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and Others. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

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. *arXiv preprint arXiv:2503.12811*, 2025. 13

Yuan Mao and Zheng-Chu Guo. Online regularized learning algorithm for functional data. *Journal of Complexity*, 82:101825, 2024. 7

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating SGD saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 13

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023. 13

Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 32, 2019. 4

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024. 2, 13

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018. 7

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951. 1

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. 1

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. *arXiv preprint arXiv:2501.18965*, 2025. 13

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pp. 79–93, 2009. 8

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. 2

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *arXiv preprint arXiv:2408.11029*, 2024. 13

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and Others. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

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. 13

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pp. 24280–24314. PMLR, 2022a. 1, 2, 13

Lei Wu, Chao Ma, et al. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018. 5

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35: 4680–4693, 2022b. 5

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. *arXiv preprint arXiv:2511.13421*, 2025. 13

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008. 7

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, 2022. 1

Haihan Zhang, Yuanshi Liu, Qianwen Chen, and Cong Fang. The optimality of (accelerated) SGD for high-dimensional quadratic optimization. *arXiv preprint arXiv:2409.09745*, 2024. 1, 8

Part I

Appendix

Table of Contents

A Related Work

Neural scaling laws. Hestness et al. (2017) first observed that the performance of deep learning models follows predictable power-law relationships in model and data size, later formalized as neural *scaling laws* (Kaplan et al., 2020). These laws have since guided large-scale training and been refined across architectures and training regimes (Henighan et al., 2020; Hoffmann et al., 2022; Kadra et al., 2023; Aghajanyan et al., 2023; Muennighoff et al., 2023; Kumar et al., 2024; Tissue et al., 2024; Luo et al., 2025), with parallel theoretical efforts explaining their origins and mechanisms (Bordelon et al., 2024; Lin et al., 2024; Bahri et al., 2024; Paquette et al., 2024; Lin et al., 2025; Yan et al., 2025). Our work fits into this line of research by providing a scaling-law analysis of the design of learning-rate schedules. In particular, we study the structure of optimal schedules by leveraging the FSL framework introduced by Li et al. (2025).

Optimal learning-rate schedules for linear regression. Previous analyses of optimal learning-rate schedules for linear regression often formulate the problem as an optimal control problem. However, such analyses are technically challenging and typically restricted to low-dimensional settings or isotropic Hessians (Li et al., 2017; Fahrbach et al., 2023). By contrast, leveraging the FSL framework, we develop a principled variational approach to derive and analyze optimal LRSs.

Understanding cosine and WSD schedules. For cosine schedules, Li et al. (2021) attributes their empirical success to a *duration-aware* property: the cumulative learning rate scales linearly with the total training horizon while the learning rate itself decays to a horizon-independent minimum. Standard polynomial decay schedules fail to satisfy this dual requirement. In this work, we generalize this property through a class of *fractional* LRS and identify the regimes in which it attains optimal rates and those in which it does not.

For WSD schedules, Wen et al. (2024) provide a river-valley landscape interpretation of their dynamical behavior; Schaipp et al. (2025); Li et al. (2025) offer theoretical evidence for the benefits of delayed decay but do not characterize the role of the *decay shape* itself. Meanwhile, existing empirical studies reach seemingly conflicting conclusions: Defazio et al. (2023) and Bergsma et al. (2025) advocate linear decay, whereas Hägele et al. (2024) report superior performance for concave decay profiles such as $1-\sqrt{x}$. In this work, we show that the optimal decay shape is determined by the profile of the forgetting kernel, which governs the rate of noise forgetting. In addition, Luo et al. (2025) numerically solve a variational problem based on a multipower-law model for LLM pre-training and find that the resulting LRS takes a WSD form with a decay shape approximately $(1-x)^{1.5}$. Our analysis provides theoretical support for this empirical observation.

One-pass SGD for kernel regression. The convergence of one-pass SGD for kernel regression—often formulated as high-dimensional feature-space linear regression—has received considerable attention. In particular, Dieuleveut & Bach (2015) and Mücke et al. (2019) showed that averaged SGD attains the minimax-optimal rates $\mathcal{O}(N^{-\frac{s\beta}{s\beta+1}})$ in the easy-task regime and $\mathcal{O}(N^{-s})$ in the hard-task regime. Subsequent work demonstrated that iterate averaging can be replaced by the more practical last iterate when learning-rate schedule is adopted, but the resulting rates typically incur logarithmic factors (Wu et al., 2022a; Lin et al., 2024; Li et al., 2025). In this work, we remove these logarithmic factors by employing power-decay schedules.

B Proof Sketch of Theorems 3.1 and 4.2

B.1 Proof Sketch of Theorem 3.1

Directly solving the variational problem equation 6 is hard due to the non-linearity. To address this, we adopt a decoupled approach:

Step 1: Deriving the optimal profile under a fixed intrinsic time budget. We first fix the total intrinsic time $t(N) = T$ and optimize the schedule profile subject *solely* to this constraint. Noting $t = \Gamma(z) := \int_0^z \eta(u)\,\mathrm{d}u$ and $\varphi(t) = \eta(\Gamma^{-1}(t))$, the intrinsic-time constraint becomes $N = \int_0^N 1\,\mathrm{d}z = \int_0^T \frac{1}{\eta(z)}\,\mathrm{d}\tau = \int_0^T \frac{1}{\varphi(\tau)}\,\mathrm{d}\tau$. For fixed intrinsic time, the signal-learning term is

constant; we only need to minimize noise term:

$$\min_{\varphi:[0,T]\to\mathbb{R}_{\geqslant 0}} \int_0^T \mathcal{K}(T-\tau)\varphi(\tau)\,\mathrm{d}\tau \qquad \text{s.t.} \qquad \int_0^T \frac{1}{\varphi(\tau)}\,\mathrm{d}\tau = N. \tag{14}$$

Applying Cauchy-Schwarz inequality yields

$$\left(\int_0^T \mathcal{K}(T-\tau)\varphi(\tau)\,\mathrm{d}\tau\right)\left(\int_0^T \frac{1}{\varphi(\tau)}\,\mathrm{d}\tau\right) \geqslant \left(\int_0^T \sqrt{\mathcal{K}(T-\tau)}\,\mathrm{d}\tau\right)^2.$$

Consequently, the noise term equation 14 is minimized when the equality condition holds, which yields the optimal intrinsic-time profile:

$$\varphi^*(\tau) = \frac{\int_0^T \sqrt{\mathcal{K}(T-u)}\,\mathrm{d}u}{N\sqrt{\mathcal{K}(T-\tau)}} \propto \frac{1}{\sqrt{\mathcal{K}(T-\tau)}} = (1+T-\tau)^{1-\frac{1}{2\beta}}. \tag{15}$$

This also translates to a power-decay in training steps: $\eta(z) \propto (1-z/N)^{2\beta-1}$.

Step 2: Determining the optimal intrinsic time budget. Substituting equation 15 into equation 14, the noise term under optimal LRS becomes $\int_0^T \mathcal{K}(T-\tau)\varphi(\tau)\,\mathrm{d}\tau = \frac{1}{N}\left(\int_0^T \sqrt{\mathcal{K}(T-\tau)}\,\mathrm{d}\tau\right)^2 \eqsim \frac{T^{\frac{1}{\beta}}}{N}$. Consequently, the total excess risk is given by $\mathcal{E}_{N,T} \eqsim T^{-s} + \frac{T^{\frac{1}{\beta}}}{N}$. Minimizing this risk with respect to T, we obtain the optimal intrinsic time horizon $T^* \eqsim N^{\frac{\beta}{1+s\beta}}$.

Step 3: Incorporating the peak learning-rate constraint. For the unconstrained solution, the implied peak learning rate scales as

$$\varphi^*(0) \eqsim N^{-\frac{1+\beta(s-1)}{s\beta+1}}.$$

This scaling exhibits two distinct regimes. (i) When $s > 1 - 1/\beta$, the exponent is negative, and hence $\varphi^*(0) \to 0$ as $N \to \infty$. In this case, the stability constraint $\eta_{\text{peak}} \leqslant \eta_{\text{stability}}$ is asymptotically inactive. (ii) When $s < 1 - 1/\beta$, the unconstrained peak diverges as $N \to \infty$, violating the stability constraint. Consequently, the peak constraint becomes *active* and must be explicitly enforced.

To handle this regime, we apply the Karush–Kuhn–Tucker (KKT) conditions to the fully constrained variational problem. Using Lagrange multipliers and exploiting the monotonicity of the optimal profile, we show that the resulting optimal LRS exhibits a WSD-like structure.

B.2 Proof Sketch of Theorem 4.2

We first transform the training-step LRS into its intrinsic-time counterpart, denoted as $\varphi(\tau) \eqsim \eta_0\bar{\zeta}(\tau/T)$. Crucially, the fractional structure preserves in intrinsic time. Specifically, if the training-step profile exhibits a power-decay tail, the intrinsic-time profile also possesses a power-decay tail, satisfying $\bar{\zeta}(x) \eqsim (1-x)^{\frac{\gamma}{\gamma+1}}$ for $x \in [\bar{\delta}, 1]$ and some $\bar{\delta} \in (0,1)$. Applying the FSL equation 4, we derive

$$\begin{aligned}\mathcal{F}[\eta_N] &\eqsim T_N^{-s} + \int_0^{T_N} \mathcal{K}(T_N-\tau)\eta_0\bar{\zeta}\left(\frac{\tau}{T_N}\right)\mathrm{d}\tau \\ &\eqsim T_N^{-s} + \frac{\eta_0}{T_N^{1-1/\beta}}\int_0^1 \frac{\bar{\zeta}(x)}{(T_N^{-1}+1-x)^{2-1/\beta}}\,\mathrm{d}x\end{aligned} \tag{16}$$

We decompose the integral into two regions: $[0, \bar{\delta}]$ and $[\bar{\delta}, 1]$. The asymptotic behavior is **dominated by the tail integral** over $[\bar{\delta}, 1]$. We analyze the convergence based on the exponent comparison:

- **Fast decay regime** ($\frac{\gamma}{\gamma+1} > 1 - \frac{1}{\beta}$): In this case, the combined exponent satisfies $\frac{\gamma}{\gamma+1} - (2 - \frac{1}{\beta}) > -1$, ensuring the integral converges absolutely:

$$\int_{\bar{\delta}}^1 \frac{\bar{\zeta}(x)}{(T_N + 1^{-1} - x)^{2-1/\beta}}\,\mathrm{d}x \eqsim \int_{\bar{\delta}}^1 (1-x)^{\frac{\gamma}{\gamma+1}-(2-\frac{1}{\beta})}\,\mathrm{d}x < \infty.$$

- **Slow decay regime** ($\frac{\gamma}{\gamma+1} < 1 - \frac{1}{\beta}$): Here, the integral diverges as $x \to 1$. The term T_N^{-1} acts as a cutoff, yielding:

$$\int_{\bar{\delta}}^{1} \frac{\bar{\zeta}(x)}{(T_N^{-1} + 1 - x)^{2-1/\beta}} \, \mathrm{d}x \eqsim \int_{\bar{\delta}}^{1} (T_N^{-1} + 1 - x)^{\frac{\gamma}{\gamma+1} - (2 - \frac{1}{\beta})} \, \mathrm{d}x \eqsim T_N^{(1-\frac{1}{\beta}) - \frac{\gamma}{\gamma+1}}.$$

Combining the two regimes gives the rate $T_N^{-\min\{1-\frac{1}{\beta}, \frac{\gamma}{\gamma+1}\}}$. At the boundary $\frac{\gamma}{\gamma+1} = 1 - \frac{1}{\beta}$, the tail integral contributes an additional $\log T_N$ factor.

Intuitively, when the LRS decays rapidly, the noise injection diminishes faster than the forgetting kernel can dissipate it; consequently, the bottleneck becomes the decay rate of the forgetting kernel itself. In contrast, when the LRS decays more slowly, the noise accumulation overwhelms the forgetting mechanism, leading to a sub-optimal total noise forgetting.

C Proof for Theorem 3.1 (Optimal LRS under FSL)

In this section, we provide a detailed proof of Theorem 3.1, structured into the following three steps. We define the auxiliary functions:

$$\Phi(T) = (1+T)^{-s}, \quad L(T, t, t') = \mathcal{K}(T-t)(t')^2.$$

The objective functional in training step is then:

$$\begin{aligned} \min_{t \in \mathrm{AC}([0,N]), T} \quad & \widetilde{\mathcal{F}}[t, T] := \Phi(T) + \int_0^N L(T, t(z), t'(z)) \, \mathrm{d}z \\ \text{s.t.} \quad & t(0) = 0, \ t(N) = T, \\ & 0 \leqslant t'(z) \leqslant \eta_{\text{stability}} \quad \text{for a.e. } z \in [0, N]. \end{aligned} \tag{17}$$

Directly solving the variational problem equation 17 is challenging due to the non-linearity of the functional with respect to the LRS function. To address this, we adopt a decoupled approach and decompose the proof into three steps.

C.1 Step 1: Deriving the Optimal Profile under a Fixed Intrinsic Time Budget

We first fix the total intrinsic time $t(N) = T$ and optimize the schedule profile subject *solely* to this constraint (temporarily omitting the peak learning rate constraint). Formally, we apply the change of variables $z = t^{-1}(\tau)$. By the inverse function theorem, the differential transforms as:

$$\mathrm{d}z = \frac{\mathrm{d}\tau}{t'(z)} = \frac{\mathrm{d}\tau}{\eta(z)} = \frac{\mathrm{d}\tau}{\varphi(\tau)}.$$

Then, the requirement imposes the following integral constraint:

$$N = \int_0^N 1 \, \mathrm{d}z = \int_0^T \frac{1}{\varphi(\tau)} \, \mathrm{d}\tau,$$

which leads to the following variational problem of noise term:

$$\begin{aligned} \min_{\varphi : [0,T] \to \mathbb{R}_{\geqslant 0}} \quad & \int_0^T \mathcal{K}(T-\tau)\varphi(\tau) \, \mathrm{d}\tau \\ \text{s.t.} \quad & \int_0^T \frac{1}{\varphi(\tau)} \, \mathrm{d}\tau = N. \end{aligned} \tag{18}$$

Applying the Cauchy-Schwarz inequality yields

$$\left(\int_0^T \mathcal{K}(T-\tau)\varphi(\tau) \, \mathrm{d}\tau \right) \left(\int_0^T \frac{1}{\varphi(\tau)} \, \mathrm{d}\tau \right) \geqslant \left(\int_0^T \sqrt{\mathcal{K}(T-\tau)} \, \mathrm{d}\tau \right)^2.$$

Consequently, the noise term is minimized when the equality condition holds, which yields the optimal intrinsic profile:

$$\begin{aligned}\varphi(\tau) &= \frac{\int_0^T \sqrt{\mathcal{K}(T-u)}\,\mathrm{d}u}{N\sqrt{\mathcal{K}(T-\tau)}} \\ &= a_{N,T}(1+T-\tau)^{1-\frac{1}{2\beta}}\end{aligned} \tag{19}$$

where $a_{N,T} := \frac{2\beta}{N}((1+T)^{\frac{1}{2\beta}} - 1)$. We have used $\mathcal{K}(t) = (1+t)^{1/\beta-2}$.

Now, we proceed to derive the LRS in terms of training steps z. Recall that $\frac{\mathrm{d}\tau}{\mathrm{d}z} = \varphi(\tau)$. Substituting the profile from equation 19, we have

$$\frac{\mathrm{d}\tau}{\mathrm{d}z} = \varphi(\tau) = a_{N,T}(1+T-\tau)^{1-\frac{1}{2\beta}}, \tag{20}$$

with boundary condition $\tau = 0$ when $z = 0$.

Solving the ODE equation 20, we derive

$$z = a_{N,T}^{-1} 2\beta(1+T)^{\frac{1}{2\beta}} - a_{N,T}^{-1} 2\beta(1+T-\tau)^{\frac{1}{2\beta}},$$

which implies the following intrinsic time function:

$$t(z) = 1 + T - \left((1+T)^{\frac{1}{2\beta}} - \frac{a_{N,T}}{2\beta} z\right)^{2\beta}.$$

Differentiating with respect to z gives the training-step LRS:

$$\eta(z) = t'(z) = \frac{2\beta}{N}\left((1+T)^{\frac{1}{2\beta}} - 1\right)^{2\beta}\left(1 + \frac{1}{(1+T)^{\frac{1}{2\beta}} - 1} - \frac{z}{N}\right)^{2\beta-1}.$$

Remark C.1. We also provide an alternative proof for this step based on the variational method. Please refer to Appendix F for details.

C.2 Step 2: Determining the Optimal Intrinsic Time Budget

Substituting the optimal intrinsic profile equation 19 into the noise term equation 18, we derive the accumulated variance:

$$\int_0^T \mathcal{K}(T-\tau)\varphi(\tau)\,\mathrm{d}\tau = \frac{1}{N}\left(\int_0^T \sqrt{\mathcal{K}(T-\tau)}\,\mathrm{d}\tau\right)^2 \eqsim \frac{T^{\frac{1}{\beta}}}{N}.$$

Consequently, the total excess risk is given by

$$\mathcal{E}_{N,T} \eqsim T^{-s} + \frac{T^{\frac{1}{\beta}}}{N}.$$

Minimizing this risk with respect to T, we obtain the optimal intrinsic time horizon $T_{\text{opt}} \eqsim N^{\frac{\beta}{1+s\beta}}$.

C.3 Step 3: Incorporating the Peak Learning Rate Constraint

Based on the unconstrained solution, the implied peak learning rate scales as:

$$\eta(0) \eqsim N^{-\frac{1+\beta(s-1)}{s\beta+1}}.$$

This scaling behavior reveals two distinct regimes:

- **Easy-task regime ($s > 1 - 1/\beta$):** In this case, the exponent is negative, meaning $\eta(0) \to 0$ as $N \to \infty$. Consequently, the physical constraint $\eta_{\text{peak}} \leqslant \eta_{\text{stability}}$ is naturally satisfied (inactive) for sufficiently large N.

- **Hard-task regime ($s < 1 - 1/\beta$):** Conversely, the optimal unconstrained peak diverges as $N \to \infty$. This violates the stability constraint $\eta(z) \leqslant \eta_{\text{stability}}$. Therefore, the peak constraint becomes **active** and must be explicitly incorporated into the optimization.

To address the hard-task regime, we apply the Karush-Kuhn-Tucker (KKT) conditions to solve the fully constrained variational problem.

To explicitly incorporate the maximal learning rate constraint $t'(z) \leqslant \eta_{\text{stability}}$, we formulate the Lagrangian in the training step domain. We introduce the Lagrange multipliers:

$$\lambda(z) \geqslant 0 \quad (\text{dual for } t'(z) \leqslant \eta_{\text{stability}}), \qquad \mu \in \mathbb{R} \quad (\text{dual for } t(N) - T = 0).$$

The generalized Lagrangian functional $\mathcal{J}$ is defined as:

$$\mathcal{J}[t, T, \lambda, \mu] \;=\; \Phi(T) \;+\; \int_0^N \Big[L(T, t, t') + \lambda(z)\big(t'(z) - \eta_{\text{stability}}\big) \Big] \,\mathrm{d}z \;+\; \mu\big(t(N) - T\big).$$

Theorem C.2 (KKT condition)**.** *A feasible pair $(t^\star, T^\star)$ is optimal only if there exist multipliers $\lambda^\star \in L^\infty([0, N])$ and $\mu^\star \in \mathbb{R}$ such that*

$$\begin{aligned}
&\textit{(1) Euler–Lagrange stationarity:} && \partial_t L - \frac{\mathrm{d}}{\mathrm{d}z}\Big(\partial_{t'} L + \lambda^\star\Big) && = 0, && z \in (0, N);\\
&\textit{(2) Boundary stationarity:} && \Big[\partial_{t'} L + \lambda^\star\Big]_{z=N} + \mu^\star && = 0;\\
&\textit{(3) Scalar stationarity (w.r.t. } T\textit{):} && \Phi'(T^\star) + \int_0^N \partial_T L \,\mathrm{d}z - \mu^\star && = 0;\\
&\textit{(4) Primal feasibility:} && t^\star(0) = 0, \quad t^\star(N) = T^\star, \quad t^{\star\prime}(z) \leqslant 1;\\
&\textit{(5) Dual feasibility:} && \lambda^\star(z) \geqslant 0;\\
&\textit{(6) Complementary slackness:} && \lambda^\star(z)\,\big(t^{\star\prime}(z) - \eta_{\text{stability}}\big) = 0; && && z \in [0, N].
\end{aligned}$$

If $t^{\star\prime}(z) < \eta_{\text{stability}}$ at some point, condition (6) forces $\lambda^\star(z) = 0$; where the derivative saturates the bound ($t^{\star\prime} = \eta_{\text{stability}}$), $\lambda^\star$ may be positive.

Direct analysis of the above KKT condition is still complicated. A key observation that make the analysis easier is the following observation:

Proposition C.3 (Monotonicity)**.** *Let t^* be the solution of equation 6. Then, $z \mapsto t'(z)$ must be non-increasing.*

Proof. To prove this, we only need to show for any fixed T, the corresponding minimizer t_T^* is decreasing. Hence, consider

$$\begin{aligned}
G_T(t) &\coloneqq \int_0^N \big(1 + T - t(z)\big)^{-(2-1/\beta)} \big(t'(z)\big)^2 \,\mathrm{d}z\\
&= \int_0^T (1 + T - \tau)^{-(2-\frac{1}{\beta})} \varphi(\tau) \,\mathrm{d}\tau\\
&=: \int_0^T w(\tau)\varphi(\tau) \,\mathrm{d}\tau,
\end{aligned}$$

where the weight function $w(\cdot)$ is increasing.

A simple "bubble-sort" argument suffice to show that the optimal φ is non-increasing. Suppose a feasible φ is not non-increasing. There must exist $\tau_1 < \tau_2$ such that $\varphi(\tau_1) < \varphi(\tau_2)$. Then, we can construct $\tilde{\varphi}$ by swapping the values on small intervals around τ_1, τ_2.

- The constraint $\int_0^T \frac{1}{\tilde{\varphi}(\tau)} \,\mathrm{d}\tau = N$ and boundedness constraint are unaffected by the swaps.

- The change of integral:

$$\Delta = [\varphi(\tau_2) - \varphi(\tau_1)](w(\tau_1) - w(\tau_2)) < 0.$$

Thus the swap strictly lowers the objective. Repeating finitely many swaps (or taking a limit) yields a decreasing function with no larger cost, contradicting optimality. Hence a minimiser must be decreasing. □

Remark C.4. The above bubble-sort argument essentially adopts the (anti)-Hardy-Littlewood inequality.

Theorem C.5 (Stable-decay shape)**.** *If $s < 1 - 1/\beta$, the optimal LRS must be stable-decay shape:*

$$t'_{a,N_1}(z) = \begin{cases} \eta_{\text{stability}} & \text{if } 0 \leqslant z \leqslant N_1 \\ a\eta_{\text{stability}} \left(1 + o_N(1) - \frac{z-N_1}{N-N_1}\right)^{2\beta-1} & \text{if } N_1 < z \leqslant N, \end{cases} \tag{21}$$

where $a \in [0, 1]$ and $N_1 \in [0, N]$.

Proof. Let $S_i = \{z \in [0, N] : t'(z) < \eta_{\text{stability}}\}, S_b = \{z \in [0, N] : t'(z) = \eta_{\text{stability}}\}$. By the monotonicity, either $S_i = [0, N]$ or there exists a N_1 such that $S_i = (N_1, N)$ and $S_b = [0, N_1]$. For $z \in S_i$, we must have $\lambda(z) \equiv 0$ by the complementary slackness. Hence, $\lambda'(z) = 0$ for $z \in S_i$ and consequently, the LRS satisfy the Euler-Lagrange equation for $z \in S_i$:

$$\partial_t L - \frac{\mathrm{d}}{\mathrm{d}z}[\partial_{t'} L] = 0.$$

Following the derivation in Step 1, the solution must take the form:

$$t'(z) = a\eta_{\text{stability}} \left(1 + o_N(1) - \frac{z - N_1}{N - N_1}\right)^{2\beta-1}.$$

Hence, we complete the proof. □

Theorem C.6. *Let $r = (N - N_1)/N$, $\mathcal{Q}_N(a, r) = \widetilde{\mathcal{F}}[t_{a,(1-r)N}, T]$ and*

$$(a_N^*, r_N^*) = \arg\min_{a,r\in[0,1]} \mathcal{Q}_N(a, r).$$

If $s < 1 - 1/\beta$ and N is suffiently large, we have $a_N^* = 1$ and

$$r_N^* \eqsim N^{-\gamma}, \quad \gamma = \frac{(1 - \frac{1}{\beta}) - s}{1 - \frac{1}{\beta} + 1}.$$

Proof. Noting that $\gamma > 0$ for the hard regime, the optimal decay duration scales sublinearly with the total number of training steps. As a result, the optimal LRS exhibits a stable–decay shape, with the decay phase occupying only a tiny fraction of the total training steps.

Simplifying the objective function. By Theorem C.5, we know that

$$\min_t \widetilde{\mathcal{F}}[t, T] = \widetilde{\mathcal{F}}[t_{a,(1-r)N}, T].$$

With this LRS equation 21, the total intrinsic time is

$$T = \eta_{\text{stability}} \left(N_1 + \frac{a(N - N_1)}{2\beta}\right).$$

Let $T_1 = N_1$, denoting the intrinsic time of the stable phase. Then, the LRS with respect to intrinsic time can be expressed as

$$\varphi(\tau) = \begin{cases} \eta_{\text{stability}} & \text{if } 0 \leqslant \tau \leqslant T_1 \\ a\eta_{\text{stability}} \left(1 - \frac{\tau - T_1}{1 + T - T_1}\right)^{1-\frac{1}{2\beta}} & \text{if } T_1 \leqslant \tau \leqslant T. \end{cases}$$

Here, we assume $\eta_{\text{stability}} = 1$ for simplicity. The noise term is given by

$$\begin{aligned}
&\int_0^N L(T,t,t')\,\mathrm{d}z \\
&= \int_0^T (1+T-t)^{-(2-\frac{1}{\beta})}\varphi(\tau)\,\mathrm{d}\tau \\
&= \int_0^{T_1} (1+T-\tau)^{-(2-\frac{1}{\beta})}\,\mathrm{d}\tau + a\int_{T_1}^T (1+T-\tau)^{-(2-\frac{1}{\beta})}\left(1-\frac{\tau-T_1}{1+T-T_1}\right)^{1-\frac{1}{2\beta}}\mathrm{d}\tau \\
&= \frac{1}{1-\frac{1}{\beta}}\left[\frac{1}{(1+T-T_1)^{1-\frac{1}{\beta}}} - \frac{1}{(1+T)^{1-\frac{1}{\beta}}}\right] + \frac{a}{(1+T-T_1)^{1-\frac{1}{\beta}}}\int_0^1 (1-u)^{-1+\frac{1}{2\beta}}\,\mathrm{d}u \\
&= \frac{1}{1-\frac{1}{\beta}}\left[\frac{1}{(1+T-T_1)^{1-\frac{1}{\beta}}} - \frac{1}{(1+T)^{1-\frac{1}{\beta}}}\right] + a\frac{2\beta}{(1+T-T_1)^{1-\frac{1}{\beta}}}.
\end{aligned}$$

Hence, the total objective becomes

$$\begin{aligned}
G(N_1,a) \coloneqq &\left(N_1 + \frac{a(N-N_1)}{2\beta}\right)^{-s} \\
&+ \frac{\beta}{\beta-1}\left([a(N-N_1)/(2\beta)]^{-(1-\frac{1}{\beta})} - \left(N_1+\frac{a(N-N_1)}{2\beta}\right)^{-(1-\frac{1}{\beta})}\right) \\
&+ 2\beta a\left(\frac{a}{2\beta}(N-N_1)\right)^{-(1-\frac{1}{\beta})}.
\end{aligned}$$

Let $a' = a/2\beta$, $N_1 = (1-r)N, \alpha = 1-\frac{1}{\beta}$. Then, $\widetilde{G}(r,a') \coloneqq G((1-r)N, 2\beta a')$ is given by

$$\widetilde{G}(r,a') = N^{-s}(1-r+a'r)^{-s} + N^{-\alpha}\left[(1+4\beta^2 a')(a'r)^{-\alpha} - (1-r+a'r)^{-\alpha}\right]$$

The optimal peak learning rate a_N^*. Let $B = 1-r+a'r$ (so $B > a'r$). Then,

$$\frac{\partial\widetilde{G}}{\partial a'} = r\left[-sN^{-s}B^{-s-1} + \alpha N^{-\alpha}\left(B^{-\alpha-1} - (a'r)^{-\alpha-1} + 4\beta^2 a'^{\alpha} r^{-\alpha-1}\right)\right].$$

When $s < \alpha$ and N is sufficiently large, we have $\frac{\partial\widetilde{G}}{\partial a'} < 0$. Hence, $a_N^* = 1$.

The optimal decay duration. Let $A(r) = \left(1-r+a'r\right)^{-s}, B(r) = (a'r)^{-\alpha} - \left(1-r+a'r\right)^{-\alpha}$. When taking $a = a_N^* = 1$, we have

$$\widetilde{G}(r) := \widetilde{G}(r, a_N^*) \;=\; N^{-s}\left(1-r+a'r\right)^{-s} + N^{-\alpha}\Big[(a'r)^{-\alpha} - \left(1-r+a'r\right)^{-\alpha}\Big] = N^{-s}A(r) + N^{-\alpha}B(r).$$

We can obtain that $r_N^* \ll 1$ and now we need to track the explicit scaling. When $r \ll 1$, we have

$$A'(r) = s(1-a') + o_N(1), \qquad B'(r) = -(a')^{-\alpha}\alpha r^{-\alpha-1} + O_N(1).$$

Then, $\widetilde{G}'(r) = N^{-s}A'(r) + N^{-\alpha}B'(r) = 0$. This up to constants leads to

$$N^{-s} - N^{-\alpha}(r_N^*)^{-\alpha-1} = 0 \implies r_N^* \eqsim N^{-\frac{\alpha-s}{\alpha+1}}.$$

This completes the proof. □

D Proofs for Section 4 (Shape-Fixed Optimality and Capacity Saturation)

D.1 The Intrinsic-Time Profile Function

Throughout this section, let η_N be a fractional LRS as is given in Definition 4.1:

$$\eta_N(z) = \eta_0\,\zeta\Big(\frac{z}{N}\Big), \qquad z \in [0,N],$$

where $\eta_0 > 0$ and $\zeta : [0,1] \to [0,1]$ satisfies $\zeta(0) = 1$. Define

$$\rho(x) \coloneqq \int_0^x \zeta(u)\ \mathrm{d}u, \qquad \rho_1 \coloneqq \rho(1) \in (0,1].$$

Since $\zeta(u) \geqslant 0$, the map $x \mapsto \rho(x)$ is non-decreasing on $[0,1]$ and admits a generalized inverse ρ^{-1} on $[0, \rho_1]$.

The intrinsic time as a function of training steps is

$$t(z) \coloneqq \int_0^z \eta_N(u)\ \mathrm{d}u = \eta_0 N\, \rho(z/N), \qquad T_N \coloneqq t(N) = \eta_0 N \rho_1.$$

Consequently, $t(\cdot)$ is non-decreasing on $[0, N]$. Let $z(t)$ denote a generalized inverse on $[0, T_N]$.

Learning rate in intrinsic time. Following the FSL notation in equation 4, define the intrinsic-time learning rate

$$\varphi(t) \coloneqq \eta_N(z(t)), \qquad t \in [0, T_N].$$

We introduce the intrinsic-time profile function $\bar{\zeta} : [0,1] \to [0,1]$ by

$$\bar{\zeta}(y) \coloneqq \zeta\Big(\rho^{-1}(y\rho_1)\Big), \qquad y \in [0,1]. \tag{22}$$

Then for any $t \in [0, T_N]$ with $y = t/T_N$, let $x = z(t)/N$. Then we have $y = \rho(x)/\rho_1$. Hence $x = \rho^{-1}(y\rho_1)$ and

$$\varphi(t) = \eta_0\, \bar{\zeta}\left(\frac{t}{T_N}\right), \qquad t \in [0, T_N].$$

Importantly, $\bar{\zeta}$ depends only on the fixed training-step profile function ζ and is independent of N.

Tail exponent under intrinsic-time The power-decay tail in Definition 4.1 is stated in training step: there exist constants $\gamma > 0$ and $\delta \in (0,1)$ such that

$$\zeta(x) \eqsim (1-x)^{\gamma}, \qquad x \in [\delta, 1].$$

The next lemma identifies the corresponding tail behavior of $\bar{\zeta}$.

Lemma D.1 (Intrinsic-time tail exponent)**.** *Assume $\zeta(x) \eqsim (1-x)^{\gamma}$ for all $x \in [\delta, 1]$ with some $\gamma > 0$. Let*

$$\bar{\delta} \coloneqq \frac{\rho(\delta)}{\rho_1} \in (0,1).$$

Then for all $y \in [\bar{\delta}, 1]$,

$$\bar{\zeta}(y) \eqsim (1-y)^{\frac{\gamma}{\gamma+1}}.$$

Proof. Fix $y \in [\bar{\delta}, 1]$ and define $x \coloneqq \rho^{-1}(y\rho_1) \in [\delta, 1]$. Then $\bar{\zeta}(y) = \zeta(x)$ by definition.

Let $A(x) \coloneqq \int_x^1 \zeta(u)\ \mathrm{d}u = \rho_1 - \rho(x)$. Since $\zeta(u) \eqsim (1-u)^{\gamma}$ on $[\delta, 1]$, integrating yields

$$A(x) \eqsim (1-x)^{\gamma+1}, \qquad x \in [\delta, 1].$$

Moreover,

$$1 - y = 1 - \frac{\rho(x)}{\rho_1} = \frac{\rho_1 - \rho(x)}{\rho_1} = \frac{A(x)}{\rho_1} \eqsim (1-x)^{\gamma+1}.$$

Hence, for $y \in [\bar{\delta}, 1]$,

$$1 - x \eqsim (1-y)^{\frac{1}{\gamma+1}}.$$

Finally, using $\zeta(x) \eqsim (1-x)^{\gamma}$ on $[\delta, 1]$ and $\bar{\zeta}(y) = \zeta(x)$, we obtain

$$\bar{\zeta}(y) = \zeta(x) \eqsim (1-x)^{\gamma} \eqsim \Big((1-y)^{\frac{1}{\gamma+1}}\Big)^{\gamma} = (1-y)^{\frac{\gamma}{\gamma+1}},$$

which holds for all $y \in [\bar{\delta}, 1]$. □

D.2 Proof of Theorem 4.2

Proof. Recall $\mathcal{K}(u) = (1+u)^{-(2-1/\beta)}$ and set

$$p \coloneqq 2 - \frac{1}{\beta} \in (1,2), \qquad q \coloneqq \frac{\gamma}{\gamma+1} \in (0,1).$$

By the intrinsic-time representation in Appendix D.1, the learning-rate function in intrinsic time satisfies

$$\varphi(t) = \eta_0 \, \bar{\zeta}\left(\frac{t}{T_N}\right), \qquad t \in [0, T_N],$$

and by Lemma D.1 there exists $\bar{\delta} \in (0,1)$ such that

$$\bar{\zeta}(x) \eqsim (1-x)^q, \qquad x \in [\bar{\delta}, 1]. \tag{23}$$

Applying the FSL equation 4 at the final intrinsic time T_N gives

$$\mathcal{F}[\eta_N] \eqsim (1+T_N)^{-s} + \int_0^{T_N} \mathcal{K}(T_N - \tau)\, \varphi(\tau)\, \mathrm{d}\tau. \tag{24}$$

For $T_N \gtrsim 1$, we have $(1+T_N)^{-s} \eqsim T_N^{-s}$, so it remains to estimate the noise term

$$\mathcal{N}(T_N) \coloneqq \int_0^{T_N} (1 + T_N - \tau)^{-p}\, \varphi(\tau)\, \mathrm{d}\tau.$$

Substituting $\varphi(\tau) = \eta_0 \, \bar{\zeta}(\tau/T_N)$ and changing variables $\tau = T_N x$ yield

$$\mathcal{N}(T_N) = \eta_0 T_N \int_0^1 \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x. \tag{25}$$

We split the integral into an early region and a tail region. Decompose

$$\int_0^1 \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x = \int_0^{\delta} \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x + \int_{\delta}^1 \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x.$$

Define

$$I_{\text{early}} \coloneqq \int_0^{\delta} \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x, \qquad I_{\text{tail}} \coloneqq \int_{\delta}^1 \big(1 + T_N(1-x)\big)^{-p}\, \bar{\zeta}(x)\, \mathrm{d}x.$$

(i) Early region. For $x \in [0, \delta]$, we have $1 - x \geqslant 1 - \delta$, hence $\big(1 + T_N(1-x)\big)^{-p} \eqsim T_N^{-p}$ uniformly for $T_N \gtrsim 1$. Therefore,

$$I_{\text{early}} \eqsim T_N^{-p} \int_0^{\delta} \bar{\zeta}(x)\, \mathrm{d}x \eqsim T_N^{-p},$$

where $\int_0^{\delta} \bar{\zeta}(x)\, \mathrm{d}x$ is a fixed positive constant absorbed into $\eqsim$. Therefore,

$$\eta_0 T_N I_{\text{early}} \eqsim \eta_0 T_N \cdot T_N^{-p} = \eta_0 T_N^{1-p} = \eta_0\, T_N^{-(1-1/\beta)}. \tag{26}$$

(ii) Tail region and the logarithmic boundary. For $x \in [\delta, 1]$, equation 23 gives $\bar{\zeta}(x) \eqsim (1-x)^q$, hence

$$I_{\text{tail}} \eqsim \int_{\delta}^1 \big(1 + T_N(1-x)\big)^{-p} (1-x)^q\, \mathrm{d}x.$$

Let $u = T_N(1-x)$, so that $\mathrm{d}x = -\,\mathrm{d}u/T_N$ and $(1-x) = u/T_N$. Then

$$I_{\text{tail}} \eqsim T_N^{-(q+1)} \int_0^{T_N(1-\delta)} (1+u)^{-p} u^q\, \mathrm{d}u.$$

This gives

$$\eta_0 T_N I_{\text{tail}} \eqsim \eta_0\, T_N^{-q} \int_0^{T_N(1-\delta)} (1+u)^{-p} u^q\, \mathrm{d}u. \tag{27}$$

The integral in equation 27 is governed by the behavior of $(1+u)^{-p} u^q \sim u^{q-p}$ as $u \to \infty$. We distinguish three cases.

- Case 1: $q - p < -1$ (equivalently $\beta < \gamma + 1$). Then $\int_0^{T_N(1-\delta)} (1+u)^{-p} u^q \, \mathrm{d}u \eqsim T_N^{q+1-p}$, so
$$\eta_0 T_N I_{\text{tail}} \eqsim \eta_0 \, T_N^{-q} \cdot T_N^{q+1-p} = \eta_0 T_N^{1-p} = \eta_0 \, T_N^{-(1-1/\beta)}.$$
- Case 2: $q - p > -1$ (equivalently $\beta > \gamma + 1$). Then the integral in equation 27 converges to a positive constant, and hence
$$\eta_0 T_N I_{\text{tail}} \eqsim \eta_0 \, T_N^{-q} = \eta_0 \, T_N^{-(1-1/(\gamma+1))}.$$
- Case 3 (boundary): $q - p = -1$ (equivalently $\beta = \gamma + 1$). In this case, $(1+u)^{-p} u^q \sim u^{-1}$ as $u \to \infty$, so
$$\int_0^{T_N(1-\delta)} (1+u)^{-p} u^q \, \mathrm{d}u \eqsim \log T_N.$$
Therefore,
$$\eta_0 T_N I_{\text{tail}} \eqsim \eta_0 \, T_N^{-q} \log T_N = \eta_0 \, T_N^{-(1-1/(\gamma+1))} \log T_N. \tag{28}$$

Combining the early contribution equation 26 with the tail analysis above, we obtain
$$\mathcal{N}(T_N) \eqsim \eta_0 \, T_N^{-(1-\frac{1}{\alpha})} \left(\log T_N\right)^{\mathbf{1}\{\beta=\gamma+1\}}, \qquad \alpha := \min\{\beta, \gamma+1\}.$$
Substituting this estimate into equation 24 and using $(1+T_N)^{-s} \eqsim T_N^{-s}$ proves the scaling law
$$\mathcal{F}[\eta_N] \eqsim T_N^{-s} + \eta_0 \, T_N^{-(1-\frac{1}{\alpha})} \left(\log T_N\right)^{\mathbf{1}\{\beta=\gamma+1\}}.$$

□

D.3 Proof of Theorem 4.3

Proof. Recall
$$T_N = \eta_0 N \int_0^1 \zeta(x) \, \mathrm{d}x \eqsim \eta_0 N, \qquad \alpha = \min\{\beta, \gamma+1\}.$$
By Theorem 4.2, the final-step loss satisfies
$$\mathcal{F}[\eta_N] \eqsim T_N^{-s} + \eta_0 \, T_N^{-(1-\frac{1}{\alpha})} \left(\log T_N\right)^{\mathbf{1}\{\beta=\gamma+1\}}. \tag{29}$$
Using $T_N \eqsim \eta_0 N$, we rewrite the right-hand side as a function of η_0:
$$\mathcal{F}[\eta_N] \eqsim (\eta_0 N)^{-s} + \eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})} \big(\log(\eta_0 N)\big)^{\mathbf{1}\{\beta=\gamma+1\}}. \tag{30}$$
We minimize equation 30 over η_0.

(i) Hard-task regime: $s < 1 - \frac{1}{\alpha}$. We minimize equation 30 over the admissible range $0 < \eta_0 \leqslant \eta_{\max}$, where $\eta_{\max} = \Theta(1)$ is the stability upper bound.

We first show that in the hard regime, the noise term is uniformly negligible. Since $0 < \eta_0 \leqslant \eta_{\max} = \Theta(1)$, we have $\eta_0^{1/\alpha} \leqslant C$ and $\log(\eta_0 N) \lesssim \log N$. Therefore, uniformly over $\eta_0 \in (0, \eta_{\max}]$,
$$\eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})} \big(\log(\eta_0 N)\big)^{\mathbf{1}\{\beta=\gamma+1\}} \lesssim N^{-(1-\frac{1}{\alpha})} (\log N)^{\mathbf{1}\{\beta=\gamma+1\}}.$$
Because $s < 1 - \frac{1}{\alpha}$, the power gap $(1 - \frac{1}{\alpha}) - s > 0$, hence $N^{-(1-\frac{1}{\alpha})} (\log N)^{\mathbf{1}\{\beta=\gamma+1\}} = o(N^{-s})$. Consequently,
$$\mathcal{F}[\eta_N] = (\eta_0 N)^{-s} \, (1 + o(1)) \qquad \text{uniformly for } \eta_0 \in (0, \eta_{\max}].$$
Since the leading term $(\eta_0 N)^{-s}$ is strictly decreasing in η_0, the minimum over $(0, \eta_{\max}]$ is attained at the largest admissible value:
$$\eta_0^* = \eta_{\max} \eqsim 1.$$
At this choice, $T_N \eqsim N$ and thus
$$\mathcal{E}_N^* = \min_{\eta_0 \in (0, \eta_{\max}]} \mathcal{F}[\eta_N] \eqsim N^{-s}.$$
This proves the hard-task statement, including the boundary case $\beta = \gamma + 1$, where the extra $\log(\eta_0 N)$ factor remains $o(N^{(1-\frac{1}{\alpha})-s})$.

(ii) Easy-task regime: $s \geqslant 1 - \frac{1}{\alpha}$**.** In this regime, the optimal choice balances the two terms in equation 30. We consider two cases.

- Case 1: $\beta \neq \gamma + 1$. Then the logarithmic factor is absent and equation 30 becomes

$$\mathcal{F}[\eta_N] \eqsim (\eta_0 N)^{-s} + \eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})}.$$

Balancing the two terms gives

$$\eta_0^{-s} N^{-s} \eqsim \eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})},$$

i.e.

$$\eta_0^{s+1/\alpha} \eqsim N^{-(s-1+1/\alpha)}.$$

Hence

$$\eta_0^* \eqsim N^{-\frac{s-1+1/\alpha}{s+1/\alpha}}. \tag{31}$$

Substituting into either term yields

$$\mathcal{E}_N^* \eqsim (\eta_0^* N)^{-s} \eqsim N^{-\frac{s\alpha}{s\alpha+1}}.$$

- Case 2 (boundary): $\beta = \gamma + 1$. Then $\alpha = \beta$ and the second term in equation 30 carries the logarithmic factor:

$$\mathcal{F}[\eta_N] \eqsim (\eta_0 N)^{-s} + \eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})} \log(\eta_0 N).$$

In the easy-task regime, the minimizer is characterized by balancing the two terms. Balancing gives

$$\eta_0^{-s} N^{-s} \eqsim \eta_0^{1/\alpha} N^{-(1-\frac{1}{\alpha})} \log(\eta_0 N),$$

equivalently,

$$\eta_0^{s+1/\alpha} \log(\eta_0 N) \eqsim N^{-(s-1+1/\alpha)}. \tag{32}$$

Let

$$a \coloneqq \frac{s-1+1/\alpha}{s+1/\alpha} \in [0, 1), \qquad b \coloneqq \frac{1}{s+1/\alpha}.$$

Then equation 32 is equivalent to

$$\eta_0 \eqsim N^{-a} \big(\log(\eta_0 N)\big)^{-b}. \tag{33}$$

Multiplying by N yields

$$\eta_0 N \eqsim N^{1-a} \big(\log(\eta_0 N)\big)^{-b}.$$

Since $1 - a = \frac{1}{s+1/\alpha} > 0$, the right-hand side diverges, hence $\eta_0 N \to \infty$ and $\log(\eta_0 N)$ is well-defined for large N. Taking logarithms in this equation gives

$$\log(\eta_0 N) = (1-a)\log N - b \log\big(\log(\eta_0 N)\big) + O(1) = (1-a)\log N + O(\log\log N).$$

In particular, $\log(\eta_0 N) \eqsim \log N$. Substituting this back into equation 33 yields

$$\eta_0^* \eqsim N^{-a} (\log N)^{-b} = N^{-\frac{s-1+1/\alpha}{s+1/\alpha}} (\log N)^{-\frac{1}{s+1/\alpha}}. \tag{34}$$

Substituting into the signal term gives

$$\mathcal{E}_N^* \eqsim (\eta_0^* N)^{-s} \eqsim N^{-\frac{s\alpha}{s\alpha+1}} (\log N)^{\frac{s\alpha}{s\alpha+1}}.$$

□

E Proofs for Section 5 (Discrete-Time SGD and Kernel Regression)

E.1 Preliminaries for Kernel Methods

In this section we review the kernel regression setting as in Section 5. Given a feature map $\phi : \mathcal{X} \to \mathbb{H}$, we define the kernel

$$K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathbb{H}}.$$

Define the sampling operator $S : \mathbb{H} \to L^2(\mathcal{D}_{\mathcal{X}})$ by

$$(S\boldsymbol{\theta})(\mathbf{x}) \coloneqq \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathbb{H}}.$$

Under $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[K(\mathbf{x}, \mathbf{x})] < \infty$, the operator S is well-defined and bounded. We introduce two self-adjoint positive operators:

$$\mathcal{T} \coloneqq S^*S : \mathbb{H} \to \mathbb{H}, \qquad \mathcal{I} \coloneqq SS^* : L^2(\mathcal{D}_{\mathcal{X}}) \to L^2(\mathcal{D}_{\mathcal{X}}),$$

where $\mathcal{T}$ is the covariance operator,

$$\mathcal{T}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\big[\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathbb{H}} \, \boldsymbol{\phi}(\mathbf{x})\big],$$

and $\mathcal{I}$ is the kernel integral operator given by

$$(\mathcal{I}g)(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{x}') \, g(\mathbf{x}') \, \mathrm{d}\mathcal{D}_{\mathcal{X}}(\mathbf{x}').$$

Spectral systems and the bridge between L^2 and $\mathbb{H}$. The nonzero eigenvalues of $\mathcal{T}$ and $\mathcal{I}$ coincide. Let $\{(\lambda_j, \mathbf{e}_j)\}_{j\geqslant 1}$ be the spectral system of $\mathcal{I}$ in $L^2(\mathcal{D}_{\mathcal{X}})$, so that

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j\geqslant 1} \lambda_j \mathbf{e}_j(\mathbf{x}) \mathbf{e}_j(\mathbf{x}').$$

with convergence in $L^2(\mathcal{D}_{\mathcal{X}} \times \mathcal{D}_{\mathcal{X}})$. Let $\{\mathbf{v}_j\}_{j\geqslant 1} \subset \mathbb{H}$ be eigenvectors of $\mathcal{T}$ such that

$$\mathcal{T}\mathbf{v}_j = \lambda_j \mathbf{v}_j, \qquad \langle \mathbf{v}_i, \mathbf{v}_j \rangle_{\mathbb{H}} = \delta_{ij}.$$

Then the two bases are linked by the standard relation

$$\mathbf{e}_j(\mathbf{x}) = \lambda_j^{-1/2} \, \langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathbb{H}}, \tag{35}$$

where the equality holds in $L^2(\mathcal{D}_{\mathcal{X}})$.

Capacity and source conditions. We recall the capacity and source condition in Assumption 5.1 and 5.2:

- *(Capacity condition)* There exists $\beta > 1$ such that $\lambda_j \lesssim j^{-\beta}$.
- *(Source condition)* There exists $s > 0$ and coefficients $\{a_j\}_{j\geqslant 1}$ with $\sum_{j\geqslant 1} a_j^2 \leqslant 1$ such that

$$f^*(\mathbf{x}) = \sum_{j\geqslant 1} a_j \, \lambda_j^{\frac{s}{2}} \, \mathbf{e}_j(\mathbf{x}).$$

We note that the source condition can be equivalently stated in the space $\mathbb{H}$. That is,

$$\boldsymbol{\theta}^* = \sum_{j\geqslant 1} a_j \, \lambda_j^{\frac{s-1}{2}} \mathbf{v}_j = \sum_{j\geqslant 1} \theta_j^* \mathbf{v}_j.$$

where the coefficient $\theta_j^* = a_j \, \lambda_j^{\frac{s-1}{2}}$. Note that when $s < 1$, the sequence $\{\theta_j^*\}$ may not be square-summable, so the corresponding $\boldsymbol{\theta}^*$ may not belong to $\mathbb{H}$ as a vector; however, each coefficient θ_j^* is well-defined, and this is sufficient for our coordinate-wise SGD analysis below. We also assume w.l.o.g. that f^* lies in the closure of $\mathrm{span}\{\mathbf{e}_j\}$ (otherwise its orthogonal component is not learnable by any $f_{\boldsymbol{\theta}}$ and only contributes an irreducible constant error).

E.2 Analysis of One-Pass SGD

We consider the standard kernel regression model

$$y = f^*(\mathbf{x}) + \epsilon, \qquad \epsilon \mid \mathbf{x} \sim \mathcal{N}(0, \sigma^2),$$

and let $\{(\mathbf{x}_k, y_k)\}_{k=0}^{N-1}$ be i.i.d. samples from $\mathcal{D}$. Define the population squared risk by

$$\mathcal{R}(f) \coloneqq \frac{1}{2}\,\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[(f(\mathbf{x}) - y)^2\big].$$

Using $y = f^*(\mathbf{x}) + \epsilon$ and $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$, we have

$$\mathcal{R}(f) = \frac{1}{2}\,\mathbb{E}_{\mathbf{x}}\big[(f(\mathbf{x}) - f^*(\mathbf{x}))^2\big] + \frac{\sigma^2}{2}.$$

The last term is the irreducible noise level. Accordingly, we define the excess risk

$$\mathcal{E}(f) \coloneqq \frac{1}{2}\,\mathbb{E}_{\mathbf{x}}\big[(f(\mathbf{x}) - f^*(\mathbf{x}))^2\big].$$

Recall the RKHS $\mathcal{H}$ induced by K, where each $f \in \mathcal{H}$ satisfies the reproducing property $f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}$.

One-pass SGD in RKHS. For a single sample $z = (\mathbf{x}, y)$, define the instantaneous loss

$$\ell(f; z) \coloneqq \frac{1}{2}\,(f(\mathbf{x}) - y)^2, \qquad f \in \mathcal{H}.$$

Its gradient in the RKHS $\mathcal{H}$ is given by

$$\nabla\ell(f; (\mathbf{x}, y)) = (f(\mathbf{x}) - y)\,K(\mathbf{x}, \cdot) \in \mathcal{H}.$$

Indeed, for any $g \in \mathcal{H}$, the directional derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell(f + tg; (\mathbf{x}, y))\bigg|_{t=0} = (f(\mathbf{x}) - y)\,g(\mathbf{x}) = \langle (f(\mathbf{x}) - y)K(\mathbf{x}, \cdot),\, g\rangle_{\mathcal{H}},$$

where we used the reproducing property $g(\mathbf{x}) = \langle g, K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}$.

Starting from $f_0 = 0 \in \mathcal{H}$, one-pass SGD updates f_k by

$$f_{k+1} = f_k - \eta_k \nabla\ell(f_k; (\mathbf{x}_k, y_k)) = f_k - \eta_k\big(f_k(\mathbf{x}_k) - y_k\big)\,K(\mathbf{x}_k, \cdot), \qquad k = 0, \ldots, N-1. \tag{36}$$

Translating the update to $\mathbb{H}$. Under the feature representation $f_{\boldsymbol{\theta}}(\cdot) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\cdot)\rangle_{\mathbb{H}}$, the update equation 36 is equivalent to the parameter update

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k\big(\langle \boldsymbol{\theta}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} - y_k\big)\,\boldsymbol{\phi}(\mathbf{x}_k), \qquad k = 0, \ldots, N-1. \tag{37}$$

with $\boldsymbol{\theta}_0 = 0$. For each $j \geqslant 1$, define the coordinates

$$\theta_{k,j} \coloneqq \langle \boldsymbol{\theta}_k, \mathbf{v}_j\rangle_{\mathbb{H}}, \qquad u_k^j \coloneqq \theta_{k,j} - \theta_j^*.$$

Using $y_k = f^*(\mathbf{x}_k) + \epsilon_k$ and taking inner products with $\mathbf{v}_j$ in equation 37, we obtain

$$u_{k+1}^j = u_k^j - \eta_k\Big(\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} - \epsilon_k\Big)\,\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}, \tag{38}$$

where $\mathbf{u}_k$ denotes the error vector in eigen-coordinates, i.e., $\langle \mathbf{u}_k, \mathbf{v}_j\rangle_{\mathbb{H}} = u_k^j$ for all $j \geqslant 1$.

Conditioned on $\mathbf{u}_k$, the drift satisfies

$$\mathbb{E}\big[\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}\,\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} \,\big|\, \mathbf{u}_k\big] = \langle \mathbf{u}_k, \mathcal{T}\mathbf{v}_j\rangle_{\mathbb{H}} = \lambda_j u_k^j.$$

Define the centered noise

$$\xi_k^j \coloneqq \langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}\,\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} - \lambda_j u_k^j - \epsilon_k\,\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}. \tag{39}$$

Then $\mathbb{E}[\xi_k^j \mid \mathbf{u}_k] = 0$, and equation 38 can be rewritten as

$$u_{k+1}^j = (1 - \eta_k\lambda_j)u_k^j - \eta_k\xi_k^j. \tag{40}$$

Our goal is to control the excess risk

$$\mathcal{E}(\boldsymbol{\theta}_k) := \frac{1}{2}\,\mathbb{E}_{\mathbf{x}}\big[(f_{\boldsymbol{\theta}_k}(\mathbf{x}) - f^*(\mathbf{x}))^2\big].$$

Using equation 35, we have the $L^2(\mathcal{D}_{\mathcal{X}})$ expansion

$$f_{\boldsymbol{\theta}_k}(\mathbf{x}) - f^*(\mathbf{x}) = \sum_{j\geqslant 1} \sqrt{\lambda_j}\, u_k^j\, \mathbf{e}_j(\mathbf{x}),$$

and therefore

$$\mathcal{E}(\boldsymbol{\theta}_k) = \frac{1}{2}\sum_{j\geqslant 1} \lambda_j (u_k^j)^2. \tag{41}$$

In particular, bounding $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)]$ reduces to bounding the second moments $\mathbb{E}[(u_k^j)^2]$ for all $j \geqslant 1$. Squaring equation 40 and conditioning on $\mathbf{u}_k$, the cross term vanishes since $\mathbb{E}[\xi_k^j \mid \mathbf{u}_k] = 0$, yielding

$$\mathbb{E}\big[(u_{k+1}^j)^2 \mid \mathbf{u}_k\big] = (1-\eta_k\lambda_j)^2 (u_k^j)^2 + \eta_k^2 \mathbb{E}\big[(\xi_k^j)^2 \mid \mathbf{u}_k\big]. \tag{42}$$

Taking expectation gives

$$\mathbb{E}\big[(u_{k+1}^j)^2\big] = (1-\eta_k\lambda_j)^2 \mathbb{E}\big[(u_k^j)^2\big] + \eta_k^2 \mathbb{E}\big[(\xi_k^j)^2\big]. \tag{43}$$

Proposition E.1. *For any $k \geqslant 1$ and any $j \geqslant 1$,*

$$\mathbb{E}\big[(u_k^j)^2\big] = \prod_{i=0}^{k-1}(1-\eta_i\lambda_j)^2 (u_0^j)^2 + \sum_{i=0}^{k-1} \eta_i^2\, \mathbb{E}\big[(\xi_i^j)^2\big] \prod_{\ell=i+1}^{k-1} (1-\eta_\ell\lambda_j)^2.$$

Proof. Iteratively unroll equation 43. □

Intrinsic time and excess risk bound Define the intrinsic time

$$t_k := \sum_{i=0}^{k-1} \eta_i, \qquad k \geqslant 1, \qquad t_0 := 0.$$

Assume $\eta_{\max} := \max_{0\leqslant i\leqslant N-1} \eta_i \leqslant \lambda_1^{-1}$. Then for any $0 \leqslant i < k$ and any $j \geqslant 1$,

$$\prod_{\ell=i}^{k-1}(1-\eta_\ell\lambda_j)^2 \leqslant \exp\Big(-2\lambda_j \sum_{\ell=i}^{k-1}\eta_\ell\Big) = \exp\big(-2\lambda_j(t_k - t_i)\big).$$

Combining this with Proposition E.1 and equation 41 yields

$$2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \leqslant \sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j t_k} (u_0^j)^2 + \sum_{i=0}^{k-1} \eta_i^2 \sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j (t_k - t_{i+1})}\, \mathbb{E}\big[(\xi_i^j)^2\big]. \tag{44}$$

Recall the hypercontractivity condition in Assumption 2.1. It gives that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\big[\langle \mathbf{u}, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2 \langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2\big] \leqslant C \cdot \mathbb{E}_{\mathbf{x}}\big[\langle \mathbf{u}, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2\big]\, \mathbb{E}_{\mathbf{x}}\big[\langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2\big]. \tag{45}$$

Proposition E.2. *Suppose equation 45 holds and $\epsilon_k \sim \mathcal{N}(0,\sigma^2)$ is independent of $\mathbf{x}_k$. Then for any $k \geqslant 0$ and any $j \geqslant 1$,*

$$\mathbb{E}\big[(\xi_k^j)^2\big] \leqslant \lambda_j\big(2C\,\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] + \sigma^2\big).$$

Proof. Recall the definition

$$\xi_k^j = \Big(\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} - \epsilon_k\Big)\, \langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} - \lambda_j u_k^j.$$

Let $A_k := \langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}} \langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}$ and $B_k := \epsilon_k \langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}$. Then $\xi_k^j = (A_k - \mathbb{E}[A_k \mid \mathbf{u}_k]) - B_k$ since $\mathbb{E}[A_k \mid \mathbf{u}_k] = \lambda_j u_k^j$.

Conditioned on $\mathbf{u}_k$, we have $\mathbb{E}[B_k \mid \mathbf{u}_k] = 0$ and $\mathbb{E}[A_k - \mathbb{E}[A_k \mid \mathbf{u}_k] \mid \mathbf{u}_k] = 0$. Moreover, ϵ_k is independent of $\mathbf{x}_k$ and $\mathbf{u}_k$, hence the cross term vanishes:

$$\mathbb{E}\big[(A_k - \mathbb{E}[A_k \mid \mathbf{u}_k])\, B_k \mid \mathbf{u}_k\big] = 0.$$

Therefore,

$$\begin{aligned}\mathbb{E}\big[(\xi_k^j)^2 \mid \mathbf{u}_k\big] &= \mathbb{E}\big[(A_k - \mathbb{E}[A_k \mid \mathbf{u}_k])^2 \mid \mathbf{u}_k\big] + \mathbb{E}\big[B_k^2 \mid \mathbf{u}_k\big] \\ &\leqslant \mathbb{E}[A_k^2 \mid \mathbf{u}_k] + \mathbb{E}[\epsilon_k^2] \cdot \mathbb{E}\big[\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2\big].\end{aligned}$$

The second term equals $\sigma^2 \lambda_j$ because

$$\mathbb{E}\big[\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2\big] = \langle \mathbf{v}_j, \mathcal{T}\mathbf{v}_j\rangle_{\mathbb{H}} = \lambda_j.$$

For the first term, apply equation 45 with $\mathbf{u} = \mathbf{u}_k$ and $\mathbf{v} = \mathbf{v}_j$:

$$\mathbb{E}[A_k^2 \mid \mathbf{u}_k] = \mathbb{E}\big[\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2 \langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2 \mid \mathbf{u}_k\big] \leqslant C \cdot \mathbb{E}\big[\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2 \mid \mathbf{u}_k\big]\, \mathbb{E}\big[\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2\big].$$

Using $\mathbb{E}[\langle \mathbf{v}_j, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2] = \lambda_j$ and

$$\mathbb{E}\big[\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x}_k)\rangle_{\mathbb{H}}^2 \mid \mathbf{u}_k\big] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\big[\langle \mathbf{u}_k, \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathbb{H}}^2\big] = 2\,\mathcal{E}(\boldsymbol{\theta}_k),$$

we obtain $\mathbb{E}[A_k^2 \mid \mathbf{u}_k] \leqslant 2C\,\lambda_j\,\mathcal{E}(\boldsymbol{\theta}_k)$. Combining the two bounds yields

$$\mathbb{E}\big[(\xi_k^j)^2 \mid \mathbf{u}_k\big] \leqslant \lambda_j\big(2C\,\mathcal{E}(\boldsymbol{\theta}_k) + \sigma^2\big).$$

Taking expectation over $\mathbf{u}_k$ gives the desired bound. □

Proposition E.3. *Assume $\eta_{\max} \leqslant \lambda_1^{-1}$ and equation 45 holds. Then for any $k \geqslant 1$,*

$$2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \leqslant \sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j t_k} (u_0^j)^2 + \sum_{i=0}^{k-1} \eta_i^2 \left(2C\,\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_i)] + \sigma^2\right) \sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j (t_k - t_{i+1})}.$$

Proof. Plug Proposition E.2 into equation 44 and use

$$\lambda_j e^{-2\lambda_j(t_k - t_{i+1})} \cdot \mathbb{E}[(\xi_i^j)^2] \leqslant \lambda_j e^{-2\lambda_j(t_k - t_{i+1})} \cdot \lambda_j\big(2C\,\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_i)] + \sigma^2\big).$$

□

The following proposition provides a uniform bound on the excess risk, when the maximal learning rate is smaller than a certain constant. We note that $\operatorname{tr}(\mathcal{T}) = \sum_{j\geqslant 1} \lambda_j = \mathbb{E}[K(\mathbf{x}, \mathbf{x})] < \infty$.

Proposition E.4 (Uniform boundedness of the excess risk)**.** *Assume equation 45 holds, and $\eta_{\max} < \frac{1}{2C\operatorname{tr}(\mathcal{T})}$. Then*

$$\sup_{k\leqslant N} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \lesssim 1.$$

Proof. From Proposition E.3, the first term satisfies

$$\sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j t_k} (u_0^j)^2 \leqslant \sum_{j\geqslant 1} \lambda_j (u_0^j)^2 = 2\,\mathcal{E}(\boldsymbol{\theta}_0).$$

For the second term, use $\eta_i^2 \leqslant \eta_{\max}\eta_i = \eta_{\max}(t_{i+1} - t_i)$ and obtain

$$\begin{aligned}\sum_{i=0}^{k-1} \eta_i^2 \sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j(t_k - t_{i+1})} &\leqslant \eta_{\max} \sum_{j\geqslant 1} \lambda_j^2 \int_0^{t_k} e^{-2\lambda_j(t_k - t)}\,\mathrm{d}t \\ &= \eta_{\max} \sum_{j\geqslant 1} \frac{\lambda_j}{2}\left(1 - e^{-2\lambda_j t_k}\right) \leqslant \frac{\eta_{\max}}{2}\operatorname{tr}(\mathcal{T}).\end{aligned}$$

Let $M_k \coloneqq \max_{0\leqslant i\leqslant k} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_i)]$. Then Proposition E.3 implies

$$2\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \leqslant 2\mathcal{E}(\boldsymbol{\theta}_0) + \frac{\eta_{\max}\operatorname{tr}(\mathcal{T})}{2}\Big(2C\,M_{k-1} + \sigma^2\Big).$$

Taking maximum over k and using $\eta_{\max} < (2C\operatorname{tr}(\mathcal{T}))^{-1}$ yields $M_k \lesssim 1$ by induction. □

With this uniform bound, we obtain the following bound on the excess risk.

Lemma E.5. *Under the assumptions of Proposition E.4, for any $k \geqslant 1$,*

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_k)] \lesssim \underbrace{\sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j t_k} (u_0^j)^2}_{S_k} + \underbrace{\sum_{i=0}^{k-1} \eta_i^2 \sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j (t_k - t_{i+1})}}_{\mathcal{N}_k}.$$

Proof. By Proposition E.4, we have $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_i)] \lesssim 1$ uniformly in $i \leqslant k$. Hence in Proposition E.3, the factor $(2C\,\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_i)] + \sigma^2)$ is bounded by a constant and can be absorbed into $\lesssim$. □

E.3 Bounding Signal Learning and Noise Accumulation

We now bound the two terms S_k and $\mathcal{N}_k$ in Lemma E.5. They correspond to the signal-learning term and the noise-accumulation term in the FSL equation 4. Since our goal is the final-step excess risk $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)]$, it suffices to bound S_N and $\mathcal{N}_N$. We denote the total intrinsic time by $T = t_N$.

E.3.1 Signal-Learning Term S_N

Proposition E.6 (Bound on S_N)**.** *Assume the source condition in Assumption 5.2. Then*

$$S_N := \sum_{j\geqslant 1} \lambda_j e^{-2\lambda_j t_N} (u_0^j)^2 \;\lesssim\; T^{-s},$$

where $T = t_N$.

Proof. Since $\boldsymbol{\theta}_0 = 0$, the source condition implies $u_0^j = -\theta_j^* = -a_j \lambda_j^{\frac{s-1}{2}}$, hence

$$S_N = \sum_{j\geqslant 1} a_j^2\, \lambda_j^s e^{-2\lambda_j T}.$$

Using $\sum_{j\geqslant 1} a_j^2 \leqslant 1$ and $\sup_{\lambda\geqslant 0} \lambda^s e^{-2\lambda T} \leqslant (s/(2eT))^s$, we obtain $S_N \lesssim T^{-s}$. □

E.3.2 Noise-Accumulation Term $\mathcal{N}_N$

We first provide a lemma that will be useful in the bound of $\mathcal{N}_N$.

Lemma E.7. *Assume the capacity condition $\lambda_j \lesssim j^{-\beta}$ for some $\beta > 1$. Then for all $\tau > 0$,*

$$\sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j \tau} \;\lesssim\; \min\left\{1,\; \tau^{-2+\frac{1}{\beta}}\right\}.$$

Proof. Let $\lambda_j \leqslant c\, j^{-\beta}$. Then

$$\sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j \tau} \leqslant c^2 \sum_{j\geqslant 1} j^{-2\beta} e^{-2c\tau j^{-\beta}} \;\lesssim\; \int_1^\infty x^{-2\beta} e^{-2c\tau x^{-\beta}}\, \mathrm{d}x.$$

With the change of variables $u = x^{-\beta}$ (so $x = u^{-1/\beta}$, $\mathrm{d}x = \frac{1}{\beta} u^{-1/\beta - 1} \mathrm{d}u$), the integral is bounded by

$$\frac{1}{\beta} \int_0^1 u^{1-\frac{1}{\beta}} e^{-2c\tau u}\, \mathrm{d}u \;\leqslant\; \frac{1}{\beta} \int_0^\infty u^{1-\frac{1}{\beta}} e^{-2c\tau u}\, \mathrm{d}u \;\asymp\; \tau^{-2+\frac{1}{\beta}}.$$

This yields the $\tau^{-2+1/\beta}$ bound. The constant bound follows from $\sum_{j\geqslant 1} \lambda_j^2 \leqslant \lambda_1 \sum_{j\geqslant 1} \lambda_j = \lambda_1 \mathrm{tr}(\mathcal{T}) < \infty$. □

Noise term $\mathcal{N}_N$ for power decay. We consider the power decay schedule

$$\eta_i = \eta_0\Big(1 - \frac{i}{N}\Big)^{\gamma}, \qquad i = 0, 1, \dots, N-1,$$

and the intrinsic time $t_k = \sum_{i=0}^{k-1} \eta_i$.

Lemma E.8 (Intrinsic time for power decay)**.** *For $\gamma > 0$, we have*

$$t_N \eqsim \eta_0 N, \qquad t_N - t_{i+1} \eqsim \eta_0 N\Big(\frac{N-i-1}{N}\Big)^{\gamma+1}, \quad i = 0, \dots, N-1.$$

Proof. Both relations follow from comparing the sums with the corresponding integrals:

$$\sum_{i=0}^{N-1}\Big(1 - \frac{i}{N}\Big)^{\gamma} \eqsim N\int_0^1 (1-x)^{\gamma}\,\mathrm{d}x, \qquad \sum_{\ell=i+1}^{N-1}\Big(1 - \frac{\ell}{N}\Big)^{\gamma} \eqsim N\int_{(i+1)/N}^1 (1-x)^{\gamma}\,\mathrm{d}x.$$

□

Proposition E.9 (Noise term for power decay)**.** *Assume the capacity condition and $\eta_0 \lesssim 1$. Let $\eta_i = \eta_0(1 - i/N)^{\gamma}$ with $\gamma > 0$. Suppose $\beta \neq \gamma + 1$. Then*

$$\mathcal{N}_N \coloneqq \sum_{i=0}^{N-1} \eta_i^2 \sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j(t_N - t_{i+1})} \lesssim \eta_0\, T^{-\min\{1-\frac{1}{\beta},\, \frac{\gamma}{\gamma+1}\}},$$

where $T = t_N$.

Proof. By Lemma E.7,

$$\sum_{j\geqslant 1} \lambda_j^2 e^{-2\lambda_j(t_N - t_{i+1})} \lesssim \min\Big\{1,\ (t_N - t_{i+1})^{-2+\frac{1}{\beta}}\Big\}.$$

Let $m = N - i - 1$ and choose a cutoff m_0 such that

$$m_0 \eqsim N\, t_N^{-\frac{1}{\gamma+1}} \qquad \Longleftrightarrow \qquad t_N\Big(\frac{m_0}{N}\Big)^{\gamma+1} \eqsim 1.$$

We split the sum into $m \leqslant m_0$ (near the end, use the bound $\leqslant 1$) and $m > m_0$ (use the polynomial bound). Using $\eta_i = \eta_0((m+1)/N)^{\gamma}$ and Lemma E.8, we have the following bound.

(i) Tail part $m \leqslant m_0$.

$$\sum_{m=0}^{m_0} \eta_0^2\Big(\frac{m+1}{N}\Big)^{2\gamma} \lesssim \eta_0^2\, N^{-2\gamma}\, m_0^{2\gamma+1} \eqsim \eta_0\, t_N^{-\frac{\gamma}{\gamma+1}}.$$

(ii) Head part $m > m_0$. Using $t_N - t_{i+1} \eqsim t_N(m/N)^{\gamma+1}$ and $\eta_i = \eta_0(m/N)^{\gamma}$, we have

$$\eta_i^2(t_N - t_{i+1})^{-2+\frac{1}{\beta}} \lesssim \eta_0^2\Big(\frac{m+1}{N}\Big)^{2\gamma} \cdot t_N^{-2+\frac{1}{\beta}}\Big(\frac{m}{N}\Big)^{(\gamma+1)(-2+\frac{1}{\beta})} = \eta_0^2\, t_N^{-2+\frac{1}{\beta}}\Big(\frac{m}{N}\Big)^{-2+\frac{\gamma+1}{\beta}}.$$

Let $\nu \coloneqq -2 + \frac{\gamma+1}{\beta}$. Then

$$\sum_{m>m_0} \eta_i^2(t_N - t_{i+1})^{-2+\frac{1}{\beta}} \lesssim \eta_0^2\, t_N^{-2+\frac{1}{\beta}} \sum_{m>m_0}\Big(\frac{m}{N}\Big)^{\nu}.$$

The head part We then consider two cases of the head part.

- Case 1: $\gamma > \beta - 1$ (i.e., $\nu > -1$). Since $(m/N)^{\nu} \leqslant 1$ for $m \leqslant N$, we have $\sum_{m>m_0}(m/N)^{\nu} \lesssim N$, and hence

  $$\sum_{m>m_0} \eta_i^2(t_N - t_{i+1})^{-2+\frac{1}{\beta}} \lesssim \eta_0^2\, t_N^{-2+\frac{1}{\beta}} \cdot N \eqsim \eta_0\, t_N^{-\left(1-\frac{1}{\beta}\right)},$$

  where we used $t_N \eqsim \eta_0 N$.

- Case 2: $\gamma < \beta - 1$ (i.e., $\nu < -1$). In this case, $\sum_{m>m_0}(m/N)^\nu \lesssim N\,(m_0/N)^{\nu+1}$, so

$$\sum_{m>m_0} \eta_i^2 (t_N - t_{i+1})^{-2+\frac{1}{\beta}} \;\lesssim\; \eta_0^2\, t_N^{-2+\frac{1}{\beta}} \cdot N\Big(\frac{m_0}{N}\Big)^{\nu+1}.$$

  Since $m_0 \eqsim N\, t_N^{-1/(\gamma+1)}$, we have

$$\Big(\frac{m_0}{N}\Big)^{\nu+1} = t_N^{-\frac{\nu+1}{\gamma+1}} = t_N^{-\left(\frac{1}{\beta}-\frac{1}{\gamma+1}\right)},$$

  and therefore

$$\sum_{m>m_0} \eta_i^2 (t_N - t_{i+1})^{-2+\frac{1}{\beta}} \;\lesssim\; \eta_0^2\, N\, t_N^{-2+\frac{1}{\gamma+1}} \;\eqsim\; \eta_0\, t_N^{-\frac{\gamma}{\gamma+1}},$$

  again using $t_N \eqsim \eta_0 N$.

Combining the two cases yields the following bound on the head part:

$$\sum_{m>m_0} \eta_i^2 (t_N - t_{i+1})^{-2+\frac{1}{\beta}} \;\lesssim\; \eta_0\, t_N^{-\min\left\{1-\frac{1}{\beta},\, \frac{\gamma}{\gamma+1}\right\}}.$$

Then we finish the proof by combining (i) and (ii). □

E.4 Proof of Theorem 5.3

Proof. By Lemma E.5 with $k = N$,

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \;\lesssim\; S_N + \mathcal{N}_N.$$

Let $T \coloneqq t_N$. Proposition E.6 gives $S_N \lesssim T^{-s}$. Proposition E.9 gives

$$\mathcal{N}_N \;\lesssim\; \eta_0\, T^{-\mu}, \qquad \mu \coloneqq \min\left\{1 - \frac{1}{\beta},\, \frac{\gamma}{\gamma+1}\right\}.$$

Moreover, Lemma E.8 yields $T = t_N \eqsim \eta_0 N$. Hence

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \;\lesssim\; (\eta_0 N)^{-s} + \eta_0 (\eta_0 N)^{-\mu}. \tag{46}$$

(i) Easy-task regime ($s \geqslant 1 - \frac{1}{\beta}$). Choose $\gamma > \beta - 1$, so that $\mu = 1 - \frac{1}{\beta}$. Let $x \coloneqq \eta_0 N$. Then equation 46 becomes

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim x^{-s} + N^{-1} x^{1/\beta}.$$

Balancing yields $x^{s+1/\beta} \eqsim N$, i.e.,

$$\eta_0 \eqsim N^{-\frac{s\beta-\beta+1}{s\beta+1}},$$

and substituting back gives $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim N^{-\frac{s\beta}{s\beta+1}}$.

(ii) Hard-task regime ($s < 1 - \frac{1}{\beta}$). Take $\eta_0 \eqsim 1$ (a sufficiently small constant so that Proposition E.4 applies). Then $T \eqsim N$ and

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim N^{-s} + N^{-\mu}.$$

If $\gamma > \frac{s}{1-s}$ then $\frac{\gamma}{\gamma+1} > s$. Together with $s < 1 - \frac{1}{\beta}$ we have $\mu > s$, so $N^{-\mu} \lesssim N^{-s}$ and therefore $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \lesssim N^{-s}$. □

E.5 Proof of Proposition 5.4

Proof. Let $\eta_{\max} \coloneqq \max_{0 \leqslant i \leqslant N-1} \eta_i$ and assume $\eta_{\max} \lesssim 1$. We prove that there exists a data distribution $\mathcal{D}$ satisfying Assumptions 5.1 and 5.2 such that $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim N^{-s}$. This implies the desired lower bound on $\sup_{\mathcal{D}} \mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)]$.

Recall from equation 40 that for each $j \geqslant 1$,

$$u_{k+1}^j = (1 - \eta_k \lambda_j) u_k^j - \eta_k \xi_k^j, \qquad \text{with } \mathbb{E}[\xi_k^j \mid \mathbf{u}_k] = 0.$$

By the second-moment update equation 43, we have

$$\mathbb{E}[(u_{k+1}^j)^2] = (1-\eta_k\lambda_j)^2\mathbb{E}[(u_k^j)^2] + \eta_k^2\mathbb{E}[(\xi_k^j)^2] \geqslant (1-\eta_k\lambda_j)^2\mathbb{E}[(u_k^j)^2].$$

Iterating this inequality yields, for any $k \geqslant 1$,

$$\mathbb{E}[(u_k^j)^2] \geqslant \prod_{i=0}^{k-1}(1-\eta_i\lambda_j)^2\,(u_0^j)^2. \tag{47}$$

Fix an operator spectrum $\lambda_j = j^{-\beta}$ (which satisfies the capacity condition with parameter β). Choose a target function supported on a single eigen-direction:

$$f^*(\mathbf{x}) = \lambda_{j^*}^{s/2}\,\mathbf{e}_{j^*}(\mathbf{x}), \quad \text{i.e.,} \quad a_{j^*} = 1,\ a_j = 0\ (j \neq j^*),$$

where the index j^* is to be determined. This target function satisfies the source condition (Assumption 5.2) with $\sum_j a_j^2 = 1$. We also take the regression noise to satisfy $\epsilon \mid \mathbf{x} \sim \mathcal{N}(0,\sigma^2)$ with $\sigma^2 > 0$. Note that the lower bound below only uses equation 47, and therefore it does not depend on the noise distribution.

We initialize $\boldsymbol{\theta}_0 = 0$, so $u_0^j = -\theta_j^*$. For the above f^*, the corresponding coefficients in $\mathbb{H}$ satisfy $\theta_{j^*}^* = \lambda_{j^*}^{\frac{s-1}{2}}$ and $\theta_j^* = 0$ for $j \neq j^*$, hence

$$(u_0^j)^2 = \begin{cases} \lambda_{j^*}^{s-1}, & j = j^*, \\ 0, & j \neq j^*. \end{cases} \tag{48}$$

Recall $\mathcal{E}(\boldsymbol{\theta}_k) = \frac{1}{2}\sum_{j\geqslant 1}\lambda_j(u_k^j)^2$ (cf. equation 41). Combining equation 47 and equation 48 gives

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \geqslant \frac{1}{2}\,\lambda_{j^*}\,\mathbb{E}[(u_N^{j^*})^2] \geqslant \frac{1}{2}\,\lambda_{j^*}\prod_{i=0}^{N-1}(1-\eta_i\lambda_{j^*})^2\lambda_{j^*}^{s-1} = \frac{1}{2}\,\lambda_{j^*}^s\prod_{i=0}^{N-1}(1-\eta_i\lambda_{j^*})^2.$$

Let $T := t_N = \sum_{i=0}^{N-1}\eta_i$. Since $T \leqslant \eta_{\max}N \lesssim N$, it suffices to show $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim T^{-s}$. Define

$$j_0 := \min\left\{j \geqslant 1 :\ \eta_{\max}\lambda_j \leqslant \tfrac{1}{2}\right\}.$$

For any $j \geqslant j_0$, we have $\eta_i\lambda_j \leqslant 1/2$ for all i, and thus $1-x \geqslant e^{-2x}$ for $x \in [0,1/2]$. This gives

$$\prod_{i=0}^{N-1}(1-\eta_i\lambda_j)^2 \geqslant \prod_{i=0}^{N-1}e^{-4\eta_i\lambda_j} = e^{-4\lambda_j T}.$$

Hence for any $j^* \geqslant j_0$,

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \geqslant \frac{1}{2}\,\lambda_{j^*}^s\,e^{-4\lambda_{j^*}T}. \tag{49}$$

Let

$$j^* := \max\left\{j_0,\ \left\lceil\left(\frac{4T}{s}\right)^{1/\beta}\right\rceil\right\}.$$

If $\left\lceil\left(\frac{4T}{s}\right)^{1/\beta}\right\rceil \geqslant j_0$, then $\lambda_{j^*} = j^{*-\beta} \eqsim \frac{s}{4T}$ and equation 49 yields

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim \left(\frac{s}{4T}\right)^s e^{-s} \gtrsim T^{-s}.$$

If instead $\left\lceil\left(\frac{4T}{s}\right)^{1/\beta}\right\rceil < j_0$, then $j^* = j_0$ and λ_{j^*} is a positive constant (depending only on $\eta_{\max}$). Moreover, in this case $T \lesssim 1/\lambda_{j_0}$, hence $\lambda_{j_0}T \lesssim 1$ and equation 49 gives $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim 1 \gtrsim T^{-s}$.

In both cases, we have shown $\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim T^{-s}$. Finally, since $T \leqslant \eta_{\max}N \lesssim N$, we conclude

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\theta}_N)] \gtrsim T^{-s} \gtrsim N^{-s}.$$

□

F An Alternative Proof of Optimal Intrinsic-Time LRS Profile (Step 1)

F.1 A Quick Introduction to Variational Calculus

Consider the simplest setting

$$\mathcal{F}[f] \;=\; \int_a^b L\big(z,\, f(z),\, f'(z)\big)\,\mathrm{d}z,$$

where $f : [a,b] \to \mathbb{R}$ is admissible[1] and L (the *Lagrangian*) is smooth.

Euler–Lagrange equation. Take a smooth perturbation h with $h(a) = h(b) = 0$ and set $f_\varepsilon = f + \varepsilon h$. The first variation is

$$\delta\mathcal{F}[f;h] = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}[f_\varepsilon]\Big|_{\varepsilon=0} = \int_a^b \Big(\partial_f L\, h + \partial_{f'} L\, h'\Big)\,\mathrm{d}z = \int_a^b \Big(\partial_f L - \frac{\mathrm{d}}{\mathrm{d}z}\,(\partial_{f'} L)\Big)h(z)\,\mathrm{d}z,$$

where the last step uses the integration by part and $h(a) = h(b) = 0$).

[1] For instance, $f \in C^1([a,b])$ with prescribed boundary values $f(a) = A,\ f(b) = B$.

Because this must vanish for *all* such h, the integrand is zero:

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\partial_{f'}L\right) - \partial_f L = 0 \qquad \text{(Euler–Lagrange)}$$

Beltrami identity. If $L = L(f, f')$ has no explicit z-dependence, multiply the Euler–Lagrange equation by f' and integrate once:

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(f'\,\partial_{f'}L - L\right) = 0 \quad \Longrightarrow \quad L - f'\,\partial_{f'}L = C \qquad \text{(Beltrami)}$$

where C is a constant. The quantity $L - f'\partial_{f'}L$ is (up to sign) the Legendre transform of L with respect to f' and is conserved – an analogue of energy conservation.

F.2 Proof of Optimal Intrinsic-Time LRS Profile

Let us start by fixing the total intrinsic time T. Then, the problem becomes

$$\min_t \int_0^N L(T, t(z), t'(z))\,\mathrm{d}z \qquad (50)$$
$$\text{s.t. } t(0) = 0, t(N) = T.$$

The Beltrami identity implies

$$\underbrace{(1+T-t)^{-(2-\frac{1}{\beta})}(t')^2}_{L} - \underbrace{t' \cdot 2(1+T-t)^{-(2-\frac{1}{\beta})}t'}_{t'\partial_{t'}L} = C.$$

This gives

$$\varphi(t) = t' = C(1+T-t)^{1-\frac{1}{2\beta}} = \varphi(0)\left(1 - \frac{t}{1+T}\right)^{1-\frac{1}{2\beta}}. \qquad (51)$$

This means that the optimal LRS follows a power law with respect to the intrinsic time. Now we turn to derive the LRS in terms of training steps. Note that equation 51 gives

$$\begin{aligned}
&\frac{\mathrm{d}t}{\mathrm{d}z} = \varphi(0)\left(1 - \frac{t}{1+T}\right)^{1-\frac{1}{2\beta}} \\
&\Longrightarrow \varphi(0)^{-1}\left(1 - \frac{t}{1+T}\right)^{-(1-\frac{1}{2\beta})}\mathrm{d}t = \mathrm{d}z \\
&\Longrightarrow -\varphi(0)^{-1}2\beta(1+T)\left(1 - \frac{t}{1+T}\right)^{\frac{1}{2\beta}} = z - C
\end{aligned}$$

By maching the end point $t(0) = 0, t(N) = T$, we obtian the following intrinsic time function:

$$t(z) = 1 + T - \left((1+T)^{\frac{1}{2\beta}} - \frac{a_{N,T}}{2\beta}z\right)^{2\beta},$$

where $a_{N,T} = \frac{2\beta}{N}((1+T)^{\frac{1}{2\beta}} - 1)$. Differentiating with respect to z gives the physical LRS:

$$\eta(z) = t'(z) = \frac{2\beta}{N}\left((1+T)^{\frac{1}{2\beta}} - 1\right)^{2\beta}\left(1 + \frac{1}{(1+T)^{\frac{1}{2\beta}} - 1} - \frac{z}{N}\right)^{2\beta-1} \qquad (52)$$

The above derivation means that for any fixed T, the solution to the Euler-Lagrange equation $\partial_t L - \frac{\mathrm{d}}{\mathrm{d}z}(\partial_{t'}L) = 0$ is given by Eq. equation 52.