Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules

Binghui Li[1,*] **Fengling Chen**[2,*] **Zixun Huang**[2,*] **Lean Wang**[3,*] **Lei Wu**[1,2,4,†]

[1]Center for Machine Learning Research, Peking University
[2]School of Mathematical Sciences, Peking University
[3]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[4]AI for Science Institute, Beijing

{libinghui, lean}@pku.edu.cn, flchen_lwycc@stu.pku.edu.cn
alexpku@stu.pku.edu.cn, leiwu@math.pku.edu.cn

We strongly recommend reading the arXiv version of this paper, available at https://arxiv.org/abs/2509.19189.

Abstract

Scaling laws have emerged as a unifying lens for understanding and guiding the training of large language models (LLMs). However, existing studies predominantly focus on the final-step loss, leaving open whether the entire *loss dynamics* obey similar laws and, crucially, how the *learning rate schedule* (LRS) shapes them. We address these gaps in a controlled theoretical setting by analyzing stochastic gradient descent (SGD) on a power-law kernel regression model. The key insight is a novel **intrinsic-time** viewpoint, which captures the training progress more faithfully than iteration count. We then establish a **Functional Scaling Law (FSL)** that captures the full loss trajectory under arbitrary LRSs, with the schedule's influence entering through a simple convolutional functional. We further instantiate the theory for three representative LRSs—constant, exponential decay, and warmup–stable–decay (WSD)—and derive explicit scaling relations in both data- and compute-limited regimes. These comparisons explain key empirical phenomena: (i) higher-capacity models are more data- and compute-efficient; (ii) learning-rate decay improves training efficiency; and (iii) WSD-type schedules outperform pure decay. Finally, experiments on LLMs ranging from 0.1B to 1B parameters demonstrate the practical relevance of FSL as a surrogate model for fitting and predicting loss trajectories in large-scale pre-training.

1 Introduction

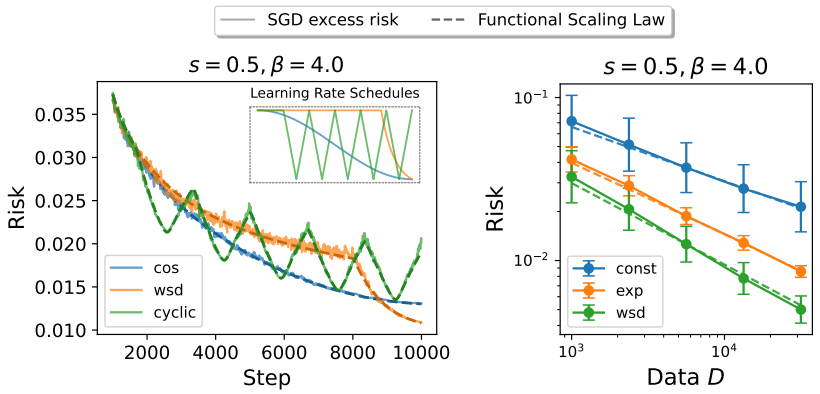
It is well established that the training of large-scale deep learning models mysteriously follows *scaling laws*, which describe how model performance scales *predictably* with available resources such as compute or data [19]. In particular, the landmark study by Kaplan et al. [25] demonstrated that, in LLM pre-training, the loss L decreases with model size M and dataset size D according to a power-law relation:

$$L(M, D) = L_0 + C_M M^{-\alpha_M} + C_D D^{-\alpha_D}, \tag{1}$$

where α_M and α_D are the scaling exponents, L_0 denotes the irreducible loss, and C_M, C_D are some constants. Such empirical relations have proven remarkably robust across scales, architec-

[*]Equal contribution.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Loss Dynamics (b) Scaling Behavior

Figure 1: **FSL accurately captures the loss dynamics and scaling behavior of SGD in PLK regression.** In both subplots, solid lines denote the results of SGD, while dashed lines represent the corresponding FSL predictions. **(a)** FSL accurately tracks the loss dynamics of SGD, averaged over 1000 runs, for three learning rate schedules: cosine, WSD-like, and a non-standard cyclic schedule. **(b)** FSL predictions (dashed) are computed using the analytical forms from Section 5, and compared with the mean of 200 SGD runs (solid).

tures, and training setups [20, 59, 36] and have become foundational principles for guiding LLM development [18, 24, 1, 5, 54, 27]. In practice, they are now routinely used to design optimal resource-allocation strategies [20] and to tune key hyperparameters such as learning rates and batch sizes [36, 29].

Despite their empirical success, the theoretical understanding of scaling laws remains limited. Recent studies have begun to illuminate the underlying mechanisms [53, 22, 39, 62, 23, 42, 43, 2, 14, 3, 7, 35, 49, 8, 69], yet two important gaps persist:

- **Determinants of scaling efficiency.** Existing studies lack a systematic characterization of how key factors—such as model capacity, task difficulty, and hyperparameter choices—govern scaling efficiency, as reflected by the exponents α_M and α_D. In particular, **learning rate schedules (LRSs)** are known to be critical in practice [40, 4, 17], but their precise role in shaping scaling behavior remains unclear.
- **Beyond the final-step loss.** The scaling law (1) focuses only on the end-of-training loss [25, 20], thus leaving open whether the *full trajectory* follows similar laws. Empirical studies [58, 38] suggest this possibility, but the fits there are still crude and lack theoretical grounding.

1.1 Our Contribution

In this paper, we take a step toward addressing these gaps in a controlled yet representative theoretical setting. We study stochastic gradient descent (SGD) training of the **power-law kernel (PLK)** regression—a widely adopted surrogate for scaling-law analysis [7, 3, 49, 35, 8]. The PLK regression is characterized by four parameters: the task difficulty s, the capacity exponent β, the model size M, and the label-noise level σ. To capture the influence of learning-rate schedules (LRSs), we model SGD via an **intrinsic-time SDE**, in which the concept of **intrinsic time** emerges as a key quantity enabling a unified characterization of how different LRSs shape the loss dynamics. Building on this formulation, we establish the **Functional Scaling Law (FSL)**, which provides a unified description of the entire *loss dynamics*—beyond the traditional final-loss prediction.

Concretely, for a general intrinsic-time LRS $\gamma : [0, \infty) \to [0, \infty)$, and under some conditions, the dynamics of the expected loss $\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)]$ (where t denotes the intrinsic time) satisfies:

$$\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \underbrace{\frac{\sigma^2}{2}}_{\text{irreducible error}} \eqsim \underbrace{\frac{1}{M^{s\beta}}}_{\text{approx. error}} + \underbrace{e(t)}_{\text{signal learning}} + \underbrace{\int_0^t \mathcal{K}(t-z)\,[e(z)+\sigma^2]\,\gamma(z)\,\mathrm{d}z}_{\text{noise accumulation}}, \tag{2}$$

where $e(t) = (1+t)^{-s}$ and $\mathcal{K}(t) = (1+t)^{-(2-1/\beta)}$. Each term in FSL admits clear interpretation: $\frac{\sigma^2}{2}$ denotes the **irreducible error** caused by label noise, $M^{-s\beta}$ represents the **approximation error**, $e(t)$ characterizes the **signal-learning dynamics** under noiseless (full-batch) gradient descent, and the final term captures the injection and dissipation of gradient noise, with the LRS γ entering through

Table 1: **Learning-rate schedule (LRS) strongly influences scaling efficiency in power-law kernel regression.** Efficiency is determined by two key factors: relative task difficulty $s \in (0, \infty)$ and model capacity $\beta > 1$. We distinguish between an *easy-learning regime* ($s \geqslant 1 - 1/\beta$) and a *hard-learning regime* ($s < 1 - 1/\beta$).

Learning Rate Schedule (LRS)	Data-Optimal Scaling Laws		Compute-Optimal Scaling Laws	
	Easy	Hard	Easy	Hard
Constant	$D^{-\frac{s}{s+1}}$		$C^{-\frac{s\beta}{1+s\beta+\beta}}$	
Exponential-decay	$D^{-\frac{s\beta}{1+s\beta}}(\log D)^{\frac{s\beta}{1+s\beta}}$	$D^{-s}(\log D)^{s}$	$C^{-\frac{s\beta}{2+s\beta}}(\log C)^{\frac{s\beta}{2+s\beta}}$	$C^{-\frac{s\beta}{1+\beta}}(\log C)^{\frac{s\beta}{1+\beta}}$
Warmup-stable-decay (WSD)	$D^{-\frac{s\beta}{1+s\beta}}(\log D)^{\frac{s\beta-s}{1+s\beta}}$	D^{-s}	$C^{-\frac{s\beta}{2+s\beta}}(\log C)^{\frac{s\beta-s}{2+s\beta}}$	$C^{-\frac{s\beta}{1+\beta}}$

a tractable **convolutional functional**. The function $\mathcal{K}$, referred to as **forgetting kernel**, quantifies how fast the injected noise dissipates during training.

Building on FSL, we derive concrete scaling laws for the final-step loss under three representative LRSs—constant, exponential decay [15], and warmup–stable–decay (WSD) [68, 21]—in both **data-limited** and **compute-limited regimes**. The results, summarized in Table 1, recover and extend prior analyses [7, 8, 49, 35], and reveal several unifying insights.

- **Scaling efficiency of different schedules.** WSD achieves the best scaling efficiency, followed by exponential decay and then constant schedules. This efficiency hierarchy provides theoretical justification for learning-rate decay and explains empirical success of WSD [68, 21, 57, 36].
- **Role of model capacity.** Higher-capacity models are consistently more efficient in both compute and data, highlighting the necessity of scaling model capacity [25].
- **Data–model trade-off.** Compute-optimal training requires scaling data more than model size, consistent with established heuristics in LLM pre-training [20].
- **Scaling law for peak learning rate.** Optimal scaling requires the peak learning rate (LR) to scale appropriately with the training budget (data or compute), revealing the importance of careful peak LR tuning [6, 29].

Beyond PLK regression, we further apply the FSL ansatz to fit and predict loss trajectories from LLM pre-training experiments with model sizes ranging from 0.1B to 1B parameters, covering both dense and MoE architectures. These results highlight the potential of FSL as a practical surrogate for understanding and guiding LLM pre-training.

To better situate our contribution, we provide a detailed comparison with related work in Appendix B.

Notation. For any $n \in \mathbb{N}$, let $[n] := \{1, 2, \ldots, n\}$. For a positive semi-definite (PSD) matrix $\mathbf{S}$, denote by $\mu_j(\mathbf{S})$ its j-th largest eigenvalue, and define the $\mathbf{S}$-induced norm $\|\mathbf{u}\|_{\mathbf{S}} := \sqrt{\mathbf{u}^\top \mathbf{S}\mathbf{u}}$ for any vector $\mathbf{u}$. We write $\mathbf{A} \preceq \mathbf{B}$ (resp. $\mathbf{A} \succeq \mathbf{B}$) if $\mathbf{B} - \mathbf{A}$ (resp. $\mathbf{A} - \mathbf{B}$) is PSD. Throughout the paper, we use $\eqsim$ to denote equivalence up to a constant factor, and $\lesssim$ (resp. $\gtrsim$) to denote an inequality up to a 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 $C_1 f(t) \leqslant g(t) \leqslant C_2 f(t),\ \forall\, t \geqslant 0$.

2 Power-Law Kernel (PLK) Regression

Let $\mathcal{X}$ denote the input domain and $\mathcal{D}$ the input distribution, and assume labels are generated by $y = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta}^* \rangle + \epsilon$, where $f^*(\mathbf{x}) := \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\theta}^* \rangle$ is the target function, and the label noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{x}$. Here $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^N$ with $N \in \mathbb{N}_+ \cup \{\infty\}$ is a feature map, satisfying the following assumption:

Assumption 2.1 (Hypercontractivity)**.** Let $\mathbf{H} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top]$ be the feature covariance. There exist constants $C_1, C_2 > 0$ such that for any PSD matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, $C_1 \operatorname{tr}(\mathbf{H}\mathbf{A})\, \mathbf{A} \preceq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\big(\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{A} \boldsymbol{\phi}(\mathbf{x})\big)\ \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top \Big] - \mathbf{H}\mathbf{A}\mathbf{H} \preceq C_2 \operatorname{tr}(\mathbf{H}\mathbf{A})\, \mathbf{A}$.

This condition ensures that the feature distribution is sufficiently regular—its fourth-order moments are controlled by the second-order ones [41]. It holds, for example, for Gaussian features $\boldsymbol{\phi}(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{H})$ with $C_1 = 1, C_2 = 2$ (see Lemma G.1). For simplicity, we also assume:

Assumption 2.2. $\mathbf{H} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$.

To learn f^*, we consider a **model of width** M: $f(\mathbf{x}; \mathbf{v}) = \sum_{j=1}^{M} v_j \mathbf{w}_j^\top \boldsymbol{\phi}(\mathbf{x}) =: \langle \mathbf{v}, \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \rangle$, where $\mathbf{v} \in \mathbb{R}^M$ denotes trainable weights and $\mathbf{W} \in \mathbb{R}^{M\times N}$ projects the N-dimensional features onto an M-dimensional subspace. We study two choices of projection $\mathbf{W}$:

- **Top-M features:** $\mathbf{w}_j = \boldsymbol{e}_j$ for $j \in [M]$, i.e., selecting the top-M features $\{\phi_j\}_{j=1}^M$;
- **Random-M features:** $\mathbf{w}_j \sim \mathcal{N}(0, I_N)$ independently for $j \in [M]$.

The top-M setting is a particularly simple yet analytically representative case, widely adopted in prior scaling-law studies [43, 13]. For random features [3, 7, 49, 35, 8], we will show that, in certain regimes, their scaling behavior closely parallels that of the top-M case. As clarified in Appendix A.3, our setup is equivalent to learning with the kernel $K_\phi(\mathbf{x}, \mathbf{x}') := \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$.

We now formalize the key notions of *model capacity* and *task difficulty*. Let $\widehat{\phi}_j := \phi_j / \lambda_j^{1/2}$ for $j \in [N]$. So $\{\widehat{\phi}_j\}_{j=1}^N$ forms an orthonormal basis of $L^2(\mathcal{D})$.

Assumption 2.3 (Model capacity)**.** The spectrum of the feature map satisfies $\lambda_j \eqsim j^{-\beta}, \beta > 1$.

The condition $\beta > 1$ ensures $\mathrm{tr}(\mathbf{H}) = \sum_{j=1}^N \lambda_j \leqslant C$ for some constant C independent of N, making our analysis *dimension-free* and applicable to the infinite-dimensional setting ($N = \infty$).

For the top-M features, the model takes the form $f(\cdot; \mathbf{v}) = \sum_{j=1}^M v_j \phi_j = \sum_{j=1}^M v_j \lambda_j^{1/2} \widehat{\phi}_j \eqsim \sum_{j=1}^M v_j\, j^{-\beta/2} \widehat{\phi}_j$ reveals that higher-index (less significant) features are increasingly down-weighted by the factor $j^{-\beta/2}$. As β increases, the spectrum decays more rapidly, and the model *effectively* relies on fewer features. Hence, the model's expressive power is governed by two complementary factors: (i) the **model size** M, which controls how many features are retained, and (ii) the **capacity exponent** β, which controls how quickly these features decay in importance.

Assumption 2.4 (Task difficulty)**.** Suppose $|\theta_j^*|^2 \eqsim j^{-1} \lambda_j^{s-1}$ for some $s > 0$.

Under Assumptions 2.3 and 2.4, the target function admits the expansion $f^* = \sum_{j=1}^N \theta_j^* \phi_j \eqsim \sum_{j=1}^N j^{-1/2} \lambda_j^{s/2} \widehat{\phi}_j \eqsim \sum_{j=1}^N j^{-(s\beta+1)/2} \widehat{\phi}_j$. Since $\{\hat{\phi}_j\}$ are orthonormal, this assumption implies that the spectral energy of f^* decays as a power law. The exponent $\alpha := s\beta$ therefore quantifies the task's **intrinsic difficulty**, which depends only on the target function itself and is independent of the model spectrum. In contrast, s measures the **relative difficulty** with respect to a model of capacity β: for a fixed f^* (fixed α), adopting a higher-capacity model (smaller β) increases $s = \alpha/\beta$, making the task relatively easier. In other words, the same task appears easier to a higher-capacity model.

We remark that similar assumptions have been widely used in the analysis of kernel methods [12, 11, 56, 9, 39]. Our work builds upon and extends this line of research.

3 One-Pass SGD and Intrinsic-Time SDE

Given a data point $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$ and a model $f(\cdot; \mathbf{v})$, define the loss $\ell(\mathbf{z}, \mathbf{v}) = \frac{1}{2}\big(f(\mathbf{x}; \mathbf{v}) - y\big)^2$. Then, the population risk is $\mathcal{R}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}}[\ell(\mathbf{z}, \mathbf{v})] = \frac{1}{2}\|\mathbf{W}^\top \mathbf{v} - \boldsymbol{\theta}^*\|_{\mathbf{H}}^2 + \frac{\sigma^2}{2} =: \mathcal{E}(\mathbf{v}) + \frac{\sigma^2}{2}$, where $\mathcal{E}(\mathbf{v})$ denotes the excess risk. We minimize $\mathcal{R}(\mathbf{v})$ via **one-pass SGD**, given by

$$\mathbf{v}_{k+1} = \mathbf{v}_k - \frac{\eta_k}{B_k} \sum_{\mathbf{z} \in S_k} \nabla_{\mathbf{v}} \ell(\mathbf{z}, \mathbf{v}_k), \tag{3}$$

where $S_k := \{(\mathbf{x}_{k,j}, y_{k,j})\}_{j=1}^{B_k}$ is a mini-batch of *i.i.d.* samples, η_k and B_k are the learning rate and batch size, respectively. The initialization is set to $\mathbf{v}_0 = \mathbf{0}$.

Throughout, we refer to $\boldsymbol{\eta} := (\eta_0, \eta_1, \ldots, \eta_{K-1})$ as the learning rate schedule (LRS). Common choices in practice include the cosine [37, 59], WSD [21], and multi-step [5] schedules (see Appendix A.2 for details). To analyze the effect of LRS, we rewrite (3) as

$$\mathbf{v}_{k+1} = \mathbf{v}_k - \eta_k\big(\nabla \mathcal{R}(\mathbf{v}_k) + \boldsymbol{\xi}_k\big), \tag{4}$$

where the gradient noise $\boldsymbol{\xi}_k = \frac{1}{B_k} \sum_{\mathbf{z} \in S_k} \nabla \ell(\mathbf{z}, \mathbf{v}_k) - \nabla \mathcal{R}(\mathbf{v}_k)$ satisfies $\mathbb{E}[\boldsymbol{\xi}_k] = 0, \mathbb{E}[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top] = \frac{1}{B_k} \boldsymbol{\Sigma}(\mathbf{v}_k)$, with $\boldsymbol{\Sigma}(\cdot)$ denoting the noise covariance for batch size 1.

Continuous-time limit. Following prior work [30, 31, 32, 33, 50, 34], we analyze the continuous-time limit of SGD rather than the discrete update (3) or (4). This perspective makes the analysis more tractable and clarifies the emergence of scaling laws. Fix a discretization step size $h > 0$ and

let $\varphi_k := \eta_k/h$ for $k \in \mathbb{N}$. Then, (4) becomes $\mathbf{v}_{k+1} = \mathbf{v}_k - \varphi_k \nabla \mathcal{R}(\mathbf{v}_k)h - \varphi_k h \boldsymbol{\xi}_k$. For sufficiently small h, this iteration is well approximated by the Itô-type SDE [31, 45]:

$$\mathrm{d}\bar{\mathbf{v}}_\tau = -\varphi(\tau)\nabla\mathcal{R}(\bar{\mathbf{v}}_\tau)\,\mathrm{d}\tau + \varphi(\tau)\sqrt{\frac{h}{b(\tau)}\boldsymbol{\Sigma}(\bar{\mathbf{v}}_\tau)}\,\mathrm{d}\mathbf{B}_\tau, \tag{5}$$

where $\mathbf{B}_\tau \in \mathbb{R}^M$ is an M-dimensional Brownian motion, and $\varphi(\cdot)$ is the continuous-time LRS satisfying $\varphi(kh) = \eta_k/h$ for all $k \in \mathbb{N}$; $b(\cdot)$ is the continuous-time batch-size schedule satisfying $b(kh) = B_k$ for all $k \in \mathbb{N}$. In (5), the learning rate affects both the drift and diffusion terms, thereby coupling the deterministic and stochastic effects.

Intrinsic-time reparametrization. In SDE (5), the physical time τ serves as the continuous analogue of the discrete step index k. However, when the learning rate varies over time, the actual training progress is determined not by the number of updates k but by the accumulated step size $\sum_{j=1}^{k}\eta_j$, which more faithfully reflects the total optimization effort. Motivated by this observation, we introduce an *intrinsic time* variable that *rescales* the physical time τ according to the LRS:

$$t = T(\tau) := \int_0^\tau \varphi(r)\,\mathrm{d}r, \tag{6}$$

which measures the LRS-adjusted training duration. Let $\boldsymbol{\nu}_t = \bar{\mathbf{v}}_{T^{-1}(t)}$. Applying Øksendal's time change formula [44] to the SDE (5) yields the **intrinsic-time SDE**:

$$\mathrm{d}\boldsymbol{\nu}_t = -\nabla\mathcal{R}(\boldsymbol{\nu}_t)\,\mathrm{d}t + \sqrt{\gamma(t)\,\boldsymbol{\Sigma}(\boldsymbol{\nu}_t)}\,\mathrm{d}\mathbf{B}_t \quad \text{with} \quad \gamma(t) = \frac{h\varphi(T^{-1}(t))}{b(T^{-1}(t))}. \tag{7}$$

Here $\gamma(t)$ quantifies the joint effect of learning-rate and batch-size scheduling. Compared with (5), the LRS dependence is absorbed from the drift and retained only in the diffusion term, thereby **decoupling the deterministic and stochastic effects**. This structural simplification greatly facilitates the subsequent scaling analysis.

For a clearer explanation of the connection between the discrete SGD (4) and the SDE formulations (5) and (7), we refer the reader to Appendix A.4.

4 Intrinsic-Time Functional Scaling Laws

In this section, we present our main results on the Functional Scaling Law (FSL). All proofs are deferred to Appendix E. We begin with assumptions on the learning-rate schedule and model size.

Assumption 4.1. Suppose Assumptions 2.1, 2.3 and 2.4 hold. Assume both M and $N-M$ are sufficiently large, and the LRS satisfies $\sup_{t\geqslant 0}\gamma(t) \leqslant C_3$ for a sufficiently small constant $C_3 > 0$.

Theorem 4.2 (Intrinsic-Time FSL, top-M features, hard-regime)**.** *Under Assumption 4.1, let $\boldsymbol{\nu}_t$ denote the solution to the intrinsic-time SDE (7) with top-M features. Then, for f^* with difficulty $s \in (0, 1-1/\beta]$ and any $\sigma \geqslant 0$, it holds for all $t \geqslant 0$ that*

$$\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \tfrac{1}{2}\sigma^2 \eqsim M^{-s\beta} + e(t) + \int_0^t \mathcal{K}(t-z)[e(z)+\sigma^2]\gamma(z)\,\mathrm{d}z, \tag{8}$$

where $e(t) := (1+t)^{-s}$, $\mathcal{K}(t) := (1+t)^{-(2-1/\beta)}$.

This theorem establishes that, for hard tasks with $s \leqslant 1-1/\beta$, the loss dynamics are fully characterized by the FSL (8). We explain the emergence of power laws in FSL from a multi-task learning perspective in Appendix A.5. Moreover, each term in the FSL (8) admits a clear interpretation:

- **Irreducible error: $\frac{1}{2}\sigma^2$.** This term is due to label noise.
- **Approximation error: $M^{-s\beta}$.** This term corresponds to the error due to finite model size, with the scaling efficiency is determined by the task's intrinsic difficulty $s\beta$.
- **Signal learning: $e(t)$.** This term corresponds to learning under full-batch gradient descent, capturing the rate at which SGD extracts the signal f^*. Moreover, the rate depends on the task's relative difficulty s. For a fixed target f^* (fixed $\alpha = s\beta$), increasing model capacity (smaller β) accelerates its convergence since $s = \alpha/\beta$ becomes larger.

- **Noise accumulation:** $\int_0^t \mathcal{K}(t-z)[e(z)+\sigma^2]\gamma(z)\,\mathrm{d}z$**.** This term characterizes how the learning-rate and batch-size schedules shape the accumulation and dissipation of stochastic noise. The integrand $[e(z)+\sigma^2]\gamma(z)$ represents the instantaneous noise magnitude, where $e(z)$ captures mini-batch noise and σ^2 captures label noise. The **forgetting kernel** $\mathcal{K}(\cdot)$ quantifies how noise injected at time z still affects the loss at time t. Due to $\mathcal{K}(t) \asymp t^{-(2-1/\beta)}$, a higher-capacity model (smaller β) tends to forget noise more slowly.

Notably, the last two terms together constitute the optimization error and two key factors govern the trade-off between the them: (i) **Model capacity:** Increasing model capacity ($\beta \downarrow$) accelerates signal learning but simultaneously slows noise forgetting. (ii) **Learning-rate and batch-size schedules:** Smaller learning rates or larger batch sizes suppress noise injection but also shorten the intrinsic training time. However, sufficient intrinsic time is important: the signal-learning term requires it to effectively reduce the risk, while the noise-forgetting term relies on it to forget noise memorized in early training. Hence, effective schedules must balance these competing objectives—*suppressing injected noise while maintaining enough intrinsic time for both learning and forgetting.*

4.1 General Results

The FSL (8) is established for the hard-learning regime where $s \leqslant 1 - 1/\beta$. We now show that an analogous FSL also holds in the general case. To state the result, we define $e_M(t) = \sum_{j=1}^M \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t}, \mathcal{K}_M(t) = \sum_{j=1}^M \lambda_j^2 e^{-2\lambda_j t}$. One can verify that both functions exhibit power-law decay for $1 \lesssim t \lesssim M^\beta$:

$$e_M(t) \asymp t^{-s}, \qquad \mathcal{K}_M(t) \asymp t^{-(2-1/\beta)}, \qquad 1 \lesssim t \lesssim M^\beta. \tag{9}$$

Consequently, $e_\infty(t) \eqsim e(t)$ and $\mathcal{K}_\infty(t) \eqsim \mathcal{K}(t)$ for $t \geqslant 0$.

The following theorem provides a characterization of the loss dynamics for general case:

Theorem 4.3 (Intrinsic-Time FSL, top-M features, general label noise)**.** *Suppose Assumption 4.1 holds. Let $\boldsymbol{\nu}_t$ denote the solution to the intrinsic-time SDE (7) with the top-M features. Define $\mathcal{F}_M(t;\gamma) = e_M(t) + \int_0^t \mathcal{K}_M(t-z)[e_M(z)+\sigma^2]\gamma(z)\,\mathrm{d}z$. There exists a $c > 0$ such that for $0 \leqslant t \leqslant cM^\beta$, it holds that*

$$\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \tfrac{1}{2}\sigma^2 \;\eqsim\; M^{-s\beta} + \mathcal{F}_\infty(t;\gamma). \tag{10}$$

For all $cM^\beta \leqslant t < \infty$, it holds that

$$M^{-s\beta} + \mathcal{F}_M(t;\gamma) \;\lesssim\; \mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \tfrac{1}{2}\sigma^2 \;\lesssim\; M^{-s\beta} + \mathcal{F}_\infty(t;\gamma). \tag{11}$$

Notably, the constants implicit in $\eqsim, \lesssim$ are independent of the noise level σ.

A proof sketch is provided in Appendix D. The above characterization is *uniform* with respect to the label-noise level σ, and holds for all $s > 0$ and $\beta > 1$. It asserts that the exact FSL relation (10) (i.e., the FSL (8)) remains valid up to the intrinsic time $t \leqslant cM^\beta =: t_M$. For later times $t > t_M$, although the FSL may no longer hold exactly, the loss dynamics remain controlled from both sides as in (11).

At the critical time t_M, we have $e_M(t_M) \asymp M^{-s\beta}$, indicating that signal learning has reached the approximation-error limit. Beyond this point, further training no longer improves the learned signal; instead, the dynamics become dominated by noise effects. Depending on the interaction between the stochastic gradient noise and the decaying learning rate, additional training may either inject more noise or dissipate it. Thus, it is a priori unclear whether the total error will significantly increase or decrease after t_M. Nevertheless, the upper bound in (11) ensures that the overall loss remains well-controlled, analogous to the behavior of the infinite-width limit ($M = \infty$).

Nevertheless, an FSL may still hold for all $t \geqslant 0$, under suitably stronger conditions. In Theorem 4.2, we considered the setting with tasks satisfying $s \leqslant 1 - 1/\beta$. The following result shows that a similar characterization extends to general task difficulty with constant label noise.

Theorem 4.4 (Intrinsic-Time FSL, top-M features, constant label noise)**.** *Under Assumption 4.1, suppose $\sigma \gtrsim 1$. Let $\boldsymbol{\nu}_t$ denote the solution to the intrinsic-time SDE (7) with the top-M features. Then, for any $s > 0$ and all $t \geqslant 0$, $\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \frac{1}{2}\sigma^2 \;\eqsim\; M^{-s\beta} + \mathcal{F}_M(t;\gamma)$.*

Theorem 4.2 implies that the finite-M functions e_M and $\mathcal{K}_M$ can be replaced by their infinite-width counterparts e_∞ and $\mathcal{K}_\infty$ in the hard-learning regime. The next result demonstrates that the same FSL characterization naturally extends to the noiseless case $\sigma = 0$.

Theorem 4.5 (Intrinsic-Time FSL, top-M features, zero label noise)**.** *Suppose Assumption 4.1 holds and $\sigma = 0$. Let $\boldsymbol{\nu}_t$ denote the solution to the intrinsic-time SDE (7) with the top-M features. If $s \in [0, 2 - 1/\beta]$, then for all $t \geqslant 0$, $\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] \eqsim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-z)\, e_M(z)\, \gamma(z)\, \mathrm{d}z$.*

Random-M features. The next theorem establishes that the same FSL characterization also holds when the top-M features are replaced by randomly selected features.

Theorem 4.6 (Intrinsic-Time FSL, random-M features)**.** *Suppose Assumption 4.1 holds and $s \in (0, 1]$. Let $\boldsymbol{\nu}_t$ denote the solution to the intrinsic-time SDE (7) with the random-M features. Then, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the projection matrix $\mathbf{W}$, the results of Theorems 4.2, 4.3, 4.4, and 4.5 continue to hold.*

This theorem implies that when the task difficulty satisfies $s \leqslant 1$, training with random-M features is equivalent to using the top-M features, up to exponentially small probability. We emphasize, however, that for easier tasks with $s > 1$, the behaviors of random and top feature may diverge—an analysis we leave to future work.

5 Learning Rate Schedules Impact Scaling Efficiency

Having established the general FSL, we now instantiate it under three representative LRSs—constant, exponential decay, and warmup–stable–decay (WSD)—to examine how schedule design influences scaling efficiency. All proofs can be found in Appendix F. For clarity, we make:

Assumption 5.1. Assume constant label noise $\sigma^2 \gtrsim 1$ and batch size $b(\tau) = B$ for all $\tau \geqslant 0$.

Under this assumption, given a physical-time LRS function $\varphi(\cdot)$, Theorem 4.4 implies that the FSL for $t \gtrsim 1$ simplifies to $\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \frac{1}{2}\sigma^2 \eqsim M^{-s\beta} + e_M(t) + \frac{\sigma^2}{B}\int_0^t \mathcal{K}_M(t-r)\varphi(T^{-1}(r))\,\mathrm{d}r$.

Let $\mathcal{E}_K = \mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_{Kh})] - \frac{1}{2}\sigma^2$ denote the expected excess risk after K training steps. For each LRS, we derive concrete scaling laws describing how $\mathcal{E}_K$ scales with the model size M, the total step count K, as well as the LRS's hyperparameters. We then reinterpret these results from a resource-allocation perspective by optimizing under two canonical constraints: (i) the data-limited regime, where the total data size $D := BK$ is fixed; and (ii) the compute-limited regime [20], where the total compute $C := MD$ is fixed. For each regime, we further examine how the **optimally tuned hyperparameters** (e.g., the peak learning rate) should scale with increasing available resources.

Finally, for clarity, we distinguish between two task regimes: an **easy-learning regime**, where $s \geqslant 1 - 1/\beta$, and a **hard-learning regime**, where $s < 1 - 1/\beta$.

5.1 Constant LRS

Theorem 5.2 (Scaling law for constant LRS)**.** *Under Assumption 5.1, we have $\mathcal{E}_K \eqsim M^{-s\beta} + (\eta K)^{-s} + \frac{\eta}{B}\sigma^2$.*

Let $\gamma := \eta/B$ be the *effective learning rate*. Then, the scaling law can be rewritten as $\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma D)^{-s} + \gamma\sigma^2 =: h(\gamma, M, D)$, where the excess risk depends the learning rate via $\gamma = \eta/B$. This suggests that we should scale the learning rate linearly with respect to batch size (a.k.a. linear scaling rule) [26, 16, 40].

Data-optimal scaling. Clearly, this involves minimizing $h(\cdot)$ while keeping D fixed. A straightforward calculation yields: $\gamma_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}, M_{\text{opt}} \gtrsim D^{\frac{1}{(1+s)\beta}}, \mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}$.

Notably, both the best achievable excess risk $\mathcal{E}_{\text{opt}}$ and optimal learning rate γ_{opt} depend exclusively on the task's relative difficulty s. For a fixed target (fixed α), a higher-capacity model (smaller β) gives a larger $s = \alpha/\beta$ and is therefore more data-efficient.

Compute-optimal scaling. This involves minimizing $h(\cdot)$ while keeping $C := DM$ fixed. The solution is summarized as follows, with the derivation deferred to Appendix F.1: $\gamma_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+(s+1)\beta}}$, $M_{\text{opt}} \eqsim C^{\frac{1}{1+(s+1)\beta}}$, $D_{\text{opt}} \eqsim C^{\frac{(s+1)\beta}{1+(s+1)\beta}}, \mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+s\beta+\beta}}$.

This shows that the performance of the compute-optimal model improves with the total compute budget C in a power law. For a fixed task ($\alpha = s\beta$ fixed), we have the following observations:

- Increasing model capacity ($\beta \downarrow$) enhances compute efficiency—the extra β in the scaling exponent $\frac{s\beta}{1+s\beta+\beta}$ quantifies this gain. This explains a well-known empirical observation in LLM pre-training: Large models are more compute-efficient than small models [25, 20].
- The optimal learning rate γ_{opt} decreases as C grows, and the compute-optimal allocation favors investing more in data than in model size—again consistent with current LLM pre-training practice [5, 54, 20].

Note that [8] also investigated compute-optimal scaling for constant LRS but assumed a fixed learning rate and no label noise. In contrast, we consider a more realistic scenario where the learning rate is optimally tuned and the irreducible risk is present, leading to a compute-optimal scaling law that matches empirical observations.

5.2 Exponential Decay LRS

For a given number of training steps K [15, 64], an exponential decay (exp-decay) LRS is given by $\varphi(\tau) = a\exp(-\lambda\tau), \varphi(Kh) = b$, where λ is chosen such that $\varphi(Kh) = b$. For brevity, we assume $h = 1$. Note that the hyperparameters a and b specify the peak and final learning rates, respectively.

Theorem 5.3 (Scaling law for exp-decay LRS). *Under Assumption 5.1, we have* $\mathcal{E}_K \eqsim M^{-s\beta} + T^{-s} + \sigma^2\left(\frac{b}{B} + (a-b)\frac{\min\{M, T^{1/\beta}\}}{BT}\right)$, *where* $T = (a-b)K/\log(a/b)$ *is the total intrinsic time.*

Let $b = a/K$. Then the intrinsic time becomes $T = a(K-1)/\log K$, whereas a constant LRS with step size $\eta = a$ yields $T = aK$. Thus, exp-decay LRS drives the learning rate down to as small as a/K, yet sacrifices only a logarithmic factor of intrinsic time compared to the constant schedule.

Data-optimal scaling. Let $\gamma = a/B$ be the effective peak learning rate. By minimizing the right hand side of the exponential decay scaling law with respect to a, b, K, B, M under the constraint $KB = D$ (see Appendix F.2), We obtain $M_{\text{opt}} = \infty$ and

- For $s \geqslant 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim (D/\log D)^{-\frac{1+s\beta-\beta}{1+s\beta}}$ and $\mathcal{E}_{\text{opt}} \eqsim (D/\log D)^{-\frac{s\beta}{s\beta+1}}$.
- For $s < 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim 1$ and $\mathcal{E}_{\text{opt}} \eqsim (D/\log D)^{-s}$.

Compared with the constant LRS, exp-decay LRS achieves a strictly faster decay of the excess risk, justifying the importance of learning-rate decay in stochastic optimization.

Compute-optimal scaling. A straightforward calculation (see Appendix F.2) yields:

- For $s \geqslant 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim (\frac{C}{\log C})^{-\frac{1+s\beta-\beta}{2+s\beta}}, M_{\text{opt}} \eqsim (\frac{C}{\log C})^{\frac{1}{2+s\beta}}, D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}}(\log C)^{\frac{1}{2+s\beta}}$, and $\mathcal{E}_{\text{opt}} \eqsim (\frac{C}{\log C})^{-\frac{s\beta}{2+s\beta}}$.
- For $s < 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim 1, M_{\text{opt}} \eqsim (\frac{C}{\log C})^{\frac{1}{1+\beta}}, D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}}(\log C)^{\frac{1}{1+\beta}}, \mathcal{E}_{\text{opt}} \eqsim (\frac{C}{\log C})^{-\frac{s\beta}{1+\beta}}$.

In the easy-learning regime, the excess-risk rate is determined solely by the intrinsic difficulty $\alpha = s\beta$; hence, increasing model capacity alone does not lead to asymptotic gains. The compute-optimal allocation consistently favors data over model and moreover, the optimal compute split depends solely on the task's intrinsic difficulty, with ratio $D_{\text{opt}}/M_{\text{opt}} \eqsim C^{\alpha/(2+\alpha)}$ *decreasing* as the task becomes harder. This implies that, for harder tasks, one should allocate more compute to increasing model size. The optimal γ_{opt} decreases with the compute budget C, and for fixed α, higher-capacity models ($\beta \downarrow$) require smaller γ_{opt}.

In the hard-learning regime, data still dominates compute allocation, but now the optimal split depends only on model capacity, independent of the task difficulty. Moreover, the optimal maximal learning rate remains constant ($\gamma_{\text{opt}} \eqsim 1$). These results imply that a single, universal choice of compute split and learning rate suffices to attain optimal scaling across all tasks satisfying $s < 1 - 1/\beta$, greatly simplifying hyperparameter tuning. Finally, in this regime, higher-capacity models (smaller β) become strictly more compute-efficient, as evidenced by the excess-risk scaling exponent $-s\beta/(1+\beta)$.

5.3 WSD-like LRS

We lastly turn to consider a WSD-like LRS [68, 21], which comprises a K_1-step **stable phase** followed by a K_2-step **decay phase**, for a total $K = K_1 + K_2$ steps, given by

$$\varphi(\tau) = \begin{cases} a & , \text{ if } \tau \leqslant K_1 h; \\ a \exp(-\lambda(\tau - K_1 h)) & , \text{ if } \tau > K_1 h. \end{cases} \tag{12}$$

where λ is chosen such that $\varphi(Kh) = b$. For brevity, we assume $h = 1$ and let $r = K_2/K$. This schedule is thus characterized by three hyperparameters: the peak learning rate a, the final learning rate b, and the decay proportion r, which controls the duration of decay-phase. (The warmup phase is omitted, as it does not affect our analysis.)

Theorem 5.4 (Scaling law for WSD-like LRS)**.** *Under Assumption 5.1, we have for the LRS (12):* $\mathcal{E}_K \eqsim M^{-s\beta} + (T_1 + T_2)^{-s} + \sigma^2 \left(\frac{b}{B} + (a-b) \frac{\min\{M, T_2^{1/\beta}\}}{BT_2} \right)$, *where* $T_1 = aK_1$ *and* $T_2 = (a-b)K_2/\log(a/b)$ *denote the intrinsic training times of the stable and decay phases, respectively.*

We see that WSD-like LRS can leverage the initial stable phase to boost the intrinsic training time. For a decay proportion $r < 1$, we have $T = T_1 + T_2 \geqslant (1-r)Ka$, which far exceeds the the intrinsic time $T \eqsim aK/\log K$ achieved by the pure exp-decay LRS. Consequently, WSD removes logarithmic factors in the full-batch GD term, without altering the noise term's order as long as $r > 0$. Building on this insights, we show that WSD can indeed improve the scaling efficiency, as detailed below.

Data-optimal scaling. Assuming $b = a/K$, we have $M_{\text{opt}} = \infty$ and

- For $s \geqslant 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim D^{-\frac{1+s\beta-\beta}{1+s\beta}} (\log D)^{\frac{\beta-1}{1+s\beta}}$, $r_{\text{opt}} \in (0,1)$, $\mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s\beta}{s\beta+1}} (\log D)^{\frac{s\beta-s}{1+s\beta}}$.
- For $s < 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim 1$, $r_{\text{opt}} \gtrsim D^{\frac{s\beta+1-\beta}{\beta-1}} \log D$, $\mathcal{E}_{\text{opt}} \eqsim D^{-s}$.

Compared with the exp-decay LRS, both regimes enjoy a logarithmic improvement in excess-risk decay. In particular, for the hard-learning regime, the logarithmic factor disappears. This improvement requires the **decay-phase duration only needs to scale sublinearly with** D, as indicated by $r_{\text{opt}} \to 0$ as $D \to \infty$. This matches the WSD practice in LLM pre-training, where the decay phase typically occupies only 10%-20% of the total training duration. Moreover, our theory suggests that for harder tasks, the decay fraction can be reduced further to enhance compute efficiency.

Compute-optimal scaling. Analogous improvements hold in the compute-limited regime. Assuming $b = a/K$ and imposing the compute constraint $MD = C$, the compute-optimal satisfies:

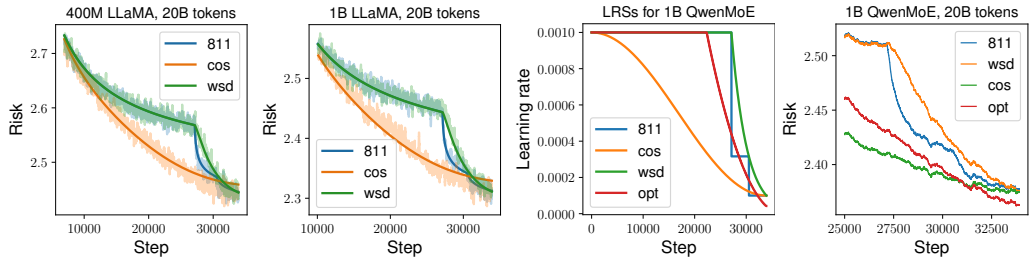
- For $s \geqslant 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim (\frac{C}{\log C})^{-\frac{1+s\beta-\beta}{2+s\beta}}$, $r_{\text{opt}} \in (0,1)$, $M_{\text{opt}} \eqsim (\frac{C}{\log C})^{\frac{1}{2+s\beta}}$, $D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1}{2+s\beta}}$, and $\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{2+s\beta}} (\log C)^{\frac{s\beta-s}{2+s\beta}}$.
- For $s < 1 - \frac{1}{\beta}$, $\gamma_{\text{opt}} \eqsim 1$, $r_{\text{opt}} \gtrsim D^{-\frac{\beta-1-s\beta}{\beta-1}} \log D$, $M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta}}$, $D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}}$, $\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta}}$.

6 Experiments

6.1 Power-Law Kernel Regression

While the FSL is derived in the continuous-time limit, we now verify that it also accurately captures the loss dynamics and scaling behavior of the discrete-time SGD (3). Specifically, we consider the PLK regression with difficulty $s = 0.5$ and capacity $\beta = 4$, corresponding to a hard-learning regime and the results are shown in Figure 1.

FSL accurately captures the loss dynamics of SGD. Figure 1(left) compares the loss dynamics of SGD with the predictions of the FSL under three representative LRSs: cosine, WSD, and an unconventional cyclical schedule [55]. Across all cases, the FSL provides a remarkably accurate description of the SGD's loss evolution. Comparing the WSD and cosine schedules, we observe that the loss under WSD exhibits a slower decay during the stable phase but undergoes a much sharper drop once the decay phase begins, ultimately yielding a lower final loss. This seemingly counterintuitive two-phase dynamical behavior of WSD aligns well with empirical observations in practical LLM pre-training [21, 63, 57].



(a) Fitting and prediction using FSL. (b) The FSL-optimal LRS and its performance

Figure 2: **Experiment on LLMs. (a)** Fitting and predictive accuracy of the FSL on dense LLaMA models. **(b)** Left: comparison of various LRSs. Right: loss trajectories of the FSL-optimal schedule versus baseline LRSs on a 1B QwenMoE model.

FSL predicts the scaling behavior of SGD. Figure 1(right) further validates the scaling laws derived in Section 5 for the three canonical LRSs—constant, exponential, and WSD-like (12). The results show that the final-step loss of SGD closely follows the theoretical predictions of FSL. Among these schedules, WSD yields the best scaling performance, followed by exponential decay, while the constant schedule performs the worst. More experiment details and additional results experiments with varying (s, β) and other LRSs are provided in Appendix C.1, and exhibit consistent behaviors.

6.2 LLM Pre-training

We now evaluate the practical utility of FSL as a surrogate model for capturing the loss dynamics of LLM pre-training. Specifically, three popular LRSs: cosine, WSD, and the 8-1-1 [5] are considered; see Figure 2b(left) for a visualization. In the 8-1-1 LRS, the learning rate is reduced by a factor $\sqrt{10}$ at 80% and 90% of the total token budget, yielding a final value that is 0.1 times the peak learning rate. For more experiment details, we refer to Appendix C.2.

FSL accurately fit and predict loss curves. We first quantify the descriptive and predictive power of FSL. Following the protocol of [58] and [38], we restrict attention to the post-warmup portion of the loss trajectory. Two Llama [59] models (400 M and 1 B) are trained on 20 B tokens under the three LRSs. For each model we (i) fit the FSL parameters on the loss curve obtained using the 8-1-1 LRS and (ii) deploy the fitted FSL to *predict* the loss curves of the cosine and WSD schedules. Figure 2a demonstrates that FSL not only fits the 8-1-1 trajectory accurately but also generalizes reliably to the unseen WSD and cosine schedules for both model sizes.

The FSL-optimal LRS is WSD-like. We next leverage the fitted FSL to *design* improved LRSs. Specifically, we numerically minimize the final-step loss over the space of LRSs using the fitted FSL. This experiment employs a 1B-parameter QwenMoE model [67], trained on 20B tokens using the same three LRSs. We fit the FSL using the trajectory from the 8-1-1 LRS and numerically solve for the FSL-optimal LRS. The model is then trained under this FSL-optimal LRS, using the same compute budget, and compared against the baseline LRSs. Figure 2b(left) shows that surprisingly, the FSL-optimal LRS is WSD-like and the decay phase drives the learning rate far below the conventional $0.1\eta_{\max}$ threshold. This echos recent empirical recommendations by [4, 17]. Furthermore, Figure 2b(right) demonstrates that the FSL-optimal schedule yields a strictly lower final loss than all baselines, substantiating its practical relevance. Taken together, these results suggest that FSL is a faithful surrogate for studying LLM training dynamics and a principled tool for interpreting and designing LRSs in large-scale pre-training.

7 Conclusion

In this paper, we present a systematic study of how LRS shapes the loss dynamics in kernel regression. Specifically, we establish a novel functional-level scaling law, which precisely characterizes the loss dynamics of SGD for general learning LRSs. The utility of our FSL is demonstrated through detailed analyses of three widely used LRSs, providing theoretical justification for several prevailing practices in LLM pre-training—most notably, offering an explanation for the effectiveness of the empirically popular but previously less-understood WSD schedules.

Acknowledgement

Lei Wu is supported by the National Natural Science Foundation of China (NSFC12522120, NSFC92470122, and NSFC12288101). Binghui Li is supported by the Elite Ph.D. Program in Applied Mathematics at Peking University. We are grateful to Kaifeng Lyu, Kairong Luo, and Haodong Wen for generously sharing their work [38], which greatly inspired this study. We also thank Tingkai Yan, Yuhao Liu, Yunze Wu, and Zean Xu for many helpful discussions, and the anonymous reviewers for their valuable feedback.

References

[1] 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*, pages 265–279. PMLR, 2023. (cited on page 2)

[2] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024. (cited on pages 2 and 19)

[3] 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. (cited on pages 2, 4, and 19)

[4] 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. (cited on pages 2 and 10)

[5] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and Others. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. (cited on pages 2, 4, 8, 10, and 17)

[6] Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal lr across token horizons. In *The Thirteenth International Conference on Learning Representations*. (cited on page 3)

[7] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024. (cited on pages 2, 3, 4, and 19)

[8] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *arXiv preprint arXiv:2409.17858*, 2024. (cited on pages 2, 3, 4, 8, and 19)

[9] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020. (cited on page 4)

[10] Sébastien Bubeck and Others. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. (cited on page 23)

[11] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007. (cited on page 4)

[12] Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm. 2005. (cited on page 4)

[13] Shihong Ding, Haihan Zhang, Hanzhen Zhao, and Cong Fang. Scaling law for stochastic gradient descent in quadratically parameterized linear regression. *arXiv preprint arXiv:2502.09106*, 2025. (cited on page 4)

[14] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024. (cited on pages 2 and 19)

[15] 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. (cited on pages 3 and 8)

[16] P Goyal. Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. (cited on page 7)

[17] Alex Hägele, Elie Bakouch, Atli Kosson, Leandro Von Werra, Martin Jaggi, and Others. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024. (cited on pages 2, 10, 17, and 20)

[18] 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. (cited on page 2)

[19] 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. (cited on page 1)

[20] 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. (cited on pages 2, 3, 7, 8, and 16)

[21] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and Others. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. (cited on pages 3, 4, 9, 17, and 20)

[22] Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021. (cited on pages 2 and 19)

[23] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint arXiv:2402.04376*, 2024. (cited on pages 2 and 19)

[24] Arlind Kadra, Maciej Janowski, Martin Wistuba, and Josif Grabocka. Power laws for hyperparameter optimization. *arXiv preprint arXiv:2302.00441*, 2023. (cited on page 2)

[25] 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. (cited on pages 1, 2, 3, and 8)

[26] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014. (cited on page 7)

[27] 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. (cited on page 2)

[28] Kiwon Lee, Andrew Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of mini-batch momentum: batch size saturation and convergence in high dimensions. *Advances in Neural Information Processing Systems*, 35:36944–36957, 2022. (cited on page 26)

[29] Houyi Li, Wenzhen Zheng, Jingcheng Hu, Qiufeng Wang, Hanshan Zhang, Zili Wang, Shijie Xuyang, Yuantao Fan, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale: Part I – optimal hyperparameter scaling law in large language model pretraining. *arXiv preprint arXiv:2503.04715*, 2025. (cited on pages 2 and 3)

[30] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017. (cited on page 4)

[31] 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. (cited on pages 4 and 5)

[32] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020. (cited on page 4)

[33] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021. (cited on page 4)

[34] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022. (cited on page 4)

[35] 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. (cited on pages 2, 3, 4, 19, 33, and 34)

[36] 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. (cited on pages 2, 3, and 20)

[37] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. (cited on pages 4 and 17)

[38] 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. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024. (cited on pages 2, 10, 11, 17, and 20)

[39] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022. (cited on pages 2, 4, and 19)

[40] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018. (cited on pages 2 and 7)

[41] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. (cited on page 3)

[42] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024. (cited on pages 2 and 19)

[43] Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, and Ard Louis. An exactly solvable model for emergence and scaling laws. *arXiv preprint arXiv:2404.17563*, 2024. (cited on pages 2, 4, and 19)

[44] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003. (cited on page 5)

[45] Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on page 5)

[46] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021. (cited on page 26)

[47] Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. *Advances in Neural Information Processing Systems*, 34:9229–9240, 2021. (cited on page 26)

[48] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, pages 1–90, 2024. (cited on page 26)

[49] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024. (cited on pages 2, 3, 4, 19, and 26)

[50] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021. (cited on page 4)

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and Others. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. (cited on page 23)

[52] 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. (cited on page 20)

[53] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020. (cited on pages 2 and 19)

[54] Xian Shuai, Yiding Wang, Yimeng Wu, Xin Jiang, and Xiaozhe Ren. Scaling law for language models training considering batch size. *arXiv preprint arXiv:2412.01505*, 2024. (cited on pages 2 and 8)

[55] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. (cited on page 9)

[56] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020. (cited on page 4)

[57] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and Others. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. (cited on pages 3, 9, and 20)

[58] Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *arXiv preprint arXiv:2408.11029*, 2024. (cited on pages 2, 10, 16, 20, and 23)

[59] 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. (cited on pages 2, 4, 10, 17, and 23)

[60] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. (cited on page 38)

[61] Mingze Wang and Lei Wu. A theoretical analysis of noise geometry in stochastic gradient descent. *arXiv preprint arXiv:2310.00692*, 2023. (cited on page 26)

[62] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, pages 23549–23588. PMLR, 2022. (cited on pages 2 and 19)

[63] 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. (cited on pages 9 and 20)

[64] 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*, pages 24280–24314. PMLR, 2022. (cited on page 8)

[65] Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pages 37656–37684. PMLR, 2023. (cited on page 28)

[66] 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, 2022. (cited on pages 26 and 28)

[67] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. (cited on pages 10 and 23)

[68] 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*, pages 12104–12113, 2022. (cited on pages 3, 9, 17, and 20)

[69] Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? *arXiv preprint arXiv:2410.21676*, 2024. (cited on page 2)

Appendix

Table of Contents

A Miscellanea

A.1 Empirical Fitting of LLM Pre-training Loss Trajectory

The Chinchilla Law [20] describes the final-step loss $\mathcal{L}$ as follows:

$$L(M, D) = L_0 + A_1 M^{-\kappa_1} + A_2 D^{-\kappa_2},$$

where L_0, A_1, A_2, κ_1, and κ_2 are constants, and D and M represent the amount of training data (tokens) and model size (number of parameters), respectively.

Later, [58] proposed the **Momentum Law**, a heuristic rule designed to capture the full loss trajectory. Given a learning rate schedule $\boldsymbol{\eta} := \{\eta_j\}_j$, the loss at the k-th step is modeled as

$$\mathcal{L}_k(\boldsymbol{\eta}) = L_0 + A S_1^{-\kappa} - C S_2,$$

where

$$S_1 = \sum_{i=1}^{k} \eta_i, \quad S_2 = \sum_{i=2}^{k} \sum_{j=2}^{i} (\eta_{j-1} - \eta_j) \lambda^{i-j}.$$

Here L_0, A, C, and κ are constants, and $\lambda \in (0, 1)$ is a hyperparameter representing the decay factor for learning rate annealing, which typically ranges from 0.99 to 0.999.

Subsequently, [38] proposed the **Multi-Power Law** (MPL), which replaces the S_2 in the Momentum Law with additional power laws to better capture the progressive loss reduction induced by learning-rate decay. Specifically, the MPL takes the following form:

$$\mathcal{L}_k(\boldsymbol{\eta}) = L_0 + AS_1^{-\kappa} - \mathrm{LD}(k), \tag{13}$$

where

$$\mathrm{LD}(k) := C\sum_{i=2}^{k}(\eta_{i-1}-\eta_i)G(\eta_i^{-\kappa'}S_i), \quad S_i := \sum_{j=1}^{i}\eta_j, \quad G(x) := 1-(C'x+1)^{-\kappa''}.$$

Here $L_0, A, C, C', \kappa, \kappa', \kappa''$ are constants.

A.2 Popular Learning Rate Schedules

Here, we introduce some widely used LRSs in the context of LLM pre-training.

- **Cosine Schedule [37].** The schedule is given by $\eta_k = \frac{1+\rho}{2}\eta_{\max} + \frac{1-\rho}{2}\eta_{\max}\cos(\frac{k-1}{K-1})$, where $\eta_{\max}$ is the maximum learning rate and the hyper-parameter ρ is usually chosen as 0.1 such that the minimum learning rate is $\eta_{\max}/10$ [59].
- **Warmup-Stable-Decay (WSD) Schedule [68, 21, 17].** The schedule consists of three phases: a warm-up phase of $K_{\text{warm-up}}$ steps, followed by a stable phase maintaining the learning rate $\eta_k = \eta_{\max}$, and finally a decay phase governed by $\eta_k = h(k-K_{\text{stable}})\eta_{\max}$ for $K_{\text{stable}} \leqslant k \leqslant K$, where K_{stable} represents the total duration of the first two phases. Here, the decay function $h(\cdot) \in (0,1)$ can be linear or exponential.
- **Multi-Step Schedule [5].** The entire schedule is divided into S stages, i.e., $[K_0, K_1] \cup [K_1, K_2] \cup \cdots \cup [K_{S-1}, K_S] = [0, K]$, where $0 = K_0 < K_1 < \cdots < K_S = K$. The schedule satisfies that $\eta_k = \eta_{K_i}$ for $K_{i-1} < k \leqslant K_i$ $(1 \leqslant i \leqslant S)$. In our LLM experiments, we consider a 8-1-1 LRS, corresponding to the case where $S = 3$ with $\eta_{K_1} = \eta_{\max}, \eta_{K_2} = \eta_{\max}/\sqrt{10}$, and $\eta_{K_3} = \eta_{\max}/10$, and $K_1 = 0.8K, K_2 = 0.9K$.

A.3 Connections to Kernel Regression

In this section, we explain how our setup in Section 2 are equivalent to learning with kernels.

Definition A.1 (Positive semidefinite (PSD) kernel)**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *continuous positive semidefinite (PSD) kernel* if it satisfies:

- Symmetry: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$;
- Positive semidefiniteness: for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} a_i a_j\, K(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0.$$

Definition A.2 (Reproducing kernel Hilbert space (RKHS))**.** Given a kernel K, the *reproducing kernel Hilbert space* $\mathcal{H}_K$ associated with K is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$\langle f, K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_K} = f(\mathbf{x}), \quad \forall f \in \mathcal{H}_K,\ \mathbf{x} \in \mathcal{X}.$$

Kernel methods learn functions from a hypothesis space defined by the associated RKHS. For instance, kernel ridge regression gives estimator:

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|_{\mathcal{H}_K}^2.$$

Hence, model capacity is determined by the size of the RKHS $\mathcal{H}_K$.

Let $\mathcal{D}$ be the input distribution. Given a kernel K, define the associated integral operator $\mathcal{T}_K : L^2(\mathcal{D}) \to L^2(\mathcal{D})$ by

$$\mathcal{T}_K f(\cdot) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[K(\cdot, \mathbf{x})f(\mathbf{x})].$$

By assuming $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[K(\mathbf{x},\mathbf{x})] < \infty$, the operator $\mathcal{T}_K$ is compact (Mercer's theorem) and consequently, the kernel admits the following eigenvalue decomposition

$$K(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j e_j(\mathbf{x}) e_j(\mathbf{x}'),$$

where $\{\lambda_j\}_{j=1}^{\infty}$ and $\{e_j\}_{j=1}^{\infty}$ denotes the eigenvalues and eigenfunctions, respectively. Moreover, $\langle e_i, e_j\rangle_{L^2(\mathcal{D})} = \delta_{i,j}$, i.e., the eigenfunctions form an orthonormal basis of $L^2(\mathcal{D})$.

Using the spectral decomposition, the RKHS admits the following representation:

$$\mathcal{H}_K = \left\{ \sum_{j=1}^{\infty} a_j e_j : \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\}.$$

To better quantify the smoothness of functions, we often consider the interpolation space $\mathcal{H}_K^s$ with $s \geqslant 0$, defined as

$$\mathcal{H}_K^s = \left\{ \sum_{j=1}^{\infty} a_j e_j : \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j^s} < \infty \right\}.$$

Clearly, $\mathcal{H}_K^1 = \mathcal{H}_K$, and

$$\mathcal{H}_K^{s_1} \subset \mathcal{H}_K^{s_2}, \qquad \forall\, s_1 > s_2 \geqslant 0.$$

Hence, the index s characterizes the smoothness of a function relative to the chosen kernel.

In the analysis of kernel methods, the following conditions are commonly used to describe the smoothness of the target function and the capacity of the kernel, respectively.

Assumption A.3 (Source condition)**.** There exists some $s > 0$ such that $f^* \in \mathcal{H}_K^s$.

Assumption A.4 (Capacity condition)**.** There exists some $\beta > 1$ such that $\lambda_j \eqsim j^{-\beta}$.

These conditions yield the following interpretation:

- A smaller s indicates that the target function f^* belongs to a larger space, corresponding to a more difficult learning problem.
- A smaller β implies a slower eigenvalue decay, meaning a richer hypothesis space $\mathcal{H}_K$ and thus higher model capacity.

Our formulation in Section 2 is equivalent to the above setting, but expressed in terms of the feature map $\boldsymbol{\phi}$. Under Assumption 2.2, we have

$$K_\phi(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{N} \phi_j(\mathbf{x})\phi_j(\mathbf{x}') = \sum_{j=1}^{N} \lambda_j\, \widehat{\phi}_j(\mathbf{x})\widehat{\phi}_j(\mathbf{x}').$$

In this case, Assumption 2.3 corresponds exactly to the above capacity condition, while the task-difficulty assumption in Section 2 can be viewed as a power-law version of the source condition. Specifically, under Assumption 2.4,

$$f^* = \sum_{j=1}^{N} \theta_j^* \phi_j = \sum_{j=1}^{N} j^{-1/2} \lambda_j^s\, \widehat{\phi}_j = \sum_{j=1}^{N} a_j\, \widehat{\phi}_j.$$

Hence, for any arbitrarily small $\delta \in (0,1)$, we have $f^* \in \mathcal{H}_{K_\phi}^{s-\delta}$, since

$$\sum_{j=1}^{N} \frac{a_j^2}{\lambda_j^{s-\delta}} = \sum_{j=1}^{N} j^{-1-\beta(s-\delta)} < \infty.$$

A.4 The SDE Modeling

The physical-time SDE. In our setup, the SGD update can be written as

$$\mathbf{v}_{k+1} = \mathbf{v}_k - \varphi_k \nabla \mathcal{R}(\mathbf{v}_k) h - \varphi_k h \boldsymbol{\xi}_k.$$

The term $\boldsymbol{\xi}_k$ is the gradient noise, whose covariance is $\frac{1}{B_k}\Sigma(\mathbf{v}_k)$. By assuming the gradient noise to be Gaussian, the SGD becomes

$$\mathbf{v}_{k+1} - \mathbf{v}_k = -\varphi_k \nabla \mathcal{R}(\mathbf{v}_k) h + \varphi_k \sqrt{h}\, \mathcal{N}\left(0, \tfrac{h}{B_k}\Sigma(\mathbf{v}_k)\right).$$

It is exactly the Euler–Maruyama discretization of the Itô-type SDE:

$$\mathrm{d}\bar{\mathbf{v}}_\tau = -\varphi(\tau)\nabla \mathcal{R}(\bar{\mathbf{v}}_\tau)\,\mathrm{d}t + \varphi(\tau)\sqrt{\frac{h}{b(\tau)}\Sigma(\bar{\mathbf{v}}_\tau)}\,\mathrm{d}\mathbf{B}_\tau,$$

where $\mathbf{B}_\tau \in \mathbb{R}^M$ denotes the M-dimensional Brownian motion, and $\varphi(\cdot)$, $b(\cdot)$ are the continuous version of LRS function and batch-size schedule function, respectively.

The intrinsic-time SDE. Intuitively, the discrete update (4) can be viewed as the Euler–Maruyama discretization of SDE (7) on the *non-uniform* grid $\{t_k = \sum_{j=0}^{k} \eta_j\}_{k\in\mathbb{N}}$ where the effective step size is $\Delta t_k = \eta_k$:

$$\mathbf{v}_{k+1} - \mathbf{v}_k = -\nabla \mathcal{R}(\mathbf{v}_k)(t_{k+1} - t_k) - \sqrt{t_{k+1} - t_k}\, \mathcal{N}\left(0, \tfrac{\eta_k}{B_k}\Sigma(\mathbf{v}_k)\right).$$

A.5 The Emergence of Power Laws

We illustrate how the power law emerges in our setting from a multi-task learning viewpoint. For brevity, consider the case of the top-M features and an infinitesimal learning rate, where the SDE (7) reduces to the gradient flow ODE: $\mathrm{d}\boldsymbol{\nu}_t = -\mathbf{W}^\top \mathbf{W}\mathbf{H}\boldsymbol{\nu}_t\,\mathrm{d}t$. Noting that $\mathbf{W}^\top\mathbf{W}\mathbf{H} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_M, 0, \cdots, 0\}$ is diagonal, consequently the ODE is solvable and gives the following expression of the excess risk:

$$\mathcal{R}(\boldsymbol{\nu}_t) - \tfrac{1}{2}\sigma^2 \eqsim \underbrace{\sum_{j=1}^{M} \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t} + \int_0^t}_{\text{learned sub-tasks}} + \underbrace{\sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2}_{\text{unlearned sub-tasks}}.$$

Intuitively, we can view the learning of each eigenfunction as a sub-task. Due to the limited model size, student model can at most learn the top-M eigenfunctions.

(i) **Intrinsic-time power law.** For each sub-task, the sub-task risk converges exponentially w.r.t. the intrinsic time t. However, owing to the power-law structure of λ_j, θ_j^*, the total multi-task risk exhibits a power-law decay for sufficiently large M due to $\sum_{j=1}^{M} \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t} \eqsim \int_0^1 u^{s-1} e^{-2ut}\,\mathrm{d}u \eqsim \frac{1}{t^s}$ if $M \gg 1$.

(ii) **Model-size power law.** Approximation error accounts for total risk of the $N - M$ unlearned sub-tasks, which follows a power-law decay due to $\sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2 \eqsim M^{-s\beta}$ if $N - M \gg 1$.

Summary. The emergence of power law arises from the accumulation effect, requiring both the number of learned tasks and unlearned tasks to be large (ideally infinite).

B Related Work

Theoretical explanation of scaling laws. Among the growing body of work seeking to theoretically explain scaling laws [53, 22, 39, 62, 23, 42, 43, 2, 14, 3, 7, 35, 49, 8], the most closely related are [7, 49, 8, 35], which also analyze PLK regression (often written in the equivalent linear-regression form). Specifically, [7] studies gradient flow, [49, 8] analyze SGD with a constant LRS, and [35]

considers an exponential-decay LRS. In contrast, we establish a unified scaling law applicable to general LRSs, which not only recovers these prior results as special cases but also substantially extends them by capturing the loss dynamics rather than only the final-step loss. This unification is enabled by introducing the key notion of *intrinsic time*, which more faithfully captures the effective training progress than the raw number of training steps.

Predicting loss trajectories in LLM pre-training. [58] presented the empirical evidence suggesting that full loss trajectories in LLM pre-training may be predictable. Subsequent work [38] proposed a heuristic called the multi-power law, achieving improved predictive accuracy. A detailed description of these fitting strategies is provided in Appendix A.1. Our analysis offers a theoretical explanation for these empirical findings.

Warmup-Stable-Decay (WSD) LRS. A WSD schedule [68, 21] maintains a constant learning rate for a long stable phase, followed by a learning rate decay only near the end of training. Although unconventional, WSD has become popular in LLM pre-training [21, 17] and is already deployed in training industry-scale LLMs such as `DeepSeek-V3` [36] and `Kimi-K2` [57]. Yet its mechanism remains poorly understood. While recent works [63, 52] offer partial insights, we show—perhaps surprisingly—that even *quadratic optimization*, corresponding to a kernel regression problem, already reproduces the essential advantage of WSD. Furthermore, we quantify this advantage through explicit comparisons of scaling efficiency against constant and exponential-decay schedules.

C Experiment Details and Additional Results

In this section, we present the details of our experiments as well as additional results.

C.1 Power-Law Kernel Regression

Physical-time FSL. The FSL (10) is presented in terms of intrinsic time, but in practice, it is often more convenient to use physical time (training steps). By a suitable change of times, after τ steps (equivalently, τ/h discrete steps), the FSL maintains the form (10), with adjustments:

$$t^{-s} = T(\tau)^{-s},$$
$$\mathcal{N}(\varphi, b) = \int_0^\tau \mathcal{K}(T(\tau) - T(u)) \left(e(T(u)) + \sigma^2\right) \frac{h\varphi(u)^2}{b(u)} \, \mathrm{d}u.$$

Fitting FSL on SGD Average-Risks. To validate that the Functional Scaling law (FSL) can accurately capture the risk curve of SGD, we conducted a series of SGD experiments under different configurations of s and β. Subsequently, we fitted the FSL to these risk curves. Our results demonstrate that FSL indeed provides a close fit to the SGD trajectories.

In each experiment, we adopt a PLKR configuration with $M = N = 128$, $\sigma = 3$ and employ the top-M projection matrix, thereby eliminating the approximation error term $M^{-s\beta}$. We explore a range of values for $s \in [0.5, 1]$ and $\beta \in [1.5, 5]$, encompassing both easy- ($s \geqslant 1 - 1/\beta$) and hard-learning ($s < 1 - 1/\beta$) regimes. For each parameter configuration, we execute 200 independent SGD runs with a batch size of 1 over 10,000 steps. The resulting average trajectory across these runs serves as the fitting target. The FSL fitting is performed using the physical-time FSL formulation.

$$\mathcal{E}_k = c_1 T(k)^{-s} + c_2 \sum_{i=1}^{k} \mathcal{K}(T(k) - T(i)) e(T(i)) \eta_i^2 + c_3 \sigma^2 \sum_{i=1}^{k} \mathcal{K}(T(k) - T(i)) \eta_i^2,$$

where c_1, c_2, c_3 are constants to fit, $\{\eta_i\}_{i=1}^k$ is the learning rate schedule, and $T(i) = \sum_{j=1}^{i} \eta_j$.

When fitting the SGD trajectory, we minimize the mean squared error (MSE) between the empirical risk trajectory of SGD (without the irreducible risk $\frac{\sigma^2}{2}$), denoted by $\mathcal{E}_{\text{SGD}}(k)$, and the theoretical prediction from FSL, $\mathcal{E}_k$. Formally, we solve the following optimization problem:

$$\min_{c_1, c_2, c_3} \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{E}_{\text{SGD}}(k) - \mathcal{E}_k\right)^2,$$

where K represents the total number of training steps. This minimization is performed using ordinary least squares (OLS), with the integrals in the FSL expression $\mathcal{E}_k$ evaluated numerically via quadrature methods.

We display the learning rate schedules (LRSs) used in the SGD experiments in the top-left panel of Figure 3. Complementing Figure 1 (middle and right), additional experimental results for various values of s and β are presented in Figure 3.

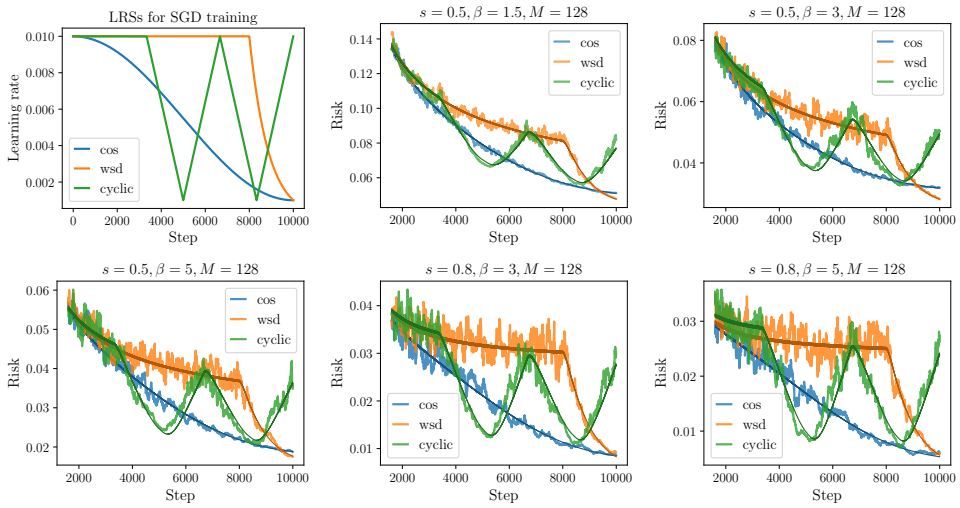


Figure 3: **Fitting results of FSL on SGD trajectories.** The shaded curves are the average over 200 independent SGD runs, while the solid curves show the predictions of FSL.

Scaling law experiments. These experiments are designed to evaluate the correctness of the scaling laws predicted by our analytical analysis. To this end, we conduct two complementary sets of experiments:

- **FSL experiments.** This experiment is intended to validate the theoretical predictions derived from the FSL. We compute the predicted risk by numerically discretizing the FSL (10), with all untracked constants set to 1. For each LRS, following the theoretical analysis, we set $\eta_{\max} = 0.05D^{-r}$, where $r = s/(1+s)$ for the constant learning rate schedule, and $r = 1$ for exponential decay and WSD schedules. We fix the batch size to $B = 1$; thus, for each data budget D, we compute the intrinsic time and evaluate the final-step loss using the discretized FSL.
- **SGD experiments.** This experiment serves to assess whether the scaling behavior predicted by our continuous-time FSL faithfully captures that of discrete-time SGD. We simulate stochastic gradient descent (SGD) with 200 independent trajectories and a fixed batch size $B = 1$. For each data budget D, the maximum learning rate is set as $\eta_{\max} = 0.05D^{-r}$, using the same theoretical values of r as in the FSL experiments. We run SGD for D steps under each corresponding LRS and record the final-step excess risk.

C.2 LLM pre-training

Practical FSL Ansatz for LLM pre-training In this section, in order to fit real LLM pre-training loss curves, we will derive an approximation form of the FSL in Theorem 4.2.

First, by the physical-time for of the FSL (10) with $h = 1$ and $B(u) \equiv B$, we have

$$\mathcal{E}_k \eqsim \frac{1}{T(k)^s} + M^{-s\beta} + \frac{1}{B}\int_0^k \mathcal{K}(T(k) - T(u))(\sigma^2 + e(T(u))) \cdot \varphi(u)^2 \,\mathrm{d}u.$$

Here we focus on the integral term. Since $\varphi(u) = \int_0^u \varphi'(r)\,\mathrm{d}r + \varphi(0)$, and that

$$\int_0^k \mathcal{K}(T(k) - T(u))(\sigma^2 + e(T(k)) \cdot \varphi(u)\varphi(0)\,\mathrm{d}u = \varphi(0)\int_0^{T(k)} \mathcal{K}(T(k) - t)(\sigma^2 + e(t))\,\mathrm{d}t. \quad (14)$$

Note that this is exactly the SGD noise term at the constant LRS $\eta(0)$ for a total intrinsic-time $T(\tau)$. By results of constant LRS (as seen in the proof of Theorem F.1), we have

$$(14) \eqsim \varphi(0)(\sigma^2 + e(T(k))).$$

As $\varphi(0) \lesssim 1$, we have

$$\mathcal{E}_k \eqsim \frac{1}{T(k)^s} + M^{-s\beta} - \mathrm{LRD}(k), \tag{15}$$

where

$$\begin{aligned}
\mathrm{LRD}(k) &:= -\frac{1}{B}\int_0^k \mathcal{K}(T(k)-T(u))(\sigma^2+e(T(u)))\varphi(u)\int_0^u \varphi'(r)\,\mathrm{d}r\,\mathrm{d}u \\
&= -\frac{1}{B}\int_0^k \varphi'(r)\int_r^k \mathcal{K}(T(k)-T(u))(\sigma^2+e(T(u)))\varphi(u)\,\mathrm{d}u\,\mathrm{d}r \\
&= -\frac{1}{B}\int_0^k \varphi'(r)\int_{T(r)}^{T(k)} \mathcal{K}(T(k)-t)(\sigma^2+e(t))\,\mathrm{d}t\,\mathrm{d}r.
\end{aligned}$$

We discretize the outer integral at integer nodes $r = 0, 1, \ldots, k$,

$$\mathrm{LRD}(k) \approx \frac{1}{B}\sum_{i=1}^{k}(\eta_{i-1}-\eta_i)\int_{T(i)}^{T(k)} \mathcal{K}(T(k)-t)(\sigma^2+e(t))\,\mathrm{d}t.$$

By the integral mean value theorem, we can take $(\sigma^2 + e(t))$ outside the integral, which gives

$$\mathrm{LRD}(k) \approx \frac{1}{B}\sum_{i=1}^{k}(\eta_{i-1}-\eta_i)(\sigma^2+e(\xi_i))\int_{T(i)}^{T(k)} \mathcal{K}(T(k)-t)\,\mathrm{d}t,$$

where $\xi_i \in [T(i), T(k)]$. Now since

$$\int_{T(i)}^{T(k)} \mathcal{K}(T(k)-t)\,\mathrm{d}t \approx \int_{T(i)}^{T(k)} \frac{1}{(1+ct)^{2-1/\beta}}\,\mathrm{d}t\,\mathrm{d}u \eqsim 1 - \frac{1}{(1+c(T(k)-T(i))^{1-1/\beta}},$$

we then further simplify it as

$$\mathrm{LRD}(k) \approx \frac{1}{B}\sum_{i=1}^{k}(\eta_{i-1}-\eta_i)(\sigma^2+e(T(i)))(1-(1+c(T(k)-T(i)))^{-\gamma}).$$

Here, we approximate ξ_i as $T(i)$ and introduce a new parameter γ to replace $1-\frac{1}{\beta}$ for simplicity.

Therefore, combining with (15), when the batch size B is fixed, after renaming some constants, the final discrete ansatz can be written as

$$\begin{aligned}
\mathcal{R}_k \approx\ & c_0 + \frac{c_1}{T(k)^s} + c_2 M^{-s\beta} \\
& - c_3 \sum_{i=1}^{k}(\eta_{i-1}-\eta_i)\left(c_4 + \frac{1}{T(i)^s}\right)\left(1-(1+c_5(T(k)-T(i)))^{-\gamma}\right),
\end{aligned} \tag{16}$$

where c_0, c_1, c_2, c_3, c_4, c_5, s, β, γ are constants to fit.

Fitting the Practical FSL The objective of this experiment is to analyze and fit the loss function using our functional scaling law, by (16), since we do not explore the effect of varying the model size M in our experiments, we drop the term $M^{-s\beta}$ and get

$$\mathcal{L}_\Theta(k) = L_0 + \frac{c_1}{T(k)^s} - \mathrm{LRD}(k)$$

where $T(k) = \sum_{i=1}^{k}\eta_i$ and

$$\mathrm{LRD}(k) := c_2 \sum_{i=1}^{k}(\eta_{i-1}-\eta_i)\left(c_3 + \frac{1}{T(i)^s}\right)\left(1 - \frac{1}{(1+c_4(T(k)-T(i)))^{\gamma}}\right),$$

and $\Theta = (L_0, c_1, c_2, c_3, c_4, s, \gamma)$.

Following [58], we utilize the Huber loss as the objective function.

$$\min_{\Theta} \sum_{k=1}^{K} \text{Huber}_{\delta} \left(\log \mathcal{L}_{\Theta}(k) - \log \mathcal{L}_{\text{gt}}(k) \right),$$

where $\delta = 1 \times 10^{-3}$, $\mathcal{L}_{\text{gt}}$ denotes the ground truth of the validation losses. We adopt the Adam optimizer, with a learning rate of 5×10^{-2} for the index parameters in our law and 5×10^{-3} for the coefficient or constant parameters. Each optimization takes over 10,000 steps.

We fit the law on the 400M model and 1B model trained with 20B tokens and an 8-1-1 LRS We then predict the loss curve for the 400M model and 1B model with cosine LRS and WSD LRS. The experiment result is present in Figure 2a.

FSL-optimal LRS via numerical variation. We propose to obtain a numerical optimal LRS by directly minimizing the final-step loss over the space of LRS using the fitted FSL, termed FSL-optimal LRS.

Step 1: Fitting FSL. Fit FSL on the loss curve of a 1B QwenMoE model trained on 20B tokens with batch size 288, maximum learning rate $\eta_0 = 0.001$, and the 8-1-1 scheduler over a total step of $K = 33907$, following the same procedure described earlier.

Step 2: Optimize LRS. To improve optimization stability, we reparameterize the learning rate schedule by defining

$$\delta_i = \eta_i - \eta_{i+1}, \text{ for } i = 0, 1, \ldots, K-1.$$

Then, the i-th step learning rate can be recovered by $\eta_i = \eta_0 - \sum_{k=0}^{i-1} \delta_k$, which defines a one-to-one correspondence between the learning rate schedule $\{\eta_i\}$ and $\{\delta_i\}$. The optimization problem is

$$\min_{\{\delta_i\}_{i=1}^{K}} \mathcal{L}_{\Theta}(\{\eta_i\}_{i=1}^{K}), \quad \text{subject to } \sum_{k=0}^{K-1} \delta_k \leqslant \eta_0,\ d\eta_i \geqslant 0, i = 0, 1, \ldots, K-1. \tag{17}$$

To solve the above constraint optimization, we use the projected gradient descent (PGD) [10]. The learning rate of PGD is searched ranging from 1×10^{-8} to 5×10^{-10}, and the optimization step number ranges from 50,000 to 100,000.

The resulting FSL-optimal LRS is presented in Figure 2b (left), where cosine, WSD, and 8-1-1 LRSs are also given for a comparison.

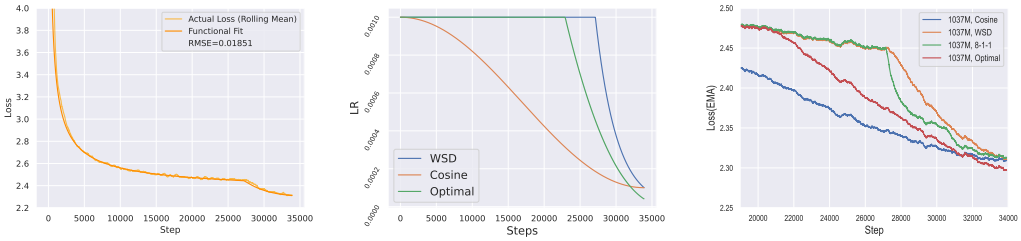
Step 3: Evaluate our LRS. We then evaluate the performance of the resulting FSL-optimal LRS, and the three LRSs in Figure 2b (left) are used as baseline. All comparisons are conducted on the same 1B QwenMoE model under identical training conditions: 33,907 total steps, batch size 288, and 20B training tokens. Full loss curves are shown in Figure 2b.

Additional Experiments We have further conducted ablation experiments with different model sizes and architectures, different total steps and different WSD schedules.

We validate our functional scaling law in models with various sizes, ranging from 100M to 1B, and diverse architectures including GPT-2 [51], LLaMA [59] and QwenMoE [67]. For each model, we first fit the FSL using the 8-1-1 LRS and subsequently employ it to predict the loss curve under a WSD LRS. Next we numerically solve the FSL-optimal LRS and empirically validate its efficacy by comparing the final pre-training loss against those obtained using other commonly adopted learning rate schedules. We present the results in Figure 4 for the 1B LLaMA dense model, Figure 5 for the 100M GPT-2 dense model. The consistent alignment between predicted and observed performance across architectures and sizes underscores the robustness and generalizability FSL.

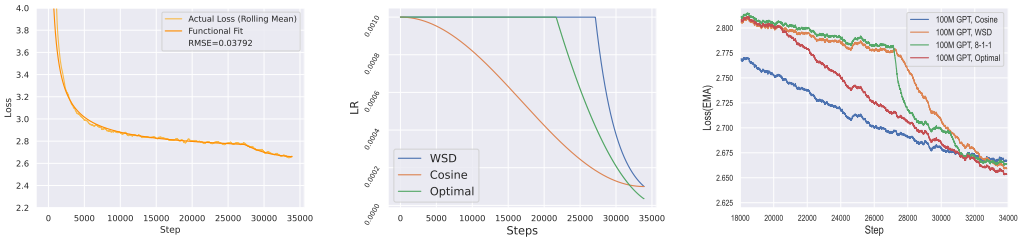
We further validate the applicability of our functional scaling law (FSL) across varying training durations. Using a 100M LLaMA dense model, we conduct experiments with total training steps set to 17k, 34k, 68k, and 134k. As demonstrated in Figures 6 and 7, our FSL accurately models the loss trajectories across all evaluated step counts, confirming its robustness to different total training steps.

Finally, we conduct a comprehensive empirical comparison between our FSL-optimal learning rate schedule and various WSD baselines, examining different decay ratios and minimum learning rate



(a) Prediction on WSD LRS (b) Optimal and Existing LRSs (c) Loss curve of LRSs

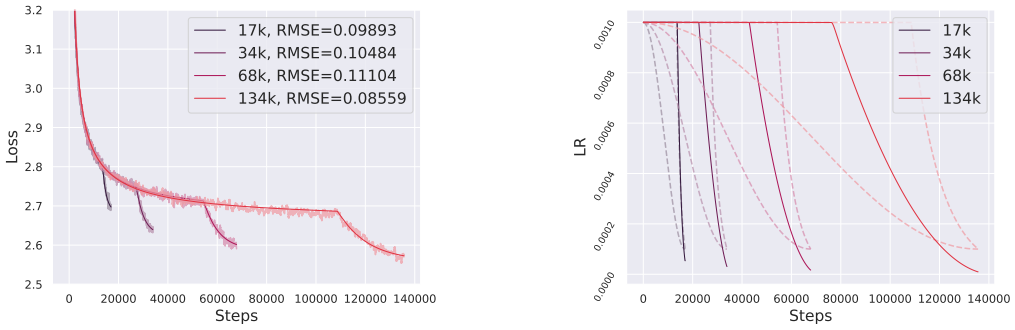
Figure 4: **Experiment on the 1B LLaMA (dense) model.** Figure (a): We fit our functional scaling law on the loss curve of 1B LLaMA (dense) model with 20B tokens training data and 8-1-1 LRS. Figures (b)(c): The comparison on the 1B model between the optimal LRS, cosine LRS, WSD LRS with exponential decay and 8-1-1 LRS.



(a) Prediction on WSD LRS (b) Optimal and Existing LRSs (c) Loss curve of LRSs

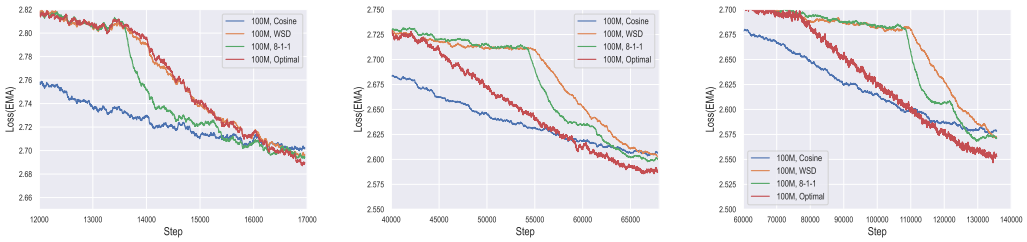
Figure 5: **Experiment on the 100M GPT2 (dense) model.** Figure (a): We fit our functional scaling law on the loss curve of 100M GPT2 (dense) model with 20B tokens training data and 8-1-1 LRS. Figures (b)(c): The comparison on the 100M model between the optimal LRS, cosine LRS, WSD LRS with exponential decay and 8-1-1 LRS.

configurations. As evidenced by Figures 8 and 9, our FSL-optimal LRS consistently outperforms all WSD variants, achieving superior final pre-training loss across all experimental conditions. This systematic evaluation demonstrates both the effectiveness of our theoretically-derived schedule and its practical advantages over conventional heuristic approaches.



(a) Prediction result (b) Optimal LRS

Figure 6: **Experiments with different total steps.** Figure (a): Fitted functional scaling laws on 100M LLaMA model with different total training steps 17k, 34k, 68k and 134k (corresponding to 10B, 20B, 40B and 80B tokens respectively). Figure (b): Optimal LRSs compared with cosine and WSD LRSs. The solid lines are optimal LRSs, and the dashed lines are cosine/WSD LRSs.



(a) 17k steps (b) 68k steps (c) 134k steps

Figure 7: **Experiments with different total steps.** We compare loss curves of existing LRSs and optimal LRS on the 100M LLaMA model with different total training steps 17k, 68k and 134k.

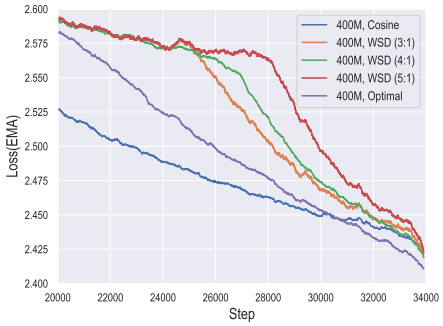


Figure 8: **WSD with different decay ratios:** We train a 400M LLaMA (dense) model with 20B tokens of training data and WSD LRSs with the ratios between stable time and decay time of 3:1, 4:1, and 5:1. All WSD LRSs exhibit a final loss similar to that of the Cosine LRS, and the optimal LRS derived from our functional scaling law outperforms all other LRSs by a loss gap of approximately 0.01.

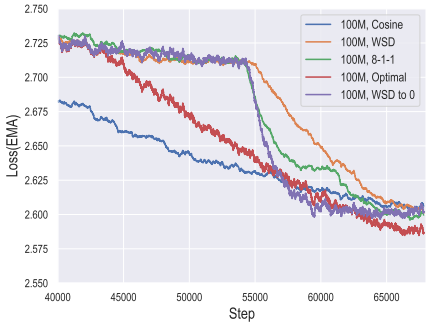


Figure 9: **Comparison between optimal LRS and WSD with a near-zero final learning rate:** We train a 100M LLaMA (dense) model with 40B tokens training data and various LRSs with the same $\eta_{\max} = 10^{-3}$, including WSD LRS with $\eta_{\min} = \frac{1}{10}\eta_{\max}$, WSD LRS with $\eta_{\min} = 10^{-7}$, cosine LRS with $\eta_{\min} = \frac{1}{10}\eta_{\max}$, 8-1-1 LRS with $\eta_{\min} = \frac{1}{10}\eta_{\max}$, and optimal LRS. The experimental results show that decaying to (near) zero does not result in significant loss reduction.

D Proof Sketch for the FSL (10)

In this section, we outline the main ideas behind the proof of the FSL (10), highlighting the key techniques. Complete proofs of the above theorems are deferred to Appendix E.

We will need the following characterization of the gradient noise structure.

Lemma D.1 (Noise structure)**.** *For any $\mathbf{v} \in \mathbb{R}^M$, it holds that*

$$(2C_1\mathcal{E}(\mathbf{v}) + \sigma^2)\,\nabla^2\mathcal{R}(\mathbf{v}) \;\preceq\; \mathbf{\Sigma}(\mathbf{v}) \;\preceq\; (2C_2\mathcal{E}(\mathbf{v}) + \sigma^2)\,\nabla^2\mathcal{R}(\mathbf{v}),$$

where $\nabla^2\mathcal{R}(\mathbf{v}) = \mathbf{W}\mathbf{H}\mathbf{W}^\top$, and the constants C_1 and C_2 are the same as in Assumption 2.1.

The proof is provided in Appendix E.1. Let $\boldsymbol{\xi}(\mathbf{v})$ denote the gradient noise at $\mathbf{v}$. Since $\mathcal{R}(\mathbf{v}) = \mathcal{E}(\mathbf{v}) + \frac{1}{2}\sigma^2$, it follows that for any direction $\boldsymbol{n} \in \mathbb{S}^{M-1}$,

$$\mathbb{E}[|\boldsymbol{\xi}(\mathbf{v})^\top \boldsymbol{n}|^2] = \boldsymbol{n}^\top \mathbf{\Sigma}(\mathbf{v})\boldsymbol{n} \;\eqsim\; \mathcal{R}(\mathbf{v})\,\boldsymbol{n}^\top \nabla^2\mathcal{R}(\mathbf{v})\boldsymbol{n},$$

where $\boldsymbol{n}^\top \nabla^2\mathcal{R}(\mathbf{v})\boldsymbol{n}$ represents the local curvature of the population risk along $\boldsymbol{n}$. Hence, the noise energy in each direction is proportional to the product of the population risk and the curvature along that direction. This **anisotropic structure** of the gradient noise—scaling with the risk and shaped by curvature—has also been reported in prior work [66, 61].

For clarity, in this section, we focus on the case of top-M features, for which the population risk takes the form

$$2\mathcal{E}(\mathbf{v}) = \sum_{j=1}^{M} \lambda_j (v_j - \theta_j^*)^2 + \sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2. \tag{18}$$

For the intrinsic-time SDE (7), each coordinate of $\boldsymbol{\nu}(t)$ evolves as

$$\mathrm{d}\nu_j(t) = -\lambda_j(\nu_j - \theta_j^*)\,\mathrm{d}t + \sqrt{\gamma(t)q_j(t)}\,\mathrm{d}B_j(t),$$

where $q_j(t) := \mathbf{e}_j^\top \mathbf{\Sigma}(\boldsymbol{\nu}_t)\mathbf{e}_j$ is the variance of gradient noise along $\mathbf{e}_j$, and $\mathbf{e}_j$ is the j-th canonical basis vector for $j \in [M]$. By applying Itô's formula to $(\nu_j - \theta_j^*)^2$ and noting $\boldsymbol{\nu}(0) = \mathbf{0}$, we obtain

$$\mathbb{E}[(\nu_j(t) - \theta_j^*)^2] = (0 - \theta_j^*)^2 e^{-2\lambda_j t} + \lambda_j \int_0^t e^{-2\lambda_j(t-\tau)}\gamma(\tau)q_j(\tau)\,\mathrm{d}\tau.$$

Let $\mathcal{E}_t = \mathcal{E}(\boldsymbol{\nu}_t)$ and plugging the above equation into (18) gives

$$2\mathbb{E}[\mathcal{E}_t] = \sum_{j=1}^{M} \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t} + \sum_{j=1}^{M} \lambda_j^2 \int_0^t e^{-2\lambda_j(t-\tau)}\gamma(\tau)\mathbb{E}[q_j(\tau)]\,\mathrm{d}\tau + \sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2. \tag{19}$$

By Lemma D.1 and noting $\nabla^2\mathcal{R}(\mathbf{v}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, we have

$$q_j = \mathbf{e}_j^\top \mathbf{\Sigma}(\boldsymbol{\nu}_t)\mathbf{e}_j \eqsim \lambda_j \mathcal{R}(\boldsymbol{\nu}_t) = \lambda_j(\mathcal{E}(\boldsymbol{\nu}_t) + \sigma^2/2).$$

Let $\delta_M = \sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2$, $e_M(t) = \sum_{j=1}^{M} \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t}$, and $\mathcal{K}_M(t) = \sum_{j=1}^{M} \lambda_j^2 e^{-2\lambda_j t}$. Plugging them back into (19) gives the following **Volterra equation**:

$$\mathbb{E}[\mathcal{E}_t] \eqsim \delta_M + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)\gamma(\tau)(\mathbb{E}[\mathcal{E}_t] + \sigma^2)\,\mathrm{d}\tau. \tag{20}$$

The above equation characterizes the expected loss dynamics of SGD under a general spectrum and has been derived in prior works such as [46, 47, 28, 48, 49]. Our key observation is that, under the power-law assumptions on $\{\theta_j^*\}_j$ and $\{\lambda_j\}_j$ (Assumptions 2.3 and 2.4), the solution to (20) admits a *sharp asymptotic characterization*, providing explicit upper and lower bounds that precisely capture its scaling behavior.

Let $f(t) := \mathbb{E}[\mathcal{E}_t]$, $g(t) := \delta_M + e_M(t) + \sigma^2 \int_0^t \mathcal{K}_M(t-\tau)\gamma(\tau)\,\mathrm{d}\tau$, and define the linear operator

$$\mathcal{T}f(t) = \int_0^t \mathcal{K}_M(t-\tau)\gamma(\tau)f(\tau)\,\mathrm{d}\tau.$$

Then, the Volterra equation (20) can be expressed in the compact form $f = g + \mathcal{T}f$. Formally, its solution can be expanded as an infinite series:

$$f = (\mathcal{I} - \mathcal{T})^{-1} g = g + \mathcal{T}g + \mathcal{T}^2 g + \mathcal{T}^3 g + \cdots . \tag{21}$$

The key observation is that, under Assumptions 2.3 and 2.4, **the higher-order terms $\mathcal{T}^k g$ for $k \geqslant 2$ can be well controlled by the first-order term $\mathcal{T}g$.**

Lemma D.2 (Half-scale comparability)**.** *For any $t \lesssim M^\beta$, it holds $\mathcal{K}_M(t/2) \lesssim \mathcal{K}_M(t)$.*

Proof. The result follows from the fact that $\mathcal{K}_M$ exhibits a power-law decay for $t \lesssim M^\beta$. Indeed,

$$\mathcal{K}_M(t) = \sum_{j=1}^{M} \lambda_j^2 e^{-2\lambda_j t} \;\eqsim\; \int_{M^{-\beta}}^{1} z^{1-\frac{1}{\beta}} e^{-2zt}\,\mathrm{d}z \;\eqsim\; t^{-(2-1/\beta)}, \qquad 1 \lesssim t \lesssim M^\beta.$$

Consequently, $\mathcal{K}_M(t/2) \eqsim (t/2)^{-(2-1/\beta)} \eqsim t^{-(2-1/\beta)} \eqsim \mathcal{K}_M(t)$, which establishes the claim. □

Corollary D.3 (Subconvolution property)**.** *For any $t \lesssim M^\beta$, it holds $\mathcal{K}_M * \mathcal{K}_M(t) \lesssim \mathcal{K}_M(t)$.*

Proof. Noting $\mathcal{K}_M$ is non-increasing and integrable. and applying Lemma D.2, we obtain

$$\begin{aligned}(\mathcal{K}_M * \mathcal{K}_M)(t) &= \int_0^{t/2} \mathcal{K}_M(t-\tau)\mathcal{K}_M(\tau)\,\mathrm{d}\tau + \int_{t/2}^{t} \mathcal{K}_M(t-\tau)\mathcal{K}_M(\tau)\,\mathrm{d}\tau \\ &\leqslant 2\,\mathcal{K}_M(t/2)\int_0^{t/2} \mathcal{K}_M(\tau)\,\mathrm{d}\tau \;\lesssim\; \mathcal{K}_M(t/2) \;\lesssim\; \mathcal{K}_M(t),\end{aligned}$$

which proves the claim. □

Let $\|\gamma\|_\infty = \max_{t \geqslant 0} \gamma(t)$. By Corollary D.3, it holds for any $t \lesssim M^\beta$ that

$$\begin{aligned}\mathcal{T}^2 g(t) &\leqslant \|\gamma\|_\infty \int_0^t \mathcal{K}_M * \mathcal{K}_M(t-\tau)\gamma(\tau)g(\tau)\,\mathrm{d}\tau \\ &\lesssim \|\gamma\|_\infty \int_0^t \mathcal{K}_M(t-\tau)\gamma(\tau)g(\tau)\,\mathrm{d}\tau = \|\gamma\|_\infty \mathcal{T}g(t).\end{aligned}$$

When $\|\gamma\|_\infty$ is sufficiently small, there exists a constant $0 < c < 1$ such that $\mathcal{T}^k g(t) \leqslant c^{k-1}\mathcal{T}g(t)$ holds for any $0 \leqslant t \lesssim M^\beta$. Hence,

$$f(t) \leqslant g(t) + \mathcal{T}g(t) + \sum_{k=2}^{\infty} c^{k-1}\mathcal{T}g(t) \lesssim g(t) + \mathcal{T}g(t) + \tfrac{c}{1-c}\mathcal{T}g(t).$$

Combining $f(t) \geqslant g(t) + \mathcal{T}g(t)$ with the above upper bound, we conclude $f(t) \eqsim g(t) + \mathcal{T}g(t)$, which completes the proof of FSL (10).

E Proofs for Section 4

E.1 Volterra Integral Equation Governing the Loss Dynamics

In this section, we derive a Volterra-type integral equation that exactly characterizes the evolution of expected loss under the intrinsic-time SDE. This equation serves as the starting point for all subsequent theoretical analysis. Recall the intrinsic-time SDE:

$$\mathrm{d}\boldsymbol{\nu}_t = -\nabla\mathcal{R}(\boldsymbol{\nu}_t)\,\mathrm{d}t + \sqrt{\gamma_t \boldsymbol{\Sigma}(\boldsymbol{\nu}_t)}\,\mathrm{d}\mathbf{B}_t, \tag{22}$$

where $\gamma_t = \gamma_{\varphi,b}(t)$. Here, we drop the dependence on φ and b for simplicity.

By the definition of $\mathcal{R}(\mathbf{v})$, we have $\nabla\mathcal{R}(\mathbf{v}) = \mathbf{W}\mathbf{H}(\mathbf{W}^\top\mathbf{v} - \theta^*)$. Let $\mathbf{u}_t = \mathbf{W}^\top\boldsymbol{\nu}_t - \boldsymbol{\theta}^*$. Then, we have

$$\mathcal{E}_t = \mathcal{E}(\mathbf{u}_t) = \frac{1}{2}\|\mathbf{u}_t\|_{\mathbf{H}}^2$$

To obtain the estimate of $\mathcal{E}_t$, we consider the intrinsic-time SDE for $\mathbf{u}_t$ given by:

Lemma E.1. *We have*

$$\mathrm{d}\mathbf{u}_t = -\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{u}_t \,\mathrm{d}t + \sqrt{\gamma_t \mathbf{W}^\top \mathbf{\Sigma}_t \mathbf{W}}\,\mathrm{d}\mathbf{B}_t, \tag{23}$$

where $\mathbf{\Sigma}_t := \mathbf{\Sigma}(\boldsymbol{\nu}_t)$.

Proof. By Eq. (22),

$$\mathrm{d}\mathbf{u}_t = \mathrm{d}(\mathbf{W}^\top \boldsymbol{\nu}_t - \mathbf{v}^*) = \mathbf{W}^\top \,\mathrm{d}\boldsymbol{\nu}_t = -\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{u}_t \,\mathrm{d}t + \mathbf{W}^\top \sqrt{\gamma_t \mathbf{\Sigma}_t}\,\mathrm{d}\tilde{\mathbf{B}}_t.$$

Here $\tilde{\mathbf{B}}_t$ is an N dimensional standard Brownian motion, we are going to replace it with an M dimensional standard Brownian motion $\mathbf{B}_t$.

It is easy to see that the diffusion term $\mathbf{W}^\top \sqrt{\gamma_t \mathbf{\Sigma}_t}\,\mathrm{d}\tilde{\mathbf{B}}_t$ has the same distribution as $\sqrt{\gamma_t \mathbf{W}^\top \mathbf{\Sigma}_t \mathbf{W}}\,\mathrm{d}\mathbf{B}_t$, hence the SDE can be written in $\mathbf{B}_t$ as

$$\mathrm{d}\mathbf{u}_t = -\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{u}_t \,\mathrm{d}t + \sqrt{\gamma_t \mathbf{W}^\top \mathbf{\Sigma}_t \mathbf{W}}\,\mathrm{d}\mathbf{B}_t.$$

□

A key insight for tractability is that the gradient noise exhibits the following anisotropic structure:

Our analytic analysis also relies on the noise structure characterized by the following lemma.

Lemma E.2 (Noise Structure)**.** *For any* $\mathbf{v} \in \mathbb{R}^M$*, it holds that*

$$(2C_1\mathcal{E}(\mathbf{v}) + \sigma^2)\,\mathbf{W}\mathbf{H}\mathbf{W}^\top \preceq \mathbf{\Sigma}(\mathbf{v}) \preceq (2C_2\mathcal{E}(\mathbf{v}) + \sigma^2)\,\mathbf{W}\mathbf{H}\mathbf{W}^\top.$$

Noting $\nabla^2\mathcal{R}(\mathbf{v}) = \mathbf{W}\mathbf{H}\mathbf{W}^\top$ and $\mathcal{R}(\mathbf{v}) = \mathcal{E}(\mathbf{v}) + \frac{1}{2}\sigma^2$, this lemma means $\mathbf{\Sigma}(\mathbf{v}) \eqsim \mathcal{R}(\mathbf{v})\nabla^2\mathcal{R}(\mathbf{v})$. That is, the gradient noise scales proportionally with the population risk and aligns with the local curvature. Notably, the noise has two distinct sources: (i) the fit-dependent term $\mathcal{E}(\mathbf{v})$, which arises purely from minibatching and persists even in the absence of label noise; (ii) the σ^2 term, which captures the contribution from label noise. This anisotropic structure of SGD noise – scaling with risk and shaped by curvature – has also been observed in prior work [66, 65].

Proof. Noting $\ell(\mathbf{z};\mathbf{v}) = \frac{1}{2}(\mathbf{v}^\top \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) - y)^2$, we have

$$\begin{aligned}
\nabla\ell(\mathbf{z};\mathbf{v}) &= \mathbf{W}\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top \left(\mathbf{W}^\top\mathbf{v} - \boldsymbol{\theta}^*\right) - \mathbf{W}\boldsymbol{\phi}(\mathbf{x})\epsilon \\
\nabla\mathcal{R}(\mathbf{v}) &= \mathbb{E}[\nabla\ell(\mathbf{z};\mathbf{v})] = \mathbf{W}\mathbf{H}\left(\mathbf{W}^\top\mathbf{v} - \boldsymbol{\theta}^*\right).
\end{aligned}$$

Hence, the covariance matrix of the noise $\boldsymbol{\xi} := \nabla\ell(\mathbf{z};\mathbf{v}) - \nabla\mathcal{R}(\mathbf{v})$ is given by

$$\begin{aligned}
\mathbf{\Sigma}(\mathbf{v}) &= \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top|\mathbf{v}] \\
&= \mathbf{W}\left(\mathbb{E}\left[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\mathbf{u}\mathbf{u}^\top\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\right] - \mathbf{H}\mathbf{u}\mathbf{u}^\top\mathbf{H}\right)\mathbf{W}^\top + \sigma^2\mathbf{W}\mathbf{H}\mathbf{W}^\top.
\end{aligned}$$

Noting

$$\mathbb{E}\left[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\mathbf{u}\mathbf{u}^\top\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\right] - \mathbf{H}\mathbf{u}\mathbf{u}^\top\mathbf{H} = \mathbb{E}\left[\boldsymbol{\phi}(\mathbf{x})^\top\mathbf{u}\mathbf{u}^\top\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top\right] - \mathbf{H}\mathbf{u}\mathbf{u}^\top\mathbf{H},$$

then applying Assumption 2.1, we have

$$\mathbf{\Sigma}(\mathbf{v}) \preceq C_2\mathbf{W}\mathrm{tr}(\mathbf{H}\mathbf{u}\mathbf{u}^\top)\mathbf{H}\mathbf{W}^\top + \sigma^2\mathbf{W}\mathbf{H}\mathbf{W}^\top = \left(2C_2\mathcal{E}(\mathbf{u}) + \sigma^2\right)\mathbf{W}\mathbf{H}\mathbf{W}^\top,$$

where the last step follows from $\mathrm{tr}(\mathbf{H}\mathbf{u}\mathbf{u}^\top) = \|\mathbf{u}\|_{\mathbf{H}}^2 = 2\mathcal{E}(\mathbf{v})$. The lower bound follows the same proof. □

The excess-risk dynamics is then given by the following Volterra integral equation:

Proposition E.3. *For the intrinsic-time SDE, we have*

$$2\mathbb{E}[\mathcal{E}_t] = \mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 + \int_0^t \mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S})\cdot\gamma_\tau(c_\tau\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau, \tag{24}$$

where $\mathbf{A}_t := e^{-\mathbf{W}^\top\mathbf{W}\mathbf{H}t}$, $\mathbf{S} := \sqrt{\mathbf{W}^\top\mathbf{W}\mathbf{H}\mathbf{W}^\top\mathbf{W}}$, *and* $2C_1 \leqslant c_\tau \leqslant 2C_2$ *are some constants.*

Proof. By Itô's formula,

$$\begin{aligned}\mathrm{d}(e^{\mathbf{W}^\top\mathbf{W}\mathbf{H}t}\mathbf{u}_t) &= e^{\mathbf{W}^\top\mathbf{W}\mathbf{H}t}(\mathbf{W}^\top\mathbf{W}\mathbf{H}\mathbf{u}_t\,\mathrm{d}t + \mathrm{d}\mathbf{u}_t)\\ &= e^{\mathbf{W}^\top\mathbf{W}\mathbf{H}t}\sqrt{\gamma_t\mathbf{W}^\top\mathbf{\Sigma}_t\mathbf{W}}\,\mathrm{d}\mathbf{B}_t.\end{aligned}$$

Integrating both sides, we get

$$e^{\mathbf{W}^\top\mathbf{W}\mathbf{H}t}\mathbf{u}_t - \mathbf{u}_0 = \int_0^t e^{\mathbf{W}^\top\mathbf{W}\mathbf{H}\tau}\sqrt{\gamma_\tau\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\,\mathrm{d}\mathbf{B}_\tau.$$

Now write $\mathbf{A}_t = e^{-\mathbf{W}^\top\mathbf{W}\mathbf{H}t}$, we have

$$\mathbf{u}_t = \mathbf{A}_t\mathbf{u}_0 + \int_0^t \mathbf{A}_{t-\tau}\sqrt{\gamma_\tau\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\,\mathrm{d}\mathbf{B}_\tau.$$

Note that the integral with respect to $\mathbf{B}_t$ always has zero expectation, therefore we have

$$\begin{aligned}2\mathbb{E}\mathcal{E}_t &= \mathbb{E}(\mathbf{u}_t^\top\mathbf{H}\mathbf{u}_t)\\ &= \mathbf{u}_0^\top\mathbf{A}_t^\top\mathbf{H}\mathbf{A}_t\mathbf{u}_0 + \mathbb{E}\int_0^t \gamma_\tau\mathrm{tr}\left(\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\right)\mathrm{d}\tau.\end{aligned}$$

By Lemma G.2 and Lemma E.2, we have

$$\mathrm{tr}\left(\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\right) \leqslant (2\mathcal{E}_\tau + \sigma^2)\mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}),$$

$$\mathrm{tr}\left(\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\right) \geqslant (\mathcal{E}_\tau + \sigma^2)\mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}).$$

Hence there exists some constant $c_\tau \in [2C_1, 2C_2]$ such that

$$\mathbb{E}\mathrm{tr}\left(\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\sqrt{\mathbf{W}^\top\mathbf{\Sigma}_\tau\mathbf{W}}\right) = (c_\tau\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top\mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}),$$

from which the lemma follows. □

E.2 The Case of Top-M Features

First, we prove the general label-noise case of FSL (Theorem 4.3). Applying this result, we then derive the constant label-noise case (Theorem 4.4), the noiseless case (Theorem 4.5) and the hard regime case (Theorem 4.2)

E.2.1 Proof of Theorem 4.3

In the top-M feature case, the matrix $\mathbf{W}$ satisfies $\mathbf{w}_j = \mathbf{e}_j$ for each $j \in [M]$, therefore we can simplify the equation for $\mathbb{E}[\mathcal{E}_t]$ as follows.

Theorem E.4 (Volterra equation of the top-M case)**.** *In the top-M case, we have*

$$\mathbb{E}[\mathcal{E}_t] \eqsim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)\gamma_\tau(\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau, \tag{25}$$

where the function e_M and $\mathcal{K}_M$ are defined as

$$e_M(t) := \sum_{j=1}^M \lambda_j(\theta_j^*)^2 e^{-2\lambda_j t}, \quad \mathcal{K}_M(t) := \sum_{j=1}^M \lambda_j^2 e^{-2\lambda_j t}. \tag{26}$$

Proof. By (24) in Proposition E.3, note that $\mathbf{W}^\top\mathbf{W}\mathbf{H} = \mathbf{H}_{0:M} \in \mathbb{R}^{N\times N}$ is the top-M part of the matrix $\mathbf{H}$, i.e. $\mathbf{H}_{0:M} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_M, 0, \ldots, 0\}$, we get

$$\mathbb{E}[\mathcal{E}_t] \eqsim \mathbf{u}_0^\top\mathbf{H}e^{-2\mathbf{H}_{0:M}t}\mathbf{u}_0 + \int_0^t \mathrm{tr}(\mathbf{H}_{0:M}^2 e^{-2\mathbf{H}_{0:M}})\gamma_\tau(\mathbb{E}[\mathcal{E}_t] + \sigma^2)\,\mathrm{d}\tau,$$

which can be further written in terms of the eigenvalues $\{\lambda_j\}$ as

$$\mathbb{E}[\mathcal{E}_t] \eqsim \sum_{j=1}^{M} \lambda_j (u_0^{(j)})^2 e^{-2\lambda_j t} + \sum_{j=M+1}^{\infty} \lambda_j (u_0^{(j)})^2 + \int_0^t \sum_{j=1}^{M} \lambda_j^2 e^{-2\lambda_j t} \gamma_\tau (\mathbb{E}[\mathcal{E}_t] + \sigma^2) \, \mathrm{d}\tau.$$

Note that $u_0^{(j)}$, the j-th component of $\mathbf{u}_0$, is equal to θ_j^* because of the zero initialization of $\boldsymbol{\nu}_0$.

Therefore by the definition of e_M and $\mathcal{K}_M$ we arrive at the Volterra-type integral equation of $\mathbb{E}[\mathcal{E}_t]$. □

Lemma E.5. *For the forgetting kernel $\mathcal{K}_M$ and $t \leqslant M^\beta$, there exists a constant C independent of t, such that*

$$\mathcal{K}_M * \mathcal{K}_M(t) \leqslant C\mathcal{K}_M(t), \quad \forall t \leqslant M^\beta.$$

where $*$ denotes convolution:

$$\mathcal{K}_M * \mathcal{K}_M(t) := \int_0^t \mathcal{K}_M(\tau)\mathcal{K}_M(t-\tau) \, \mathrm{d}\tau.$$

Proof. Observe that $\mathcal{K}_M(t)$ is a monotonically decreasing function, by the symmetry of the convolution, we have

$$\begin{aligned} \mathcal{K}_M * \mathcal{K}_M(t) &= \int_0^t \mathcal{K}_M(\tau)\mathcal{K}_M(t-\tau) \, \mathrm{d}\tau \\ &= 2\int_0^{t/2} \mathcal{K}_M(\tau)\mathcal{K}_M(t-\tau) \, \mathrm{d}\tau. \end{aligned}$$

Since $\mathcal{K}_M$ is decreasing, for $0 \leqslant \tau \leqslant t/2$ we have $\mathcal{K}_M(t-\tau) \leqslant \mathcal{K}_M(t/2)$, thus

$$\mathcal{K}_M * \mathcal{K}_M(t) \leqslant 2\mathcal{K}_M\left(\tfrac{t}{2}\right) \int_0^\infty \mathcal{K}_M(\tau) \, \mathrm{d}\tau \leqslant C\mathcal{K}_M\left(\tfrac{t}{2}\right),$$

where

$$C = 2\int_0^\infty \mathcal{K}_M(\tau) \, \mathrm{d}\tau \leqslant 2\int_0^\infty \sum_{j=1}^{\infty} \lambda_j^2 e^{-2\lambda_j \tau} \, \mathrm{d}\tau = \sum_{j=1}^{\infty} \lambda_j = \mathrm{tr}(\mathbf{H}) < \infty.$$

It remains to show that when $t \leqslant M^\beta$,

$$\mathcal{K}_M\left(\frac{t}{2}\right) \leqslant C'\mathcal{K}_M(t)$$

for some constant $C' > 0$. Recall that

$$\mathcal{K}_M(t) = \sum_{j=1}^{\infty} \lambda_j^2 e^{-2\lambda_j t} = \sum_{j=1}^{M} j^{-2\beta} e^{-2j^{-\beta}t} \eqsim \int_1^M x^{-2\beta} e^{-2x^{-\beta}t} \, \mathrm{d}x.$$

By the change of variable $y = x^{-\beta}t$, one obtains

$$\mathcal{K}_M(t) \eqsim \frac{1}{\beta} \, t^{-2+\frac{1}{\beta}} \int_{tM^{-\beta}}^{t} y^{1-\frac{1}{\beta}} e^{-2y} \, \mathrm{d}y.$$

When $t \leqslant M^\beta$, the lower limit $tM^{-\beta} \leqslant 1$, and the integrand is smooth and bounded on $[tM^{-\beta}, t]$. Hence the integral remains comparable up to a constant factor, i.e., there exists a constant $C_1 > 0$ such that

$$\mathcal{K}_M\left(\tfrac{t}{2}\right) \leqslant C_1\mathcal{K}_M(t), \quad \forall \, t \leqslant M^\beta.$$

Combining the above estimates, we obtain

$$(\mathcal{K}_M * \mathcal{K}_M)(t) \leqslant C \, \mathcal{K}_M\left(\tfrac{t}{2}\right) \leqslant CC_1 \, \mathcal{K}_M(t) =: C'\mathcal{K}_M(t).$$

The proof is complete. □

We now prove Theorem 4.3.

Proof. The lower bound is trivial by $\mathbb{E}[\mathcal{E}_t] \geqslant e_M(t)$ and the Volterra equation (25). For the upper bounds, first we prove the weaker bound with $\mathcal{K}_\infty$.

Proof of weaker upper bound. By Equation (25), we have

$$\begin{aligned}\mathbb{E}[\mathcal{E}_t] &\eqsim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)\gamma_\tau(\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau \\ &\leqslant M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_\infty(t-\tau)\gamma_\tau(\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau.\end{aligned}$$

Define $f(t) := \int_0^t \mathcal{K}_\infty(t-\tau)\gamma_\tau(\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau$, substitute the above inequality into the right-hand side, we get

$$\begin{aligned}f(t) &\lesssim \int_0^t \mathcal{K}_\infty(t-\tau)(M^{-s\beta} + e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau \\ &\quad + \int_0^t \int_0^\tau \mathcal{K}_\infty(t-\tau)\mathcal{K}_\infty(\tau-r)\gamma_\tau\gamma_r(\mathbb{E}[\mathcal{E}_r] + \sigma^2)\,\mathrm{d}r\,\mathrm{d}\tau \\ &\lesssim \gamma_{\max}M^{-s\beta} + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau \\ &\quad + \gamma_{\max}\int_0^t \int_r^t \mathcal{K}_\infty(t-\tau)\mathcal{K}_\infty(\tau-r)\,\mathrm{d}\tau\gamma_r(\mathbb{E}[\mathcal{E}_r] + \sigma^2)\,\mathrm{d}r \\ &= \gamma_{\max}M^{-s\beta} + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau \\ &\quad + \gamma_{\max}\int_0^t (\mathcal{K}_\infty * \mathcal{K}_\infty)(t-r)\gamma_r(\mathbb{E}[\mathcal{E}_r] + \sigma^2)\,\mathrm{d}r \\ &\overset{\text{Lemma E.5}}{\lesssim} \gamma_{\max}M^{-s\beta} + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau + \gamma_{\max}f(t)\end{aligned}$$

Therefore when $\gamma_{\max}$ is sufficiently small ($\gamma_{\max} \leqslant \frac{c}{\mathrm{tr}(\mathbf{H})}$ for some absolute constant c), the constant factor of $f(t)$ on the right-hand side will be less than $\frac{1}{2}$, hence we can substract it from both sides and get

$$\int_0^t \mathcal{K}_\infty(t-\tau)\gamma_\tau\mathbb{E}[\mathcal{E}_\tau]\,\mathrm{d}\tau \lesssim \gamma_{\max}M^{-s\beta} + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau.$$

Therefore substitude this back to the Volterra equation yields

$$\mathbb{E}[\mathcal{E}_t] \lesssim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau\,\mathrm{d}\tau.$$

■

Note that in the above proof, the only properties of $\mathcal{K}_\infty$ we used are Lemma E.5 and the convergence of integral $\int_0^\infty K_\infty(t)\,\mathrm{d}t$.

Now when $t \leqslant M^\beta$, as the kernel $\mathcal{K}_M$ also satisfies these two properties, the proof of the stronger bound is identical with the above, except we replace $\mathcal{K}_\infty$ with $\mathcal{K}_M$. □

E.2.2 Proof of Theorem 4.4

Proof. It is clear that $e_M(t)$, γ_τ are all bounded from above by constants, thus by the upper bound in Theorem 4.3, noting that $\sigma^2 \gtrsim 1$,

$$\mathbb{E}[\mathcal{E}_t] \lesssim 1 + (\sigma^2 + 1)\int_0^t \mathcal{K}_\infty(t-\tau)\,\mathrm{d}\tau \lesssim \sigma^2,$$

since the integral $\int_0^\infty \mathcal{K}_\infty(t)\,\mathrm{d}t$ is convergent.

Therefore by the Volterra equation (25), note that $\mathbb{E}[\mathcal{E}_\tau] + \sigma^2 \eqsim \sigma^2 \eqsim e_M(\tau) + \sigma^2$,

$$\mathbb{E}[\mathcal{E}_t] \eqsim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)(e_M(\tau)+\sigma^2)\gamma_\tau\,\mathrm{d}\tau.$$

□

E.2.3 Proof of Theorem 4.5

Proof. When $\sigma^2 = 0$, by FSL in general case (Theorem 4.3), we have

$$\mathbb{E}[\mathcal{E}_t] \lesssim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_\infty(t-\tau)e_M(\tau)\gamma_\tau\,\mathrm{d}\tau.$$

Claim. We will bound the gap introduced by $\mathcal{K}_\infty$ and $\mathcal{K}_M$:

$$\int_0^t (\mathcal{K}_\infty(t-\tau) - \mathcal{K}_M(t-\tau))e_M(\tau)\gamma_\tau\,\mathrm{d}\tau \lesssim M^{\max\{-s\beta,-2\beta+1\}}. \tag{27}$$

Proof of Claim. First note that

$$\mathcal{K}_\infty(t) - \mathcal{K}_M(t) = \sum_{j=M+1}^\infty \lambda_j^2 e^{-2\lambda_j t} \lesssim \sum_{j=M+1}^\infty j^{-2\beta} \eqsim M^{-2\beta+1}.$$

Therefore we can bound the integral as

$$\begin{aligned}
&\int_0^t (\mathcal{K}_\infty(t-\tau) - \mathcal{K}_M(t-\tau))e_M(\tau)\gamma_\tau\,\mathrm{d}\tau \\
&\lesssim M^{-2\beta+1}\int_0^t e_M(\tau)\gamma_\tau\,\mathrm{d}\tau \\
&\leqslant \gamma_{\max} M^{-2\beta+1}\int_0^\infty e_M(\tau)\,\mathrm{d}\tau \\
&= \gamma_{\max} M^{-2\beta+1}\int_0^\infty \sum_{j=1}^M \lambda_j(\theta_j^*)^2 e^{-2\lambda_j\tau}\,\mathrm{d}\tau \\
&\eqsim \gamma_{\max} M^{-2\beta+1}\sum_{j=1}^M j^{-s\beta-1+\beta} \\
&\lesssim \gamma_{\max} M^{\max\{-s\beta-\beta+1,-2\beta+1\}} \lesssim \gamma_{\max} M^{\max\{-s\beta,-2\beta+1\}}
\end{aligned}$$

■

By $s \leqslant 2 - \frac{1}{\beta}$ and $\gamma_{\max}$ is sufficiently small, we can combine the upper bound with (27), and directly conclude that

$$\mathbb{E}[\mathcal{E}_t] \lesssim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)e_M(\tau)\gamma_\tau\,\mathrm{d}\tau.$$

Now with the lower bound in FSL Theorem 4.3, the result follows. □

E.2.4 Proof of Theorem 4.2

Proof. In the hard regime $s \leqslant 1 - \frac{1}{\beta}$, by FSL in general case (Theorem 4.3), we have

$$\mathbb{E}[\mathcal{E}_t] \lesssim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_\infty(t-\tau)(e_M(\tau)+\sigma^2)\gamma_\tau\,\mathrm{d}\tau.$$

Claim. We will bound the gap introduced by $\mathcal{K}_\infty$ and $\mathcal{K}_M$:

$$\int_0^t (\mathcal{K}_\infty(t-\tau) - \mathcal{K}_M(t-\tau))\gamma_\tau \,\mathrm{d}\tau \lesssim M^{-\beta+1}. \tag{28}$$

Proof of Claim. First note that

$$\mathcal{K}_\infty(t) - \mathcal{K}_M(t) = \sum_{j=M+1}^{\infty} \lambda_j^2 e^{-2\lambda_j t}$$

Therefore we can bound the integral as

$$\begin{aligned}
&\int_0^t (\mathcal{K}_\infty(t-\tau) - \mathcal{K}_M(t-\tau))\gamma_\tau \,\mathrm{d}\tau \\
&= \gamma_{\max} \int_0^\infty \sum_{j=1}^{M} \lambda_j^2 e^{-2\lambda_j \tau} \,\mathrm{d}\tau \\
&\eqsim \gamma_{\max} \sum_{j=1}^{M} j^{-\beta} \lesssim \gamma_{\max} M^{-\beta+1}.
\end{aligned}$$

■

By $s \leqslant 1 - \frac{1}{\beta}$ and $\gamma_{\max}$ is sufficiently small, we can combine the upper bound with (28), (27), and directly conclude that

$$\mathbb{E}[\mathcal{E}_t] \lesssim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)(e_M(\tau) + \sigma^2)\gamma_\tau \,\mathrm{d}\tau.$$

Now with the lower bound in FSL Theorem 4.3, the result follows. □

E.3 The Case of Random-M Features

First, for the random-M feature setting, we can establish the following Volterra equation.

Proposition E.6. *Suppose $0 < s \leqslant 1$. Then, with probability at least $1 - e^{-\Omega(M)}$, the Volterra equation derived in Theorem E.4 continues to hold for the random-feature case; that is,*

$$\mathbb{E}[\mathcal{E}_t] \;\eqsim\; M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)\,\gamma_\tau\big(\mathbb{E}[\mathcal{E}_\tau] + \sigma^2\big)\,\mathrm{d}\tau. \tag{29}$$

Theorem 4.6 then follows by applying exactly the same argument as in the top-M case. It therefore remains to prove Proposition E.6. The key idea is to show that the spectrum of the random matrix $\mathbf{W}\mathbf{H}\mathbf{W}^\top \in \mathbb{R}^{M\times M}$ closely matches that of the top-M truncation, namely,

$$\mu_j(\mathbf{W}\mathbf{H}\mathbf{W}^\top) \;\eqsim\; \lambda_j, \qquad 1 \leqslant j \leqslant M.$$

E.3.1 Concentration Inequalities

Recall that we derived the following recursive equation in Eq. (24):

$$2\mathbb{E}\mathcal{E}_t = \mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 + \int_0^t \mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}) \cdot \gamma_\tau(c_\tau \mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau,$$

where $\mathbf{A}_t = e^{-\mathbf{W}^\top\mathbf{W}\mathbf{H}t}$ and $\mathbf{S} = (\mathbf{W}^\top\mathbf{W}\mathbf{H}\mathbf{W}^\top\mathbf{W})^{\frac{1}{2}}$.

We first introduce the following notation: for integers $0 \leqslant a < b \leqslant N$ (we allow $b = \infty$, in this case we regard it as the same as $b = N$),

$$\mathbf{H}_{a:b} = \mathrm{diag}\{\lambda_{a+1}, \ldots, \lambda_b\} \in \mathbb{R}^{(b-a)\times(b-a)}, \quad \mathbf{u}_{a:b} = ((\mathbf{u})_{a+1}, \ldots, (\mathbf{u})_b) \in \mathbb{R}^{b-a},$$

while

$$\mathbf{W}_{a:b} = [\mathbf{W}_{a+1}, \ldots, \mathbf{W}_b] \in \mathbb{R}^{M\times(b-a)}$$

is the $(a+1)$-th to b-th columns of $\mathbf{W}$.

To understand this equation with random projection matrix $\mathbf{W}$, we leverage the following concentration results developed in [35].

Lemma E.7 (Lemma G.4 in [35]). *There exist β-dependent constants $0 < c_1 < c_2$ such that it holds with probability at least $1 - e^{-\Omega(M)}$ for all $j \in [M]$ that*

$$c_1 j^{-\beta} \leqslant \mu_j(\mathbf{W}\mathbf{H}\mathbf{W}^\top) \leqslant c_2 j^{-\beta}$$

Lemma E.8 (Lemma G.5 in [35]). *There exists some β-dependent constant c such that for all $k \geqslant 1$, the ratio between the $\frac{M}{2}$-th and M-th eigenvalue*

$$\frac{\mu_{\frac{M}{2}}(\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top)}{\mu_M(\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top)} \leqslant c$$

with probability at least $1 - e^{-\Omega(M)}$.

E.3.2 Upper and Lower Bounds

Let $\hat{\lambda}_j = \mu_j(\mathbf{W}\mathbf{H}\mathbf{W}^\top)$.

Lemma E.9. *With probability at least $1 - e^{-\Omega(M)}$, for $s > 0$ we have*

$$\operatorname{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}) = \sum_{j=1}^{M} e^{-2(t-\tau)\hat{\lambda}_j}\hat{\lambda}_j^2 \eqsim \mathcal{K}_M(t-\tau).$$

Proof. We can compute that

$$\begin{aligned}
\operatorname{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}) &= \operatorname{tr}(\mathbf{W}^\top\mathbf{W}\mathbf{H}\mathbf{W}^\top\mathbf{W}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}) \\
&= \operatorname{tr}\left(\mathbf{W}^\top\mathbf{W}\mathbf{H}\mathbf{W}^\top\mathbf{W}\sum_{a,b=0}^{\infty}\frac{1}{a!b!}(-(t-\tau))^{a+b}(\mathbf{H}\mathbf{W}^\top\mathbf{W})^a\mathbf{H}(\mathbf{W}^\top\mathbf{W}\mathbf{H})^b\right) \\
&= \operatorname{tr}\left(\sum_{a,b=0}^{\infty}\frac{1}{a!b!}(-t+\tau)^{a+b}\cdot\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b+1}\mathbf{W}\mathbf{H}\right) \\
&= \operatorname{tr}\left(\sum_{a,b=0}^{\infty}\frac{1}{a!b!}(-t+\tau)^{a+b}\cdot(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b+2}\right) \\
&= \sum_{a,b=0}^{\infty}\frac{1}{a!b!}(-t+\tau)^{a+b}\sum_{j=1}^{M}\hat{\lambda}_j^{a+b+2} \\
&= \sum_{j=1}^{M} e^{-2(t-\tau)\hat{\lambda}_j}\hat{\lambda}_j^2.
\end{aligned}$$

By Lemma E.7, $\hat{\lambda}_j \eqsim \lambda_j \eqsim j^{-\beta}$, hence the summation can be estimated by the integral

$$\sum_{j=1}^{M} e^{-2(t-\tau)\hat{\lambda}_j}\hat{\lambda}_j^2 \eqsim \int_1^M e^{-2(t-\tau)x^{-\beta}}x^{-2\beta}\,\mathrm{d}x = \int_{M^{-\beta}}^1 e^{-2(t-\tau)u}u^{2-\frac{1}{\beta}}\,\mathrm{d}u = \mathcal{K}_M(t-\tau),$$

where the last but second equality is a change of variable $u = x^{-\beta}$ in the integral. □

For the first term in Eq. (24), following [35], we have

Lemma E.10. *With probability at least $1 - e^{-\Omega(M)}$, for $s \leqslant 1$ we have*

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 \lesssim M^{-s\beta} + e(t).$$

Proof. Note

$$\mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t = \sum_{a,b=0}^{\infty}\frac{1}{a!b!}(-t)^{a+b}(\mathbf{H}\mathbf{W}^\top\mathbf{W})^a\mathbf{H}(\mathbf{W}^\top\mathbf{W}\mathbf{H})^b$$

$$= \sum_{a,b=0}^{\infty} \frac{1}{a!b!}(-t)^{a+b}\mathbf{H}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b-1}\mathbf{W}\mathbf{H}$$
$$= \mathbf{H}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}$$

where $\mathbf{M}_t = P_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)$ with P_t being the power series

$$P_t(x) = \sum_{a+b\geqslant 0} \frac{1}{a!b!}(-t)^{a+b}x^{a+b+1}.$$

Note that when $x \in \mathbb{R}$, we have $P_t(x) = xe^{-2tx}$, hence the eigenvalues of $\mathbf{M}_t$ is exactly $P_t(\hat{\lambda}_j)$. Since

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 = \mathbf{u}_0^\top(\mathbf{H}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H})\mathbf{u}_0,$$

for any positive integer $k \leqslant \frac{M}{2}$, note that $\mathbf{W}\mathbf{H}\mathbf{u} = \mathbf{W}_{0:k}\mathbf{H}_{0:k}\mathbf{u}_{0:k} + \mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{u}_{k:\infty}$, we have

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 \leqslant 2(T_1 + T_2),$$

where

$$T_1 = \mathbf{u}_{0:k}^\top(\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k})\mathbf{u}_{0:k},$$
$$T_2 = \mathbf{u}_{k:\infty}^\top(\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty})\mathbf{u}_{k:\infty}.$$

Then by Lemma E.13, we can derive an upper bound. Since $s \leqslant 1 < \beta$,

$$T_1 + T_2 \lesssim \frac{1}{t}\|\mathbf{u}_{0:k}\|_2^2 + \|\mathbf{u}_{k:\infty}\|_{\mathbf{H}_{k:\infty}}^2$$
$$\eqsim \frac{1}{t}\sum_{j=1}^{k} j^{-1-s(\beta-1)} + \sum_{j=k+1}^{N} j^{-1-s\beta} \eqsim \frac{k^{-s(\beta-1)}}{t} + k^{-s\beta}.$$

By setting $k = \min\{t^{1/\beta}, \frac{M}{3}\}$, we have

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 \lesssim \max\{t^{-s}, M^{-s\beta}\}.$$

Now when $M^\beta < t$, we have

$$e(t) \eqsim \frac{1}{t^s}.$$

and then the conclusion follows. □

Lemma E.11. *For $s > 0$, it holds with probability at least $1 - e^{-\Omega(M)}$ that*

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 \gtrsim \max\{e(t), M^{-s\beta}\}.$$

Proof. Following the proof of Lemma E.10, we have

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 = \mathbf{u}_0\mathbf{H}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}\mathbf{u}_0$$
$$= \operatorname{tr}\left((\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1} \cdot \mathbf{W}\mathbf{H}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}\mathbf{W}^\top\right)$$
$$\geqslant \sum_{i=1}^{M} \mu_{M-i+1}\left((\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\right) \cdot \mu_i\left(\mathbf{W}\mathbf{H}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}\mathbf{W}^\top\right),$$

where the last inequality is by Von Neumann's trace inequality. Note that $\mathbf{M}_t = P_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)$, we then get

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t\mathbf{u}_0 \geqslant \sum_{i=1}^{M} \mu_i\left((\mathbf{W}\mathbf{H}\mathbf{W}^\top)^2\mathbf{M}_t^{-1}\right)^{-1} \mu_i\left(\mathbf{W}\mathbf{H}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}\mathbf{W}^\top\right)$$
$$= \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i}\hat{\lambda}_i^{-1}\mu_i\left(\mathbf{W}\mathbf{H}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}\mathbf{W}^\top\right).$$

Note that $\mathbf{u}_0 = \mathbf{v}^*$, by Assumption 2.3 and 2.4,

$$\begin{aligned}
\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H} \mathbf{A}_t \mathbf{u}_0 &\geqslant \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i} \hat{\lambda}_i^{-1} \mu_i \left(\mathbf{W}\mathbf{H}\mathbf{u}_0\mathbf{u}_0^\top \mathbf{H}\mathbf{W}^\top \right) \\
&\eqsim \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i} \hat{\lambda}_i^{-1} \mu_i (\mathbf{W}\mathbf{H}^{1+s+1/\beta}\mathbf{W}^\top) \\
&\gtrsim \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i} \hat{\lambda}_i^{s+1/\beta} \\
&\eqsim \int_1^M e^{-2tx^{-\beta}} x^{-1-s\beta} \, \mathrm{d}x \\
&= \int_{M^{-\beta}}^1 e^{-2tu} u^{s-1} \, \mathrm{d}u = e(t).
\end{aligned}$$

Here we used Lemma E.7.

On the other hand, we prove the lower bound on $M^{-s\beta}$. First, we claim that

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H} \mathbf{A}_t \mathbf{u}_0 \geqslant \|(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})\mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\|^2 =: T_3,$$

which will be proved in the Lemma E.12

Notice that

$$T_3 = \left\langle \mathbf{I}_N - \mathbf{H}^{1/2}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{1/2}, \mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}^{\frac{1}{2}} \right\rangle$$

Therefore note that $\mu_i(\mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}^{\frac{1}{2}}) = i^{-1-s\beta}$ by source and capacity conditions,

$$\begin{aligned}
T_3 &\geqslant \sum_{i=1}^{N} \mu_i \left(\mathbf{I}_N - \mathbf{H}^{1/2}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{1/2} \right) \cdot \mu_{N+1-i}(\mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\mathbf{u}_0^\top\mathbf{H}^{\frac{1}{2}}) \\
&\gtrsim \sum_{i=1}^{N} \mu_i(\mathbf{M}) \cdot (N+1-i)^{-1-s\beta}
\end{aligned}$$

where the third line follows from Von-Neuman's Inequality. Since $\mathbf{M} = \mathbf{I}_N - \mathbf{H}^{1/2}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{1/2}$ is a projection matrix such that $\mathbf{M}^2 = \mathbf{M}$ and $\mathrm{rank}(\mathbf{I}_N - \mathbf{M}) = M$ with probability 1, in this case $\mathbf{M}$ has M eigenvalues 0 and $N - M$ eigenvalues 1.

Hence we have

$$T_3 \gtrsim \sum_{i=M}^{N} i^{-1-s\beta} \gtrsim M^{-s\beta}.$$

□

Lemma E.12.

$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H} \mathbf{A}_t \mathbf{u}_0 \geqslant \|(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})\mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\|^2$$

Proof. By the definition of positive semi-definite, we only need to prove that

$$\mathbf{A}_t^\top \mathbf{H} \mathbf{A}_t \succeq \mathbf{H}^{\frac{1}{2}}(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})^2\mathbf{H}^{\frac{1}{2}}$$

Notice that

$$\begin{aligned}
\mathbf{A}_t^\top \mathbf{H} \mathbf{A}_t &= e^{-\mathbf{H}\mathbf{W}^\top\mathbf{W}t}\mathbf{H}e^{-\mathbf{W}^\top\mathbf{W}\mathbf{H}t} \\
&= \mathbf{H}^{\frac{1}{2}} \left(\mathbf{I} + \sum_{a+b\geqslant 1} \frac{1}{a!b!}(-t)^{a+b}\mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}}
\end{aligned}$$

Notice that $\mathbf{H}$ is a positive definite matrix, and now we only need to prove

$$\mathbf{I}+\sum_{a+b\geqslant 1}\frac{1}{a!b!}(-t)^{a+b}\mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}}\succeq(\mathbf{I}-\mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})^2.$$

Let $\mathbf{P}=\mathbf{W}\mathbf{H}^{\frac{1}{2}}$. After simplification, we only need to prove that

$$\mathbf{I}+\sum_{a+b\geqslant 1}\frac{1}{a!b!}(-t)^{a+b}\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{a+b-1}\mathbf{P}\succeq\mathbf{I}-\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\mathbf{P}.$$

Notice that, by the definition of matrix exponential, we have

$$\begin{aligned}
&\mathbf{I}+\sum_{a+b\geqslant 1}\frac{1}{a!b!}(-t)^{a+b}\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{a+b-1}\mathbf{P}\\
&=\mathbf{I}-\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\mathbf{P}+\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\left(\sum_{a+b\geqslant 0}\frac{2^{a+b}}{(a+b)!}(-t)^{a+b}(\mathbf{P}\mathbf{P}^\top)^{a+b}\right)\mathbf{P}\\
&=\mathbf{I}-\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\mathbf{P}+\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}e^{-2\mathbf{P}\mathbf{P}^\top t}\mathbf{P}.
\end{aligned}$$

Notice that the matrix $\mathbf{P}^\top\mathbf{P}$ and $e^{-2\mathbf{P}\mathbf{P}^\top t}$ are both positive semi-definite, we have $\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}e^{-2\mathbf{P}\mathbf{P}^\top t}\mathbf{P}$ is positive semi-definite. As a result,

$$\mathbf{I}+\sum_{a+b\geqslant 1}\frac{1}{a!b!}(-t)^{a+b}\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{a+b-1}\mathbf{P}\succeq\mathbf{I}-\mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\mathbf{P}.$$

which completes the proof. □

Lemma E.13. *With probability $1-e^{-\Omega(M)}$, we have*

$$T_1\leqslant c\frac{\|\mathbf{u}_{0:k}\|_2^2}{t}\left(\frac{\mu_{\frac{M}{2}}(\mathbf{W}_{0:k}\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top)}{\mu_M(\mathbf{W}_{0:k}\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top)}\right)^2,\quad T_2\leqslant\|\mathbf{u}_{k:\infty}\|_{\mathbf{H}_{k:\infty}}^2.$$

where c is some constant.

Proof. First, we prove that

$$\|\mathbf{M}_t\|_2\leqslant\frac{c}{t}.$$

Note that the eigenvalues of $\mathbf{M}_t$ is $P_t(\hat{\lambda}_j)=\frac{f(2t\hat{\lambda}_j)}{2t}$, where

$$f(x_0)=x_0e^{-x_0}\leqslant\frac{1}{e}.$$

So we have

$$\|\mathbf{M}_t\|_2\leqslant\max_{1\leqslant j\leqslant M}P_t(\hat{\lambda}_j)\leqslant\frac{1}{2et}.$$

By definition of T_1, we have

$$\begin{aligned}
T_1&\leqslant\|\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|\|\mathbf{u}_{0:k}\|_2^2\\
&\leqslant\|\mathbf{M}_t\|_2\|(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|_2^2\|\mathbf{u}_{0:k}\|_2^2\\
&\leqslant\frac{c}{t}\|(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|_2^2\|\mathbf{u}_{0:k}\|_2^2.
\end{aligned}$$

We only need to show

$$\|(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|_2\leqslant c\left(\frac{\mu_{\frac{M}{2}}(\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top)}{\mu_M(\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top)}\right).$$

We denote $\mathbf{A}_k=\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top$, and since $\mathbf{W}\mathbf{H}\mathbf{W}^\top=\mathbf{W}_{0:k}\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top+\mathbf{A}_k$, we have

$$(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}=(\mathbf{A}_k^{-1}-\mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1}+\mathbf{W}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\mathbf{W}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{W}_{0:k})\mathbf{W}_{0:k}\mathbf{H}_{0:k}$$

$$
\begin{aligned}
&= \mathbf{A}_k^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k} - \mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k} \\
&= \mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1}\mathbf{H}_{0:k} \\
&= \mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}
\end{aligned}
$$

where the first line uses Woodbury's identity. Since

$$
\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k} \succeq \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}.
$$

it follows that

$$
\|[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2 \leqslant \|[\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2.
$$

Therefore, with probability at least $1 - e^{-\Omega(M)}$

$$
\begin{aligned}
\|\mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2 &\leqslant \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{W}_{0:k}\|_2 \cdot \|[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2 \\
&\leqslant \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{W}_{0:k}\|_2 \cdot \|[\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2 \\
&\leqslant \frac{\|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{W}_{0:k}\|_2}{\mu_{\min}(\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k})}.
\end{aligned}
$$

Assume $k \leqslant \frac{M}{2}$ and with probability at least $1 - e^{-\Omega(M)}$ for some constant $c > 0$, $\|\mathbf{W}_{0:k}\|_2 \leqslant c$.

We may write $\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k} = \sum_{i=1}^M \frac{1}{\hat{\lambda}_{M-i}} \boldsymbol{s}_i \boldsymbol{s}_i^\top$, where $\boldsymbol{s}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_k/N)$ and $(\hat{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{A}_k$ in non-increasing order. Therefore, for $k \leqslant M/3$,

$$
\sum_{i=1}^{M} \frac{1}{\hat{\lambda}_{M-i}} \boldsymbol{s}_i \boldsymbol{s}_i^\top \succeq \sum_{i=1}^{M/2} \frac{1}{\hat{\lambda}_{M-i}} \boldsymbol{s}_i \boldsymbol{s}_i^\top \succeq \frac{1}{\hat{\lambda}_{M/2}} \sum_{i=1}^{M/2} \boldsymbol{s}_i \boldsymbol{s}_i^\top \succeq \frac{c\mathbf{I}_k}{\hat{\lambda}_{M/2}}.
$$

with probability at least $1 - e^{-\Omega(M)}$, where in the last inequality we again use the concentration properties of Gaussian covariance matrices (see e.g., Theorem 6.1 in [60]).

$$
\begin{aligned}
\|\mathbf{A}_k^{-1}\mathbf{W}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k}]^{-1}\|_2 &\lesssim \frac{\|\mathbf{A}_k^{-1}\|_2}{\mu_{\min}(\mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1}\mathbf{W}_{0:k})} \\
&\leqslant \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}.
\end{aligned}
$$

Now we focus on T_2, by definition of T_2 we have

$$
\begin{aligned}
T_2 &= \mathbf{u}_{k:\infty}{}^\top \mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top (\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1/2} \exp(-2t\mathbf{W}\mathbf{H}\mathbf{W}^\top)(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1/2}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{u}_{k:\infty} \\
&\leqslant \mathbf{u}_{k:\infty}{}^\top \mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top (\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{u}_{k:\infty} \\
&\leqslant \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top (\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\| \cdot \|\mathbf{u}_{k:\infty}\|_{\mathbf{H}_{k:\infty}}^2 \\
&\leqslant \|\mathbf{u}_{k:\infty}\|_{\mathbf{H}_{k:\infty}}^2,
\end{aligned}
$$

where the last line follows from

$$
\begin{aligned}
&\|\mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top (\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\|_2 \\
&\qquad = \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top (\mathbf{W}_{0:k}\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top + \mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{W}_{k:\infty}^\top)^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\|_2 \\
&\qquad \leqslant \|\mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\|_2 = 1.
\end{aligned}
$$

The last line is because a nonzero projection matrix has norm 1. □

Combining Lemma E.9, E.10 and E.11, we get with probability at least $1 - e^{-\Omega(M)}$,

$$
\mathbb{E}[\mathcal{E}_t] \eqsim M^{-s\beta} + e_M(t) + \int_0^t \mathcal{K}_M(t-\tau)\gamma_\tau(c_\tau \mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau, \tag{30}
$$

which is of the same form as in Theorem E.4. From here following the same proof as before, we get the functional scaling laws for random projection matrices.

F Proofs for Section 5

When the condition $\sigma \gtrsim 1$ holds – indicating a constant label-noise level – the FSL simplifies to

$$\mathbb{E}[\mathcal{E}(\boldsymbol{\nu}_t)] \eqsim \frac{1}{M^{s\beta}} + \frac{1}{t^s} + \frac{\sigma^2}{B}\mathcal{N}(\varphi), \quad \text{with} \quad \mathcal{N}(\varphi) = \int_0^t \mathcal{K}_M(t-r)\varphi(T^{-1}(r))\,\mathrm{d}r, \tag{31}$$

where $e_M(t) + M^{-s\beta} \eqsim e_\infty(t) + M^{-s\beta} \eqsim t^{-s} + M^{-s\beta}$ as $e_\infty(t) - e_M(t) \lesssim M^{-s\beta}$, and the fit-dependent noise term $e_M(r)$ is absorbed by the label noise term due to $\sigma \gtrsim 1$. Extending to the full range $\sigma \geqslant 0$ is possible but makes the statements and derivations much more involved. We therefore focus on the above case to streamline the exposition.

F.1 Proofs for Constant LRS

In this section, we prove Theorem 5.2 and present the data-optimal scaling strategy, as well as some results related to the compute-optimal allocation.

Theorem F.1 (Restatement of Theorem 5.2)**.** *Under Assumption 5.1, when the learning rate $\eta(k) \equiv \eta$, for the top-M selection of the projection matrix $\mathbf{W}$ or for the random case with probability at least $1 - e^{-\Omega(M)}$, we have*

$$\mathbb{E}[\mathcal{R}_K] - \frac{\sigma^2}{2} \eqsim \frac{1}{(\eta K)^s} + \frac{\eta}{B}\sigma^2 + M^{-s\beta}.$$

Proof. By our main Theorem 4.2, when the learning rate $\eta(k) \equiv \eta$, denote $\gamma := \frac{\eta}{B}$, by (31) we have

$$\mathbb{E}[\mathcal{E}_K] \eqsim \gamma\sigma^2 + \frac{1}{t^s} + M^{-s\beta}.$$

Now we may write it as

$$\mathbb{E}[\mathcal{R}_K] - \frac{\sigma^2}{2} \eqsim \gamma\sigma^2 + \frac{1}{t^s} + M^{-s\beta}.$$

Notice that $t = \eta K$ and $\gamma = \frac{\eta}{B}$, we have

$$\mathbb{E}[\mathcal{R}_K] - \frac{\sigma^2}{2} \eqsim \frac{1}{(\eta K)^s} + \frac{\eta}{B}\sigma^2 + M^{-s\beta}.$$

□

Theorem F.2. *Given a total data size of $D \gg 1$, the optimal strategy for minimizing the final population risk, in terms of the effective learning rate γ and model size M is:*

$$\gamma_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}, \quad M_{\text{opt}} \gtrsim D^{\frac{1}{(1+s)\beta}}, \qquad \mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}. \tag{32}$$

Proof. Since we have

$$\mathbb{E}[\mathcal{E}_K] \eqsim \gamma\sigma^2 + \frac{1}{(\gamma D)^s} + M^{-s\beta},$$

By weighted AM-GM inequality, we have that when $\mathcal{E}_K$ is minimized, it must hold that

$$\gamma\sigma^2 \eqsim \frac{1}{(\gamma D)^s}$$

which gives

$$\gamma_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}.$$

Substituting this into the error expression yields

$$\mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s}{s+1}} + M^{-s\beta}.$$

To balance the two terms and achieve the optimal rate, we require

$$M_{\text{opt}} \gtrsim D^{\frac{1}{(1+s)\beta}}.$$

Consequently, the optimal loss rate becomes

$$\mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s}{s+1}}.$$

□

Next we consider the compute optimal strategy for constant learning rates. We define the compute $C = MKB$ to be the product of the model size, training steps and batch size.

Theorem F.3. *Given a total compute budget of $C \gg 1$, the optimal strategy for minimizing the final population risk, in terms of the effective learning rate γ, model size M, and data size $D := BK$, is:*

$$\gamma_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta+s\beta}}, M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta+s\beta}}, D_{\text{opt}} \eqsim C^{\frac{\beta+s\beta}{1+\beta+s\beta}},$$

Proof. Since we have

$$\mathbb{E}[\mathcal{E}_K] \eqsim \gamma\sigma^2 + \frac{1}{(\eta K)^s} + M^{-s\beta},$$

substituting $K = \frac{C}{MB}$, we get

$$\mathbb{E}[\mathcal{E}_K] \eqsim \gamma\sigma^2 + \frac{M^s}{(C\gamma)^s} + M^{-s\beta}.$$

By weighted AM-GM inequality, we have that when $\mathcal{E}_K$ is minimized, it must hold that

$$\gamma\sigma^2 \eqsim \frac{M^s}{(C\gamma)^s}, \quad \frac{M^s}{(C\gamma)^s} \eqsim M^{-s\beta},$$

which gives

$$\gamma_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta+s\beta}}, \quad M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta+s\beta}}.$$

Now we can further compute $D = BK = CM^{-1} \eqsim C^{\frac{\beta+s\beta}{1+\beta+s\beta}}$. □

F.2 Proof for The Exponential-Decay LRS

Recall that the LRS given by

$$\varphi(\tau) = ae^{-\lambda\tau}, \text{ with } \varphi(K) = b,$$

where $\lambda = \log(a/b)/K =: 1/\bar{K}$. Note that the intrinsic-time transform is given by

$$T(\tau) = \int_0^\tau \varphi(r)\,\mathrm{d}r = \frac{a}{\lambda}\left(1 - e^{-\lambda\tau}\right).$$

Thus, we have

- The total intrinsic time is:

$$T(K) = \frac{a}{\lambda}(1 - e^{-\lambda K}) = \frac{K}{\log(a/b)}(a - b) =: \bar{K}(a - b).$$

  For simplicity, we shall write $T = T(K)$ in what follows.
- The LRS-adjusted function in intrinsic time is given by

$$\gamma_\varphi(t) = \varphi(T^{-1}(t)) = a - \lambda t.$$

Lemma F.4. *The noise term satisfies $\mathcal{N}(\varphi) = bI_1 + (a - b)I_2$ with*

$$I_1 = \int_{M^{-\beta}}^1 \frac{1 - e^{-2uT}}{2u^{1/\beta}}\,\mathrm{d}u, \qquad I_2 = \int_{M^{-\beta}}^1 \left(\frac{1 - e^{-2uT} - 2uTe^{-2uT}}{4Tu^{1+1/\beta}}\right)\mathrm{d}u.$$

Proof. We use the integral to approximate the forgetting kernel $\mathcal{K}_M$ as

$$\mathcal{K}_M(t) \eqsim \sum_{j=1}^M j^{-2\beta}e^{-2j^{-\beta}t} \eqsim \int_1^M x^{-2\beta}e^{-2x^{-\beta}t}\,\mathrm{d}x \eqsim \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2ut}\,\mathrm{d}u.$$

Noticing $b = a - \lambda T$ and $\lambda T = a - b$, we have

$$\int_0^T \mathcal{K}_M(T - t)\gamma_\varphi(t)\,\mathrm{d}t = \int_0^T \left(\int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2u(T-t)}\,\mathrm{d}u\right)(a - \lambda t)\,\mathrm{d}t$$

$$
\begin{aligned}
&= \int_{M^{-\beta}}^{1} u^{1-1/\beta} e^{-2uT} \left(\int_0^T e^{2ut}(a-\lambda t)\,\mathrm{d}t \right) \mathrm{d}u \\
&= \int_{M^{-\beta}}^{1} u^{1-1/\beta} e^{-2uT} \left[\frac{a}{2u}\left(e^{2uT}-1\right) - \frac{\lambda}{2u}\left(Te^{2uT} - \frac{e^{2uT}-1}{2u} \right) \right] \mathrm{d}u \\
&= \int_{M^{-\beta}}^{1} \left[\frac{a}{2u^{1/\beta}} - \frac{ae^{-2uT}}{2u^{1/\beta}} - \frac{\lambda T}{2u^{1/\beta}} + \frac{\lambda(1-e^{-2uT})}{4u^{1+1/\beta}} \right] \mathrm{d}u \\
&= \int_{M^{-\beta}}^{1} \left[\frac{a-\lambda T}{2u^{1/\beta}} - \frac{(a-\lambda T+\lambda T)e^{-2uT}}{2u^{1/\beta}} + \frac{\lambda(1-e^{-2uT})}{4u^{1+1/\beta}} \right] \mathrm{d}u \\
&= (a-\lambda T) \int_{M^{-\beta}}^{1} \frac{1-e^{-2uT}}{2u^{1/\beta}}\,\mathrm{d}u + \lambda T \int_{M^{-\beta}}^{1} \left(-\frac{e^{-2uT}}{2u^{1/\beta}} + \frac{1-e^{-2uT}}{4Tu^{1+1/\beta}} \right) \mathrm{d}u.
\end{aligned}
$$

Thus, we complete the proof. □

We next bound I_1 and I_2 separately.

Lemma F.5. *If T and M is sufficiently large, then $I_1 = \frac{\beta}{2\beta-1} + o_{T,M}(1)$.*

Proof. Note that

$$
\int_{M^{-\beta}}^{1} \frac{1}{2u^{1/\beta}}\,\mathrm{d}u = \frac{\beta(1-M^{-(\beta-1)})}{2(\beta-1)} =: A.
$$

and

$$
\int_{M^{-\beta}}^{1} \frac{e^{-2uT}}{2u^{1/\beta}}\,\mathrm{d}u = \frac{1}{2(2T)^{1-1/\beta}} \int_{T/M^{\beta}}^{T} \frac{e^{-r}}{r^{1/\beta}}\,\mathrm{d}r \leqslant \frac{\Gamma(1+\frac{1}{\beta})}{2^{2-1/\beta}} \frac{1}{T^{1-1/\beta}} =: B.
$$

Then, we complete the proof by noting $I_1 = A - B$. □

Lemma F.6. *If T and M is sufficiently large, then*

$$
I_2 \eqsim \frac{\beta \min(M, T^{1/\beta})}{4T}.
$$

Proof. Let $r = uT$. Then, by a change of variable, we obtain

$$
I_2 = \frac{1}{4T^{1-1/\beta}} \int_{\frac{T}{M^{\beta}}}^{T} \frac{1-e^{-2r}-2re^{-2r}}{r^{1+1/\beta}}\,\mathrm{d}r =: \frac{1}{4T^{1-1/\beta}} \int_{\frac{T}{M^{\beta}}}^{T} q_\beta(r)\,\mathrm{d}r.
$$

It is easy to verify that for any $\beta \geqslant 1$, $\inf_{r\geqslant 0} q_\beta(r) \geqslant 0$ and $q_\beta(r) \eqsim r^{-1-1/\beta}$ when r is sufficiently large. We refer to Figure 10 for an illustration of $q_\beta(\cdot)$.

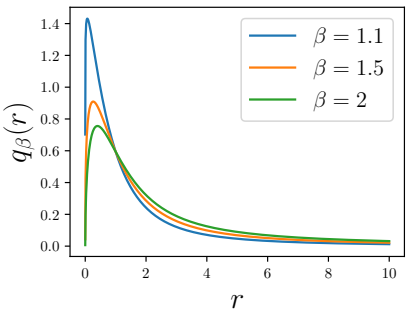


Figure 10: Illustration of the function $q_\beta(\cdot)$.

- When $T/M^{\beta} \leqslant 1$ and T is sufficiently large such that $\int_{\frac{T}{M^{\beta}}}^{T} q_\beta(r)\,\mathrm{d}r \eqsim \beta$ and thus we have

$$
I_2 = \frac{\beta + o_{M,T}(1)}{4T^{1-1/\beta}}.
$$

- When $T/M^\beta > 1$, it holds for all $r \geqslant 1$ that $0.5 \leqslant 1 - e^{-2r} - 2re^{-2r} \leqslant 1$. Thus, there exists a $C_{T,M} \in [0.5, 1]$ such that

$$I_2 = C_{T,M} \frac{1}{4T^{1-1/\beta}} \int_{\frac{T}{M^\beta}}^{T} r^{-1-1/\beta} \, \mathrm{d}r$$
$$= \frac{C_{T,M}\beta}{4T^{1-1/\beta}} \left(\left(\frac{T}{M^\beta} \right)^{-1/\beta} - T^{-1/\beta} \right) = \frac{C_{T,M}\beta(M-1)}{4T}.$$

Combining the two cases, we complete the proof. □

Theorem F.7 (Theorem 5.3 in the main paper)**.** *We consider the exponentially decaying learning rate schedule*

$$\varphi(\tau) = ae^{-\lambda\tau}, \text{ with } \varphi(K) = b,$$

Under this learning rate schedule, for the top-M projection matrix or the random projection with probability at least $1 - e^{-\Omega(M)}$, we have

$$\mathcal{E}_K \eqsim M^{-s\beta} + T^{-s} + \frac{\sigma^2}{B} \left(b + (a-b) \frac{\min\{M, T^{1/\beta}\}}{T} \right),$$

where $T = (a-b)K/\log(a/b)$ is the total intrinsic training time.

Proof. By the functional scaling laws (31),

$$\mathcal{E}_K \eqsim M^{-s\beta} + T^{-s} + \frac{\sigma^2}{B} \mathcal{N}(\varphi).$$

The noise term $\mathcal{N}(\varphi)$ is estimated by Lemma F.4 and the bound on I_1, I_2 as

$$\mathcal{N}(\varphi) = bI_1 + (a-b)I_2 \eqsim b + (a-b) \frac{\min(M, T^{1/\beta})}{T},$$

which gives

$$\mathcal{E}_K \eqsim M^{-s\beta} + T^{-s} + \frac{\sigma^2}{B} \left(b + (a-b) \frac{\min(M, T^{1/\beta})}{T} \right),$$

so we complete the proof. □

Theorem F.8. *Given a total data size $D \gg 1$, the optimal strategy for minimizing the final population risk when $b = \frac{a}{K}$ is given by $M_{\text{opt}} = \infty$ and*

- *If $s > 1 - \frac{1}{\beta}$, then $\gamma_{\text{opt}} \eqsim (D/\log D)^{-\frac{1+s\beta-\beta}{1+s\beta}}$ and $\mathcal{E}_{\text{opt}} \eqsim (D/\log D)^{-\frac{s\beta}{s\beta+1}}$.*
- *If $s \leqslant 1 - \frac{1}{\beta}$, then $\gamma_{\text{opt}} \eqsim 1$ and $\mathcal{E}_{\text{opt}} \eqsim (D/\log D)^{-s}$.*

Proof. Denote $\tilde{D} := \frac{D}{\log K}$, then by Theorem 5.3,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma\tilde{D})^{-s} + \frac{\min(M, (\gamma\tilde{D})^{\frac{1}{\beta}})}{\tilde{D}}.$$

Case 1. When $M^\beta \leqslant \gamma\tilde{D}$,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma\tilde{D})^{-s} + \frac{M}{\tilde{D}}.$$

We see that in this case γ should be as large as possible, since $a \lesssim 1$, we set $\gamma \eqsim 1$ accordingly.

In this case $M^{-s\beta} + \frac{M}{\tilde{D}} \gtrsim \tilde{D}^{-\frac{s\beta}{1+s\beta}}$, with equality at $M \eqsim \tilde{D}^{\frac{1}{1+s\beta}}$.

When $s > 1 - \frac{1}{\beta}$, the above equality condition can be acheived as $M^\beta = \tilde{D}^{\frac{\beta}{1+s\beta}} < \tilde{D}$. Hence we have that

$$M_{\text{opt}} \eqsim \tilde{D}^{\frac{1}{1+s\beta}}, \quad \gamma_{\text{opt}} \eqsim 1, \quad \mathcal{E}_{\text{opt}} \eqsim \tilde{D}^{-\frac{s\beta}{1+s\beta}}.$$

Note that $\gamma = \frac{a}{B} \eqsim 1$ and $a \lesssim 1$, which forces $B \eqsim 1$, hence $\tilde{D} \eqsim \frac{D}{\log D}$.

When $s \leqslant 1 - \frac{1}{\beta}$, the quantity $M^{-s\beta} + \frac{M}{\tilde{D}}$ is decreasing with respect to M, hence the optimal M in this case is $M = (\gamma\tilde{D})^{\frac{1}{\beta}}$, which transfers to case 2.

Case 2. When $M^\beta > \gamma\tilde{D}$,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma\tilde{D})^{-s} + \gamma^{\frac{1}{\beta}} \frac{1}{\tilde{D}^{1-\frac{1}{\beta}}}.$$

Clearly in this case $M_{\text{opt}} = \infty$, and by AM-GM inequality,

$$(\gamma\tilde{D})^{-s} + \gamma^{\frac{1}{\beta}} \frac{1}{\tilde{D}^{1-\frac{1}{\beta}}} \gtrsim \tilde{D}^{-\frac{s\beta}{1+s\beta}},$$

with equality at $\gamma \eqsim \tilde{D}^{\frac{\beta-1-s\beta}{1+s\beta}}$.

When $s > 1 - \frac{1}{\beta}$, the equality can be achieved, hence we have

$$M_{\text{opt}} = \infty, \quad \gamma_{\text{opt}} \eqsim \tilde{D}^{-\frac{1+s\beta-\beta}{1+s\beta}}, \quad \mathcal{E}_{\text{opt}} \eqsim \tilde{D}^{-\frac{s\beta}{1+s\beta}}.$$

When $s \leqslant 1 - \frac{1}{\beta}$, since $\gamma \lesssim 1$, we must have

$$M_{\text{opt}} = \infty, \quad \gamma_{\text{opt}} \eqsim 1, \quad \mathcal{E}_{\text{opt}} \eqsim \tilde{D}^{-s}.$$

Similarly, as $a \lesssim 1$, we have $B \lesssim \tilde{D}^{1-\frac{\beta}{1+s\beta}}$, which means $K \gtrsim \tilde{D}^{\frac{\beta}{1+s\beta}}$, hence $\log K \eqsim \log D$, $\tilde{D} \eqsim \frac{D}{\log D}$.

Summary. Combining the two cases together, we see that $M_{\text{opt}} = \infty$ can always achieves the optimal rate, hence the conclusion follows. □

Theorem F.9. *Given a large total compute budget $C \gg 1$, the optimal strategy for minimizing the final population risk – expressed in terms of the effective maximum learning rate γ, model size M, and data size D – is given by:*

- *When $s > 1 - \frac{1}{\beta}$, the optimal scaling laws are:*

  $$\gamma_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{-\frac{1+\beta(s-1)}{2+s\beta}}, \quad M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{2+s\beta}}, \quad D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1}{2+s\beta}},$$

  which leads to the following optimal final population risk:

  $$\mathcal{E}_{\text{opt}}(C) \eqsim (\tfrac{C}{\log C})^{-\frac{s\beta}{2+s\beta}}.$$

- *When $s \leqslant 1 - \frac{1}{\beta}$, the optimal scaling laws are*

  $$\gamma_{\text{opt}} \eqsim 1, \quad M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{1+\beta}}, \quad D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}} (\log C)^{\frac{1}{1+\beta}},$$

  which leads to the following optimal final population risk:

  $$\mathcal{E}_{\text{opt}}(C) \eqsim (\tfrac{C}{\log C})^{-\frac{s\beta}{1+\beta}}.$$

Proof. Denote $\tilde{D} = D/\log K$. For similar reasons as in the derivation of data-optimal scaling, we may assume $\log K \eqsim \log C$ to simplify the proof. At this point, the loss can be reformulated as follows.

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma\tilde{D})^s} + \sigma^2 \frac{\min\{M, (\gamma\tilde{D})^{1/\beta}\}}{\tilde{D}}.$$

Case 1. $M^\beta < \gamma\tilde{D}$ and we have

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma\tilde{D})^s} + \sigma^2 \frac{M}{\tilde{D}}$$

As γ only appears in the second term, and $\frac{1}{(\gamma\tilde{D})^s}$ is monotone decreasing with γ, we have that when $\mathcal{E}_K$ is minimized, it must hold that

$$M = (\gamma\tilde{D})^{1/\beta}.$$

When $s > 1 - \frac{1}{\beta}$, we then consider a weighted AM-GM inequality, we have

$$M^{-s\beta} = \sigma^2 \frac{M}{\tilde{D}}.$$

Combining with $C = MD$ and $M = (\gamma\tilde{D})^{1/\beta}$, we have

$$\gamma_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{-\frac{1+\beta(s-1)}{2+s\beta}}, \quad M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{2+s\beta}}, \quad D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1}{2+s\beta}},$$

and

$$\mathcal{E}_{\text{opt}}(C) \eqsim (\tfrac{C}{\log C})^{-\frac{s\beta}{2+s\beta}}.$$

When $s \leqslant 1 - \frac{1}{\beta}$, since $a \lesssim 1$, we set $\gamma_{\text{opt}} \eqsim 1$ accordingly, and proceed as follows:

$$M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{1+\beta}}, \quad D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}} (\log C)^{\frac{1}{1+\beta}},$$

and

$$\mathcal{E}_{\text{opt}}(C) \eqsim (\tfrac{C}{\log C})^{-\frac{s\beta}{1+\beta}}.$$

Case 2. $M^\beta \geqslant \gamma\tilde{D}$ and we have

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma\tilde{D})^s} + \sigma^2 \frac{(\gamma\tilde{D})^{1/\beta}}{\tilde{D}}$$

As M only appears in the second term, and $M^{-s\beta}$ is monotonically decreasing in M, we have that when $\mathcal{E}_K$ is minimized, it must hold that

$$M = (\gamma\tilde{D})^{1/\beta}.$$

And then the rest is identical to the first case. □

F.3 Proof for the WSD-Like LRS

To prove Theorem 5.4, we first present the following lemma, which gives an upper bound for the SGD noise induced by the stable phase.

Lemma F.10. *For $T_2 > 0$, we have*

$$\int_0^\infty \mathcal{K}_M(T_2 + t)\,\mathrm{d}t \lesssim \frac{\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}.$$

Proof. Similar to the previous section, we use integral to approximate the forgetting kernel $\mathcal{K}_M$ and get

$$\begin{aligned}
\int_0^\infty \mathcal{K}_M(T_2 + t)\,\mathrm{d}t &= \int_0^\infty \int_{M^{-\beta}}^1 u^{1-\frac{1}{\beta}} e^{-2u(T_2+t)}\,\mathrm{d}u\,\mathrm{d}t \\
&= \int_{M^{-\beta}}^1 u^{1-\frac{1}{\beta}} e^{-2uT_2} \frac{\mathrm{d}u}{2u} \\
&\eqsim \frac{1}{T_2^{1-\frac{1}{\beta}}} \int_{T_2 M^{-\beta}}^{T_2} u^{-\frac{1}{\beta}} e^{-2u}\,\mathrm{d}u.
\end{aligned}$$

Since the integral $\int_0^\infty u^{-\frac{1}{\beta}} e^{-2u}\,\mathrm{d}u$ is convergent, we have

$$\int_0^\infty \mathcal{K}_M(T_2 + t)\,\mathrm{d}t \lesssim \frac{1}{T_2^{1-\frac{1}{\beta}}}.$$

When $T_2 > M^\beta$, similarly we have

$$\int_0^\infty \mathcal{K}_M(T_2 + t)\,\mathrm{d}t \eqsim M^{1-\beta} \int_1^{M^\beta} u^{-\frac{1}{\beta}} e^{-2u\frac{T_2}{M^\beta}}\,\mathrm{d}u.$$

Let $p = \frac{T_2}{M^\beta} \geqslant 1$, we have

$$\begin{aligned}\int_0^\infty \mathcal{K}_M(T_2+t)\,\mathrm{d}t &\eqsim \frac{M}{T_2}p\int_1^{M^\beta} u^{-\frac{1}{\beta}}e^{-2up}\,\mathrm{d}u \\ &\lesssim \frac{M}{T_2}\int_1^{M^\beta} u^{-\frac{1}{\beta}}e^{-2u}\,\mathrm{d}u \eqsim \frac{M}{T_2}.\end{aligned}$$

Where the last line is because pe^{-2up} is decreasing in p when $u, p \geqslant 1$. □

Theorem F.11 (Theorem 5.4 in the main paper)**.** *Suppose the FSL (10) hold and M, K are sufficiently large. Then, we have*

$$\mathcal{E}_K \eqsim M^{-s\beta} + T^{-s} + \sigma^2\left(\frac{b}{B} + (a-b)\frac{\min\{M, T_2^{1/\beta}\}}{BT_2}\right),$$

where $T = aK_1 + (a-b)K_2/\log(a/b)$ is the total intrinsic training time, and $T_2 = (a-b)K_2/\log(a/b)$ is the decay-phase intrinsic training time.

Proof. By the results of the exponential decay LRS, let $\lambda = \log(a/b)/K_2$, we have

$$\int_0^{T(K)} \mathcal{K}_M(T(K)-t)\gamma_\varphi(t)\,\mathrm{d}t = \int_0^{T_1} \mathcal{K}_M(T(K)-t)a\,\mathrm{d}t + \int_0^{T_2} \mathcal{K}_M(T_2-t)(a-\lambda t)\,\mathrm{d}t,$$

Hence by the estimation of the noise term of the exponential decay LRS (see the proof of Theorem 5.3), we have

$$\int_0^{T_2} \mathcal{K}_M(T_2-t)(a-\lambda t)\,\mathrm{d}t \eqsim b + \frac{(a-b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}.$$

Thus, we know

$$\begin{aligned}\int_0^{T(K)} \mathcal{K}_M(T(K)-t)\gamma_\varphi(t)\,\mathrm{d}t &\eqsim \int_0^{T_1} \mathcal{K}_M(T(K)-t)a\,\mathrm{d}t + b + \frac{(a-b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2} \\ &\eqsim a\int_0^{T_1} \mathcal{K}_M(T_2+t)\,\mathrm{d}t + b + \frac{(a-b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2} \\ &\eqsim b + \frac{(a-b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}. \qquad \text{(by using Lemma F.10)}\end{aligned}$$

Hence the loss is given by

$$\mathcal{E}_K \eqsim \frac{1}{T^s} + M^{-s\beta} + \frac{\sigma^2}{B}\left(b + (a-b)\frac{\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}\right).$$

□

Theorem F.12. *Assume $b = \frac{a}{K}$, then we have the following data-optimal strategy:*

- *If $s \geqslant 1 - 1/\beta$, we have $\gamma_{\text{opt}} \eqsim D^{-\frac{1+s\beta-\beta}{1+s\beta}}(\log D)^{-\frac{\beta-1}{1+s\beta}}$, $(D_1)_{\text{opt}}, (D_2)_{\text{opt}} \eqsim D$ and $\mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s\beta}{s\beta+1}}(\log D)^{\frac{s\beta-s}{1+s\beta}}$.*
- *If $s < 1 - 1/\beta$, we have $\gamma_{\text{opt}} \eqsim 1$, $(D_1)_{\text{opt}} \eqsim D$, $(D_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}}\log D$ and $\mathcal{E}_{\text{opt}} \eqsim D^{-s}$.*

Proof. Since the total intrinsic time $T \lesssim \gamma D$, we can always take $D_1 \eqsim D$ to ensure $T \eqsim \gamma D$. Denote $\tilde{D}_2 := \frac{D_2}{\log K}$, then by Theorem F.11,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma D)^{-s} + \frac{\min(M, (\gamma\tilde{D}_2)^{\frac{1}{\beta}})}{\tilde{D}_2}.$$

Case 1. When $M^\beta \leqslant \gamma \tilde{D}_2$,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma D)^{-s} + \frac{M}{\tilde{D}_2}.$$

We see that in this case γ should be as large as possible, since $a \lesssim 1$, we set $\gamma \eqsim 1$ accordingly.

In this case $M^{-s\beta} + \frac{M}{\tilde{D}_2} \gtrsim \tilde{D}_2^{-\frac{s\beta}{1+s\beta}}$, with equality at $M \eqsim \tilde{D}_2^{\frac{1}{1+s\beta}}$.

When $s > 1 - \frac{1}{\beta}$, the above equality condition can be achieved as $M^\beta = \tilde{D}_2^{\frac{\beta}{1+s\beta}} < \tilde{D}_2$. Hence we have that

$$M_{\text{opt}} \eqsim \tilde{D}_2^{\frac{1}{1+s\beta}}, \quad \gamma_{\text{opt}} \eqsim 1, \quad \mathcal{E}_{\text{opt}} \eqsim \tilde{D}_2^{-\frac{s\beta}{1+s\beta}}.$$

Therefore $(D_2)_{\text{opt}} \eqsim D$. Note that $\gamma = \frac{a}{B} \eqsim 1$ and $a \lesssim 1$, which forces $B \eqsim 1$, hence $\tilde{D}_2 \eqsim \frac{D}{\log D}$.

When $s \leqslant 1 - \frac{1}{\beta}$, the quantity $M^{-s\beta} + \frac{M}{\tilde{D}_2}$ is decreasing with respect to M, hence the optimal M in this case is $M = (\gamma \tilde{D}_2)^{\frac{1}{\beta}}$, which transfers to case 2.

Case 2. When $M^\beta > \gamma \tilde{D}_2$,

$$\mathcal{E}_K \eqsim M^{-s\beta} + (\gamma D)^{-s} + \gamma^{\frac{1}{\beta}} \frac{1}{\tilde{D}_2^{1-\frac{1}{\beta}}}.$$

Clearly in this case $M_{\text{opt}} = \infty$, and by AM-GM inequality,

$$(\gamma D)^{-s} + \gamma^{\frac{1}{\beta}} \frac{1}{\tilde{D}_2^{1-\frac{1}{\beta}}} \gtrsim D^{-\frac{s}{1+s\beta}} \tilde{D}_2^{-\frac{s\beta-s}{1+s\beta}},$$

with equality at $\gamma \eqsim D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}}$.

When $s > 1 - \frac{1}{\beta}$, the equality can be achieved, hence we have that $(D_2)_{\text{opt}} \eqsim D$, so $\tilde{D}_2 \eqsim \frac{D}{\log K}$,

$$M_{\text{opt}} = \infty, \quad \gamma_{\text{opt}} \eqsim D^{-\frac{1+s\beta-\beta}{1+s\beta}} (\log K)^{-\frac{\beta-1}{1+s\beta}}, \quad \mathcal{E}_{\text{opt}} \eqsim D^{-\frac{s\beta}{1+s\beta}} (\log K)^{\frac{s\beta-s}{1+s\beta}}.$$

When $s \leqslant 1 - \frac{1}{\beta}$, since $\gamma \lesssim 1$, we must have either $\gamma \eqsim 1$ or $\gamma \eqsim D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}} \lesssim 1$. To reach the minimum risk, in both cases we require $(\tilde{D}_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}}$ (this gives $(D_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}} \log D$), and

$$M_{\text{opt}} = \infty, \quad \gamma_{\text{opt}} \eqsim 1, \quad \mathcal{E}_{\text{opt}} \eqsim D^{-s}.$$

Similarly, as $a \lesssim 1$, we have $B \lesssim_{\log} D^{1-\frac{\beta}{1+s\beta}}$, which means $K \gtrsim_{\log} D^{\frac{\beta}{1+s\beta}}$, hence $\log K \eqsim \log D$, which gives the desired rate.

Summary. Combining the two cases together, we see that $M_{\text{opt}} = \infty$ (case 2) always achieves the optimal rate, hence the conclusion follows. □

Theorem F.13. *Assume $b = \frac{a}{K}$, under the compute constraint $C \gg 1$, the optimal strategy for minimizing the final population risk—expressed in terms of the effective maximum learning rate γ, model size M, and data size D—is given by:*

- *When $s > 1 - 1/\beta$, the optimal scaling laws are:*

  $$\gamma_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{-\frac{1+s\beta-\beta}{2+s\beta}}, M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{2+s\beta}}, D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1}{2+s\beta}}, (D_1)_{\text{opt}} \eqsim D, (D_2)_{\text{opt}} \eqsim D,$$

  which leads to the following optimal final population risk:

  $$\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{2+s\beta}} (\log C)^{\frac{s\beta-s}{2+s\beta}}.$$

- *When $s \leqslant 1 - 1/\beta$, the optimal scaling laws are:*

$$\gamma_{\text{opt}} \eqsim 1, M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta}}, D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}}, (D_1)_{\text{opt}} \eqsim D, (D_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}} \log D,$$

which leads to the following optimal final population risk:

$$\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta}}.$$

Proof. Since the total intrinsic time $T \lesssim \gamma D$, we can always take $D_1 \eqsim D$ to ensure $T \eqsim \gamma D$. Denote $\tilde{D}_2 := \frac{D_2}{\log K}$, the loss can be reformulated as follows.

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma D)^s} + \sigma^2 \frac{\min\{M, (\gamma \tilde{D}_2)^{1/\beta}\}}{\tilde{D}_2}.$$

Case 1. $M^\beta < \gamma \tilde{D}_2$ and we have

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma D)^s} + \frac{M}{\tilde{D}_2}.$$

As γ only appears in the second term, and $\frac{1}{(\gamma D)^s}$ is monotone decreasing with γ, we have that when $\mathcal{E}_K$ is minimized, it must hold that

$$M = (\gamma \tilde{D}_2)^{1/\beta}.$$

When $s > 1 - \frac{1}{\beta}$, we then consider a weighted AM-GM inequality, we have

$$M^{-s\beta} = \frac{M}{\tilde{D}_2}.$$

Combining with $M = (\gamma \tilde{D}_2)^{1/\beta}$, we have

$$\gamma_{\text{opt}} \eqsim \tilde{D}_2^{-\frac{1+\beta(s-1)}{1+s\beta}}, \quad M_{\text{opt}} \eqsim \tilde{D}_2^{\frac{1}{1+s\beta}}$$

and

$$\mathcal{E}_{\text{opt}} \eqsim \tilde{D}_2^{s-\frac{s\beta}{1+s\beta}} D^{-s}.$$

Notice that

$$C \eqsim \tilde{D}_2^{\frac{1}{1+s\beta}} D \geqslant \tilde{D}^{\frac{2+s\beta}{1+s\beta}} \Longrightarrow \mathcal{E} \gtrsim C^{-\frac{s\beta}{2+s\beta}}.$$

Note that this implies $D^{\frac{2+s\beta}{1+s\beta}} \gtrsim C \gtrsim D \implies \log D \eqsim \log C$, and by similar reasons $\log K \eqsim \log D$ (the max learning rate $B\gamma \lesssim 1$).

Hence when $\mathcal{E}$ is optimized, we have $\tilde{D}_2 \eqsim D/\log C$ and

$$\gamma_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{-\frac{1+\beta(s-1)}{2+s\beta}}, \quad M_{\text{opt}} \eqsim (\tfrac{C}{\log C})^{\frac{1}{2+s\beta}}, \quad D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1}{2+s\beta}},$$

and

$$\mathcal{E}_{\text{opt}}(C) \eqsim (\tfrac{C}{\log C})^{-\frac{s\beta}{2+s\beta}} (\log C)^s.$$

When $s \leqslant 1 - \frac{1}{\beta}$, since $a \lesssim 1$, we set $\gamma_{\text{opt}} \eqsim 1$ accordingly, and proceed as follows:

$$M_{\text{opt}} \eqsim \tilde{D}_2^{\frac{1}{\beta}}$$

and

$$\mathcal{E}_{\text{opt}} \eqsim D^{-s}.$$

Notice that

$$C \eqsim \tilde{D}_2^{\frac{1}{\beta}} D \gtrsim \tilde{D}_2^{\frac{1+\beta}{\beta}} \Longrightarrow \mathcal{E} \gtrsim C^{-\frac{s\beta}{1+\beta}}.$$

Hence when $\mathcal{E}$ is optimized, we have $\tilde{D}_2 \eqsim D/\log C$ and

$$\gamma_{\text{opt}} \eqsim 1, M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta}}, D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}},$$

and

$$\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta}}.$$

Case 2. $M^\beta \geqslant \gamma \tilde{D}_2$ and we have

$$\mathcal{E}_K \eqsim M^{-s\beta} + \frac{1}{(\gamma D)^s} + \frac{(\gamma \tilde{D}_2)^{\frac{1}{\beta}}}{\tilde{D}_2}.$$

By AM-GM inequality,

$$(\gamma D)^{-s} + \gamma^{\frac{1}{\beta}} \frac{1}{\tilde{D}_2^{1-\frac{1}{\beta}}} \gtrsim D^{-\frac{s}{1+s\beta}} \tilde{D}_2^{-\frac{s\beta-s}{1+s\beta}},$$

with equality at $\gamma \eqsim D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}}$.

When $s > 1 - \frac{1}{\beta}$, the equality can be achieved, hence $(D_2)_{\text{opt}} \eqsim D$, and the loss can be written as follows.

$$\mathcal{E}_K \eqsim M^{-s\beta} + D^{-\frac{s}{1+s\beta}} \tilde{D}_2^{-\frac{s\beta-s}{1+s\beta}}$$

Combining with $C = MD$, we have the optimal scaling laws as follows:

$$\gamma_{\text{opt}} \eqsim C^{-\frac{1+s\beta-\beta}{2+s\beta}} (\log C)^{-\frac{\beta-1}{1+s\beta}}, M_{\text{opt}} \eqsim C^{\frac{1}{2+s\beta}} (\log C)^{-\frac{1-1/\beta}{2+s\beta}}, D_{\text{opt}} \eqsim C^{\frac{1+s\beta}{2+s\beta}} (\log C)^{\frac{1-1/\beta}{2+s\beta}},$$

which leads to the following optimal final population risk:

$$\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{2+s\beta}} (\log C)^{\frac{s\beta-s}{2+s\beta}}.$$

When $s \leqslant 1 - \frac{1}{\beta}$, since $\gamma \lesssim 1$, we must have either $\gamma \eqsim 1$ or $\gamma \eqsim D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}} \lesssim 1$. To reach the minimum risk, in both cases we require $(\tilde{D}_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}}$ (this gives $(D_2)_{\text{opt}} \gtrsim D^{\frac{s\beta}{\beta-1}} \log D$), and

$$\gamma_{\text{opt}} \eqsim 1, \quad \mathcal{E}_K \eqsim M^{-s\beta} + D^{-s}.$$

Combining with $C = MD$, we have the optimal scaling laws as follows:

$$\gamma_{\text{opt}} \eqsim 1, M_{\text{opt}} \eqsim C^{\frac{1}{1+\beta}}, D_{\text{opt}} \eqsim C^{\frac{\beta}{1+\beta}},$$

which leads to the following optimal final population risk:

$$\mathcal{E}_{\text{opt}} \eqsim C^{-\frac{s\beta}{1+\beta}}.$$

Summary. Combining the results of each case, we get the desired optimal scaling strategy stated in the theorem. □

G Auxiliary Lemmas

Lemma G.1. *For any PSD matrix* $\mathbf{A}$ *and a random gaussian vector* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$*,*

$$\text{tr}(\mathbf{HA})\mathbf{H} \preceq \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top - \mathbf{HAH}\right] = \text{tr}(\mathbf{HA})\mathbf{H} + \mathbf{HAH} \preceq 2\text{tr}(\mathbf{HA})\mathbf{H}$$

Proof. Assume $\mathbf{A} = (A_{ij})_{i,j=1,\dots,M}$. The (i,j)-th entry of $\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top$ is

$$\sum_{k,l} \mathbf{x}_i \mathbf{x}_k A_{kl} \mathbf{x}_l \mathbf{x}_j.$$

If $i \neq j$,

$$\mathbb{E}\left[\sum_{k,l} \mathbf{x}_i \mathbf{x}_k A_{kl} \mathbf{x}_l \mathbf{x}_j\right] = 2\mathbb{E}\left[A_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2\right] = 2A_{ij}\lambda_i\lambda_j = 2\mathbf{HAH}(i,j).$$

If $i = j$

$$\mathbb{E}\left[\sum_{k,l} \mathbf{x}_i \mathbf{x}_k A_{kl} \mathbf{x}_l \mathbf{x}_j\right] = \mathbb{E}\left[\sum_{k=1}^{M} A_{kk} \mathbf{x}_i^2 \mathbf{x}_k^2\right] = \sum_{k \neq i} A_{kk}\lambda_i\lambda_k + 3A_{ii}\lambda_i^2 = 2\mathbf{HAH}(i,i) + \text{tr}(\mathbf{HA})\mathbf{H}.$$

By the trace inequality we have

$$\mathbf{HA} \preceq \mathrm{tr}(\mathbf{HA}).$$

Multiplying $\mathbf{H}$ at both sides,

$$\mathbf{HAH} \preceq \mathrm{tr}(\mathbf{HA})\mathbf{H}.$$

Combining the results, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] = \mathrm{tr}(\mathbf{HA})\mathbf{H} + 2\mathbf{HAH} \preceq 2\mathrm{tr}(\mathbf{HA})\mathbf{H} + \mathbf{HAH}.$$

□

Lemma G.2. *Let $\mathbf{P} \preceq \mathbf{Q}$ be two PSD matrices. Then for any PSD matrix $\mathbf{U}$, we have*

$$\mathrm{tr}(\sqrt{\mathbf{P}}\mathbf{U}\sqrt{\mathbf{P}}) \leqslant \mathrm{tr}(\sqrt{\mathbf{Q}}\mathbf{U}\sqrt{\mathbf{Q}}).$$

Proof. It is clear that $\mathrm{tr}(\sqrt{\mathbf{P}}\mathbf{U}\sqrt{\mathbf{P}}) = \mathrm{tr}(\mathbf{UP})$ and

$$\mathrm{tr}(\mathbf{UQ}) - \mathrm{tr}(\mathbf{UP}) = \mathrm{tr}(\mathbf{U}(\mathbf{Q} - \mathbf{P})) \geqslant 0,$$

since $\mathbf{U}$ and $\mathbf{Q} - \mathbf{P}$ are both PSD matrices. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Specifically, we state that our work introduces a novel Functional Scaling Law (FSL) that captures the impact of learning-rate and batch-size schedules. These claims are substantiated by rigorous theoretical analysis (e.g., Theorem F.1), concrete examples (Section 5) and experiments in Section 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss several limitations and future directions in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper presents a complete set of assumptions and rigorous theoretical proofs for all main results. Assumptions 2.1, 2.3 and 2.4 define the problem setup, model capacity, and task difficulty. Detailed proofs of all the theoretical results are provided in Appendix E and F.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix C provides detailed specifications of the learning rate schedules, model sizes, number of steps, averaging procedures, and other hyper-parameters. The information provided is sufficient to reproduce the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While we provide detailed descriptions of all experimental setups and hyperparameters in Section 6 and Appendix C, we do not currently release code or datasets.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all necessary training details for reproducing the teacher–student kernel regression experiments in Appendix C.1. For LLM experiments, we specify the details in Appendix C.2

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the standard deviation as the error bar. This is clearly stated in Section 6 and Appendix C.1. While we do not explicitly verify the normality of the error distribution, the large number of samples ensures that the mean and standard deviation are reliable indicators of statistical trends.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: While we describe the experimental setup in full detail, we do not currently report the specific compute resources used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work does not involve human subjects, sensitive data, or deployment in real-world applications. All claims are rigorously supported by mathematical derivations and empirical validation, and we have taken care to ensure transparency, reproducibility, and fairness throughout the study.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our paper is a theoretical work and there is no societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not involve the release of any pretrained models, generative systems, or scraped datasets that pose a risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The code packages and open-source models used in this paper are all properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any datasets, pretrained models, or external code packages.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research does not involve human subjects or crowdsourced data collection. No participant interaction or compensation is involved at any stage of the work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects or any data collection from individuals.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not involve LLMs as any important, original or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.