Unveiling the Role of Learning Rate Schedules via Functional Scaling Laws

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]School of Computer Science, Peking University
[4]AI for Science Institute, Beijing
{libinghui, lean}@pku.edu.cn, alexpku@stu.pku.edu.cn
flchen_lwycc@stu.pku.edu.cn, leiwu@math.pku.edu.cn

Abstract

Scaling laws have played a cornerstone role in guiding the training of large language models (LLMs). However, most existing works on scaling laws primarily focus on the final-step loss, overlooking the loss dynamics during the training process and, crucially, the impact of learning rate schedule (LRS). In this paper, we aim to bridge this gap by studying a teacher–student kernel regression setup trained via online stochastic gradient descent (SGD). Leveraging a novel **intrinsic time** viewpoint and stochastic differential equation (SDE) modeling of SGD, we introduce the **Functional Scaling Law (FSL)**, which characterizes the evolution of population risk during the training process for general LRSs. Remarkably, the impact of the LRSs is captured through an explicit convolution-type functional term, making their effects fully tractable. To illustrate the utility of FSL, we analyze three widely used LRSs – constant, exponential decay, and warmup-stable-decay (WSD) – under both data-limited and compute-limited regimes. We provide theoretical justification for widely adopted empirical practices in LLMs pre-training such as (i) higher-capacity models are more data- and compute-efficient; (ii) learning rate decay can improve training efficiency; (iii) WSD-like schedules can outperform direct-decay schedules. Lastly, we explore the practical relevance of FSL as a surrogate model for fitting, predicting and optimizing the loss curves in LLM pre-training, with experiments conducted across model sizes ranging from 0.1B to 1B parameters. We hope our FSL framework can deepen the understanding of LLM pre-training dynamics and provide insights for improving large-scale model training.

1   Introduction

It has been empirically observed that the training of large-scale deep models follows certain *scaling laws* [23, 29, 24]: as model and dataset sizes increase, the performance improves *predictably* in a power-law fashion, i.e., the population loss L satisfies:

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

where M and D represent the model and dataset sizes respectively, α_M and α_D are the corresponding scaling exponents, L_0, C_M and C_D are constants. In practice, these parameters can be estimated from smaller-scale pilot experiments and then extrapolated to accurately predict performance at larger scales [61, 40, 32].

---

*Equal contribution.
†Corresponding author.

Table 1: **LRS impacts the scaling efficiency in the teacher-student kernel regression.** Easy-learning regime: $s \geqslant 1 - 1/\beta$; Hard-learning regime: $s < 1 - 1/\beta$, where s gauges task difficulty and β reflects model capacity.

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	$(D/\log D)^{-\frac{s\beta}{1+s\beta}}$	$(D/\log D)^{-s}$	$(C/\log C)^{-\frac{s\beta}{2+s\beta}}$	$(C/\log C)^{-\frac{s\beta}{1+\beta}}$
Warmup-stable-decay (WSD)	$D^{-\frac{s\beta}{1+s\beta}}(\log D)^{\frac{s\beta-s}{1+s\beta}}$	D^{-s}	$C^{-\frac{s\beta}{2+s\beta}}(\log C)^{\frac{s\beta-s}{2+s\beta}}$	$C^{-\frac{s\beta}{1+\beta}}$

Scaling laws have since become foundational principles in guiding the training of modern LLMs [22, 70, 28, 1, 9, 57, 31]. Practitioners heavily rely on these laws to optimally allocate computational resources between data volume and model size prior to training [24], as well as to tune optimization hyper-parameters such as learning rates and batch sizes [40, 32].

Given the cornerstone role of scaling laws in developing modern LLMs, explaining how these laws emerge – especially how the scaling exponents α_M and α_D depend on factors such as task difficulty and model capacity – has become a central topic of modern learning theory. Recently, a growing body of work has tackled this problem from diverse theoretical perspectives [56, 26, 43, 64, 27, 46, 48, 3, 18, 5, 10, 39, 51, 11, 71]. It is worth noting that unlike classical worst-case generalization-bound analyses [55], studying scaling laws demands explicitly tracking the learning trajectory for each instance. For this reason, nearly all recent theoretical studies adopt a teacher–student kernel regression setup due to its analytical tractability – a setting we also follow in this work.

A key limitation of existing theoretical studies is their focus on the scaling behavior of the final-step risk, while overlooking the loss dynamics during training process and, crucially, the role of the learning rate schedule (LRS). Yet, it is well established that LRS significantly affects training efficiency in deep learning [6, 44, 7, 21]. As a result, **different LRSs can induce distinct scaling behaviors, with the scaling exponent itself potentially depending on the schedule.** Developing a theoretical framework that captures the influence of LRS on scaling laws is not only of academic interest, but also practically relevant for designing effective training strategies.

1.1 Our Contribution

In this paper, we provide the first systematic study of the interaction between LRS and scaling laws. Working on the setup of teacher–student kernel regression – widely used in theoretical studies of scaling laws [43, 10, 5, 51, 39, 11], we introduce a novel **Functional Scaling Law (FSL)** that characterizes the risk dynamics under general LRSs. The impact of the schedules is captured through an explicit convolution-type functional term, making their effects fully transparent (Theorem 4.1).

To demonstrate the applicability and insights offered by FSL, we analyze three widely used LRSs, investigating their influence on scaling laws in both the traditional data-limited and modern compute-limited regimes. Our key findings are summarized as follows:

- **Constant LRS.** We recover prior results [10, 11, 51] and extend them by incorporating the effect of label noise, which models the irreducible risk presented in real-world LLM pre-training. Notably, compute-optimal training demands a learning rate that *decreases* with compute budget and higher-capacity models are more compute-efficient, matching empirical observations [9, 57].

- **Exponential-decay LRS.** We recover and refine the results of [39], now incorporating the effect of batch size. In particular, we show that exponential decay significantly boosts the scaling exponent compared to the constant LRS, providing a theoretical explanation for the cruciality of learning rate decay; see Table 1.

- **Warm-stable-decay (WSD) LRS.** We also analyze a WSD-like LRS that maintains a constant learning rate for a prolonged stable phase before applying exponential decay near the end of training. We reveal that such schedules can substantially outperform direct exponential decay in terms of optimal scaling efficiency (see Table 1), offering theoretical justification for the widespread empirical success of WSD strategies [70, 65, 25].

2

Across all three LRSs, we provide a detailed analysis of how the scaling exponents depend on task difficulty and model capacity. Our results show that **compute-optimal allocation consistently favors data over model parameters**, with the precise trade-off determined by the interplay between task difficulty and model capacity. Crucially, **higher-capacity models are uniformly more compute-efficient** across all settings considered. These findings offer a theoretical explanation for the widely adopted practice in scaling LLM training [24].

Finally, we also demonstrate the practical relevance of our FSL by applying it to LLM pre-training scenarios, providing theoretical insights into the empirical success of multiple-power-law heuristics proposed in [42].

From a technical perspective, our FSL is derived by modeling online SGD using an intrinsic-time stochastic differential equation (SDE). A key innovation in our analysis is the introduction of the concept of **intrinsic time**, which serves as a fundamental quantity for capturing the scaling behavior induced by learning-rate schedules. We believe that both the intrinsic-time SDE framework and the resulting FSL provide a versatile foundation for future investigations into the theoretical principles underlying learning rate scheduling and LLM pre-training dynamics.

1.2 Related Work

Theoretical explanation of scaling laws. Among the growing body of work seeks to theoretically explain scaling laws [56, 26, 43, 64, 27, 46, 48, 3, 18, 5, 10, 39, 51, 11], the most closely related to our work are [10, 51, 39, 11], which also analyze scaling laws for teacher–student kernel regression. [10] consider full-batch gradient flow, while [51] and [11] focus on online SGD. A key limitation of these studies is their omission of label noise, which models irreducible risk and plays a critical role in determining compute-optimal scaling. [39] considers exponential-decay LRS [19], but do not account for the impact of batch size tuning. Building on this line of work, we propose a unified FSL that applies to general LRSs, which only recovers prior results as special cases but also substantially extends them by capturing the compounding effects of LRS and label noise.

LRS-aware scaling laws. Recently, [60] heuristically proposed a momentum law to predict loss trajectories in LLM pre-training. Building on this, [42] introduced a heuristic multi-power law that demonstrates improved empirical performance. These empirical studies highlight the existence of scaling laws that explicitly depend on LRS. In this work, we provide the first theoretical explanation for this phenomenon, establishing a principled foundation for LRS-aware scaling behavior.

2 Preliminaries

Notations. For a $n \in \mathbb{N}$, let $[n] := \{1, 2, \ldots, n\}$. For a positive semi-definite (psd) matrix $\mathbf{S}$, we use $\mu_j(\mathbf{S})$ to denote its j-th largest eigenvalue and let $\|\mathbf{u}\|_{\mathbf{S}} = \sqrt{\mathbf{u}^\top \mathbf{S} \mathbf{u}}$ for a vector $\mathbf{u}$. For simplicity, we use $\eqsim$ to denote equivalence up to a constant factor, and $\lesssim$ (resp. $\gtrsim$) to indicate an upper (resp. lower) bound up to a constant factor.

2.1 Teacher–Student Kernel Regression

In this paper, we build upon the teacher–student kernel regression, a widely adopted, analytically tractable setup for studying scaling laws [43, 5, 10, 51, 39, 11]. In contrast to previous works, we provide an explicit analysis of how the LRS influences the scaling behavior. Formally, let $\mathbf{x} \in \mathbb{R}^d$ denote the data input that follows a distribution $\mathcal{D}$. The observed label is generated by $y := f^*(\mathbf{x}) + \epsilon$, where $f^* : \mathbb{R}^d \to \mathbb{R}$ is the target function and ϵ denotes the label noise. We assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and is independent of $\mathbf{x}$. The target function is assumed to be given by a **teacher model**:

$$f^*(\mathbf{x}) := \langle \phi(\mathbf{x}), \boldsymbol{\theta}^* \rangle,$$

where the target weight $\boldsymbol{\theta}^* \in \mathbb{R}^N$ with $N \in \mathbb{N}_+ \cup \{\infty\}$, and $\phi(\cdot) = (\phi_1(\cdot), \phi_2(\cdot), \cdots, \phi_N(\cdot))^\top : \mathbb{R}^d \to \mathbb{R}^N$ represents a feature map. We also assume the covariance matrix

$$\mathbf{H} := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\phi(\mathbf{x})\phi(\mathbf{x})^\top] = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$$

with eigenvalues $\lambda_1 \geqslant \lambda_2 \cdots \geqslant \lambda_N$. This diagonal assumption is made purely for simplicity of presentation, and all our results can be straightforwardly extended to the non-diagonal case.

Assumption 2.1 (Hypercontractivity condition). There exists constants $C_1, C_2 > 0$ such that it holds for any p.s.d. matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ that

$$C_1 \mathrm{tr}(\mathbf{HA})\mathbf{A} \preceq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\phi(\mathbf{x})^\top \mathbf{A} \phi(\mathbf{x}) \cdot \phi(\mathbf{x})\phi(\mathbf{x})^\top] - \mathbf{HAH} \preceq C_2 \mathrm{tr}(\mathbf{HA})\mathbf{A}.$$

This hypercontractivity condition [45] essentially requires that the fourth-order moments of $\phi(\mathbf{x})$ can be controlled by its second-order moments. It is satisfied, for example, by Gaussian features $\phi(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{H})$, as justified by Lemma E.1.

To learn f^*, we consider a finite-width **student model** defined by

$$f(\mathbf{x}; \mathbf{v}) := \langle \mathbf{v}, \mathbf{W}\phi(\mathbf{x})\rangle, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{M \times N}$ represents the projection matrix that projects the N-dimensional features into a M-dimensional subspace and $\mathbf{v} \in \mathbb{R}^M$ denotes the trainable weights. Here M can be interpreted as the *model size*. Concretely, we consider two choices of projector $\mathbf{W}$:

- **Top-M Features.** $\mathbf{W} = [\mathbf{I}_M, \mathbf{0}_{M \times (N-M)}]$. In this case, the student model essentially selects the top-M eigenfunctions $\{\phi_j\}_{j=1}^M$ as the projected features,

- **Random Features.** Each entry of $\mathbf{W}$ is i.i.d. sampled from $\mathcal{N}(0, 1/M)$.

Note that top-M features gives a theoretically convenient setting that simplifies our analysis – widely adopted in prior studies of scaling laws to facilitate tractable analysis [48, 17]. For random features [5, 10, 51, 39, 11], we will show that, in certain learning scenarios, their scaling behavior closely mirrors that of the top-M case.

2.2 The Model Capacity and Task Difficulty

We now formalize our key notions, *model capacity* and *task difficulty*, that will be pivotal in our subsequent scaling-law analysis. Let $e_j := \phi_j/\lambda_j^{1/2}$ for $j \in [N]$. Then, $\{e_j\}_{j=1}^N$ forms an orthonormal basis in $L^2(\mathcal{D})$. When $M = \infty$, the student's hypothesis space coincides with the reproducing-kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated with the kernel $k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}')\rangle = \sum_{j=1}^N \lambda_j e_j(\mathbf{x})e_j(\mathbf{x}')$ [4] (see a detailed discussion in Section C.1.1.

Assumption 2.2 (Model Capacity). Suppose $\lambda_j \asymp j^{-\beta}$ for some $\beta \in (1, \infty)$.

This condition ensures $\mathrm{Tr}(\mathbf{H}) := \sum_{j=1}^N \lambda_j < \infty$ and thus, the RKHS $\mathcal{H}_k$ is well defined. A smaller β implies slower eigenvalue decay, hence a larger RKHS $\mathcal{H}_k$. Consequently, the expressiveness of the student model (1) is governed by two orthogonal factors: **model size** M (the number of features used) and **capacity exponent** β (the intrinsic richness of the features). Conceptually, β can be viewed as an architecture-dependent quantity: models with different architectures may differ substantially in expressiveness even they have the same number of parameters.

Assumption 2.3 (Task difficulty). Suppose $\theta_j^* \asymp j^{-1/2}\lambda_j^{(s-1)/2}$ for all $j \in [N]$ for some $s \in (0, \infty)$.

Under Assumptions 2.2 and 2.3, the target function can be expressed as

$$f^* = \sum_{j=1}^N \theta_j^* \phi_j \asymp \sum_{j=1}^N j^{-1/2}\lambda_j^{s/2} e_j \overset{(i)}{\asymp} \sum_{j=1}^N j^{-(s\beta+1)/2} e_j,$$

where (i) follows from the assumption $\lambda_j \asymp j^{-\beta}$. Thus, the parameter $\alpha := s\beta$ characterizes the task's **intrinsic difficulty**, while s measures the **relative difficulty** with respect to the chosen model. For a fixed target (i.e., fixed α), choosing a higher-capacity model (smaller β) increases the value of $s = \alpha/\beta$, effectively making the task relatively easier. We will analyze, for a fixed target, how varying β, i.e., the model capacity, influences the corresponding scaling behavior.

We remark that similar capacity and source conditions have been widely used in the analysis of kernel methods [15, 14, 58, 12, 43], and recently applied to study their scaling behavior [51, 11]. Our work builds upon and extends this line of research.

2.3 Online Minibatch SGD

Given a data point $\mathbf{z} = (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and a student model $f(\cdot; \mathbf{v})$, we use the square loss to measure fitting error $\ell(\mathbf{z}, \mathbf{v}) := \frac{1}{2}(f(\mathbf{x}; \mathbf{v}) - y)^2$. Then, the population risk is given by

$$\mathcal{R}(\mathbf{v}) := \mathbb{E}_{\mathbf{z}}[\ell(\mathbf{z}, \mathbf{v})] = \frac{1}{2}\|\mathbf{W}^\top \mathbf{v} - \boldsymbol{\theta}^*\|_{\mathbf{H}}^2 + \frac{\sigma^2}{2} =: \mathcal{E}(\mathbf{v}) + \frac{\sigma^2}{2}, \tag{2}$$

where $\mathcal{E}(\mathbf{v})$ is the *excess risk*.

We consider to minimize $\mathcal{R}(\cdot)$ using online minibatch SGD. Specifically, at each step, we randomly sample a batch of *i.i.d.* samples $S_k := \{(\mathbf{x}_{k,j}, y_{k,j})\}_{j=1}^{B_k}$, where B_k is the batch size at the k-th step. Then, SGD iterates as follows:

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

where $\mathbf{v}_k$ denotes the weight of student model at the k-th step with $\mathbf{v}_0 = \mathbf{0}$, and $\eta_k \geqslant 0$ is the learning rate at the k-th step. Rewrite the SGD update (3) as:

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

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

Throughout this paper, we refer to $\boldsymbol{\eta} := (\eta_1, \eta_2, \cdots, \eta_K)^\top \in \mathbb{R}_{\geqslant 0}^K$ as the **learning rate schedule (LRS)**. In practical training of large-scale deep models, popular LRSs include the cosine schedule [41, 61], the warmup-stable-decay (WSD) schedule [25], and the multistep schedule [9] (see details in Appendix B.1). In addition, we refer to $\mathbf{b} = (B_1, B_2, \ldots, B_K)^\top \in \mathbb{N}_+^K$ as the batch size schedule.

A key insight for tractability is that the gradient noise exhibits the following anisotropic structure, whose proof is deferred to Appendix C.1.

Lemma 2.4 (Noise structure). *For any $\mathbf{v} \in \mathbb{R}^M$, it holds that*

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

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 $\Sigma(\mathbf{v}) \asymp \mathcal{R}(\mathbf{v}) \nabla^2 \mathcal{R}(\mathbf{v})$. 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 [68, 67].

3 The Intrinsic-Time Stochastic Differential Equations

We will leverage the view of SDE modeling of SGD, a framework that has proven effective in characterizing the dynamical behavior of SGD [33, 34, 35, 36, 52, 37]. By moving to continuous time [59], the dynamics become both more tractable and more interpretable.

Fix a discretization step size h and let $\varphi_k = \eta_k / h$ for all $k \in \mathbb{N}$. Then, (4) can be written as $\mathbf{v}_k = \mathbf{v}_{k-1} - \varphi_k \nabla \mathcal{R}(\mathbf{v}_{k-1}) h - \varphi_k h \boldsymbol{\xi}_k$. For small h, this recursion is approximated by the Itô-type *time-inhomogeneous* SDE [34, 50, 2]:

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

where $\mathbf{B}_\tau \in \mathbb{R}^M$ denotes the M-dimensional Brownian motion, and

- $\varphi \in C(\mathbb{R}_{\geqslant 0})$ is the continuous LRS function with $\varphi(kh) = \eta_k / h$ for all $k \in \mathbb{N}$;
- $b \in C(\mathbb{R}_{\geqslant 0})$ is the continuous batch-size schedule function with $b(kh) = B_k$ for all $k \in \mathbb{N}$.

One can explain τ as the continuous steps or physical time.

The intrinsic-time SDE. To isolate the effect of LRS, we introduce the concept of *intrinsic time*

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

which counts the intrinsic time accumulated by training for τ steps.

Let $\boldsymbol{\nu}_t = \bar{\mathbf{v}}_{T^{-1}(t)}$ and applying the Øksendal's time change formula [49] to (5) gives the intrinsic-time SDE:

$$\mathrm{d}\boldsymbol{\nu}_t = -\nabla\mathcal{R}(\boldsymbol{\nu}_t)\, \mathrm{d}t + \sqrt{\gamma_{\varphi,b}(t)\Sigma(\boldsymbol{\nu}_t)}\, \mathrm{d}\mathbf{B}_t \quad \text{with} \quad \gamma_{\varphi,b}(t) = \frac{h\varphi(T^{-1}(t))}{b(T^{-1}(t))}. \tag{6}$$

Note that $\gamma_{\varphi,b}(t)$ captures the effective gradient noise intensity, modulated by both the learning rate and batch size scheduling. Small learning rates and large batch sizes make this term smaller. Alternatively, the discrete SGD update (4) can be intuitively viewed as an "Euler–Maruyama discretization" of the SDE (6) on the *non-uniform* grid $\{t_k = \sum_{j=0}^k \eta_j\}_{k\in\mathbb{N}}$, as given by

$$\mathbf{v}_k - \mathbf{v}_{k-1} = -\nabla\mathcal{R}(\mathbf{v}_{k-1})(t_k - t_{k-1}) - \sqrt{\eta_k}\sqrt{t_k - t_{k-1}}\boldsymbol{\xi}_k.$$

4 Functional Scaling Laws

Before stating our results, we introduce two key functions:

$$e(t) := \int_{M^{-\beta}}^1 u^{s-1} e^{-2ut}\, \mathrm{d}u, \quad \mathcal{K}(t) := \int_{M^{-\beta}}^1 u^{1-\frac{1}{\beta}} e^{-2ut}\, \mathrm{d}u.$$

The function $e(\cdot)$ characterizes the risk convergence of full-batch GD in terms of intrinsic time. The function $\mathcal{K}(\cdot)$, termed the **forgetting kernel**, is a decreasing function that captures how the model gradually forgets the noise memorized in earlier training. When $M = \infty$, $e(t) \asymp t^{-s}, \mathcal{K}(t) \asymp t^{-(2-1/\beta)}$. For finite M, see Figure 1(left) for an illustration.
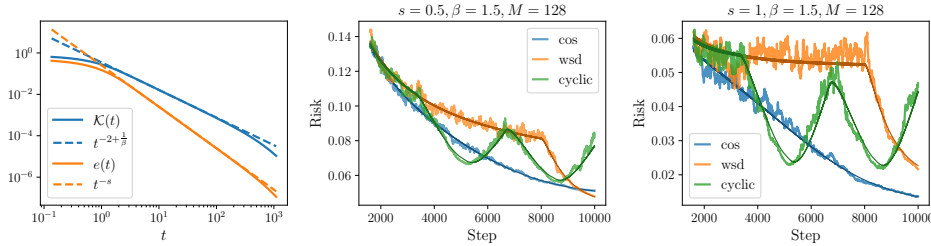

Figure 1: **(left)** The functions $e(\cdot)$ and $\mathcal{K}(\cdot)$. **(middle, right)** FSL accurately captures the average-risk dynamics of discrete SGD. Shaded regions indicate the mean over 200 independent runs, while solid lines represent FSL predictions. Results are shown for three LRSs: cosine, WSD-like, and a non-standard cyclic schedule. Additional experimental details and results for various values of s and β are provided in Appendix B.2.

Our main result is stated in the following theorem, whose proof can be found in Appendix C.

Theorem 4.1 (Intrinsic-time Functional Scaling Law (FSL))**.** *Suppose Assumptions 2.2 and 2.3 hold. Then, for the intrinsic-time SDE with top-M features and $s \in (0, 2 - 1/\beta]$, the following FSL holds for sufficiently large M and t:*

$$\mathbb{E}[\mathcal{R}(\boldsymbol{\nu}_t)] - \underbrace{\frac{1}{2}\sigma^2}_{\text{irreducible}} \asymp \underbrace{\frac{1}{M^{s\beta}}}_{\text{approximation}} + \underbrace{\frac{1}{t^s}}_{\text{full−batch GD}} + \underbrace{\int_0^t \mathcal{K}(t-r)(e(r) + \sigma^2)\gamma_{\varphi,b}(r)\, \mathrm{d}r}_{\text{noise term } \mathcal{N}(\varphi,b)} \tag{7}$$

For the random feature case, the same FSL holds for $s \in [0, 1]$ with probability at least $1 - e^{-\Omega(M)}$ with respect to the randomness of $\mathbf{W}$.

6

This theorem establishes a scaling law for the expected population risk of the intrinsic-time SDE (6), where the influence of scheduling learning rate and batch size is explicitly described via the convolution-type *noise functional:* $(\varphi, b) \mapsto \mathcal{N}(\varphi, b)$. Notably, each term of our FSL is interpretable:

- **The irreducible risk** $(= \frac{1}{2}\sigma^2)$ is due to the label noise.
- **The model-size scaling** (approximation error): The corresponding scaling exponent depends only on the intrinsic difficulty $s\beta$.
- **The intrinsic-time scaling** (full-batch GD term): The scaling exponent depends only on the relative difficulty s. Thus, for a fixed target, choosing a higher-capacity model (smaller β) make it converge faster.
- **The noise term** $\mathcal{N}(\varphi, b)$ exhibits a simple convolution form, capturing the risk arising from fitting the noise injected by SGD during training. Specifically, $(e(r) + \sigma^2)\gamma_{\varphi,b}(r)$ denotes the magnitude of gradient noise, adjusted by the learning-rate and batch-size schedules, at time r. More precisely, $e(r)\gamma_{\varphi,b}(r)$ represents the noise arising from mini-batch approximation, while $\sigma^2\gamma_{\varphi,b}(r)$ corresponds to the noise introduced by label noise. The forgetting kernel $\mathcal{K}(t - r)$ represents how much fitting noise at time r is forgotten by time t.

FSL captures the risk dynamics of SGD. Figure 1 (middle, right) compares the average risk dynamics of discrete SGD with the predictions of FSL. Despite being derived in continuous time, FSL closely captures the empirical behavior of SGD across different LRSs, task difficulties, and model capacities.

The trade-off between increasing intrinsic time and reducing noise injection. The adjustment factor $\gamma_{\varphi,b}(\cdot)$ encodes the influence of learning-rate and batch-size schedules for noise injection. Smaller learning rates or larger batches reduce this factor, but they also shrink the intrinsic training time. This time budget is crucial in two ways: (i) the full-batch GD term needs it to drive the risk down, and (ii) the forgetting kernel needs it to wash out noise memorized in the early iterations. A good schedule choice should therefore strike a compromise – **suppressing injected noise while still preserving enough intrinsic time to accomplish both objectives**.

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 (6) reduces to the gradient flow ODE: $d\boldsymbol{\nu}_t = -\mathbf{W}^\top\mathbf{W}\mathbf{H}\boldsymbol{\nu}_t\, dt$. 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:

$$\mathcal{R}(\boldsymbol{\nu}_t) - \frac{1}{2}\sigma^2 \asymp \sum_{j=1}^{M} \lambda_j |\theta_j^*|^2 e^{-2\lambda_j t} + \sum_{j=M+1}^{N} \lambda_j |\theta_j^*|^2.$$

Intuitively, we can view the learning of each eigenfunction as a subtask. 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 $\sum_{j=1}^{M} \lambda_j(\theta_j^*)^2 e^{-2\lambda_j t} \asymp \int_0^1 u^{s-1} e^{-2ut}\, du \asymp \frac{1}{t^s}$.

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

5 The Role of Learning Rate Schedules

In this section, we illustrate how our FSL can help understand the influences of LRS on scaling laws. For simplicity, we make the following assumption:

Assumption 5.1. Assume $s \in (0, 2 - 1/\beta]$ and $b(\tau) = B$ for all $\tau \geqslant 0$, and the FSL (7) holds.

When the condition $s \leqslant 2 - 1/\beta$ holds – indicating a relatively hard task – the FSL simplifies to

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

where the fit-dependent noise term $e(r)$ is absorbed by the full-batch GD term due to $s \leqslant 2 - 1/\beta$. Extending to the full range $s \geqslant 0$ is possible but makes the statements and derivations much more involved. We therefore focus on the above case to streamline the exposition.

To connect our continuous-time analysis with discrete SGD, we study the excess risk after $K = \tau/h$ steps and write $\mathcal{E}_K := \mathbb{E}[\mathcal{E}(\boldsymbol{\nu}_{Kh})]$. We then derive scaling laws characterizing $\mathcal{E}_K$ as a function of the model size M and total training step K, and how the associated scaling exponents vary with the model capacity, characterized by β. Specifically, two canonical resource-constrained settings are analyzed: (i) the traditional data-limited regime, where the data size $D := BK$ is fixed; and (ii) the modern compute-limited regime [24], where the compute $C := MD$ is fixed.

5.1 Constant LRS

For constant LRS, we have the following result, with the proof deferred to Appendix D.1.

Theorem 5.2 (Scaling law for constant LRS). *Under Assumption 5.1, we have*

$$\mathcal{E}_K \asymp M^{-s\beta} + (\eta K)^{-s} + \frac{\eta}{B}\left(\sigma^2 + (\eta K)^{-(2-\frac{1}{\beta})}\right).$$

In this case, the noise term is simplified to $\frac{\eta}{B}\left(\sigma^2 + (\eta K)^{-(2-\frac{1}{\beta})}\right)$. The fit-dependent component $(\eta K)^{-(2-\frac{1}{\beta})}$ is indeed dominated by the full-batch GD term $(\eta K)^{-s}$ when $s \leqslant 2 - \frac{1}{\beta}$. In the subsequent statement, we shall omit this term for clarity. Let $\gamma := \eta/B$ be the *effective learning rate*. Then, the scaling law can be rewritten as

$$\mathcal{E}_K \asymp M^{-s\beta} + (\gamma D)^{-s} + \gamma\sigma^2 =: h(\gamma, M, D),$$

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) [30, 20, 44].

Data-optimal scaling. Clearly, this involves minimizing $h(\cdot)$ while keeping D fixed. A straightforward calculation yields:

$$\gamma_{\mathrm{opt}} \asymp D^{-\frac{s}{s+1}}, \quad M_{\mathrm{opt}} \gtrsim D^{\frac{1}{(1+s)\beta}}, \qquad \mathcal{E}_{\mathrm{opt}} \asymp D^{-\frac{s}{s+1}}. \tag{8}$$

We observe that both the best achievable excess risk $\mathcal{E}_{\mathrm{opt}}$ 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 D.1:

$$\gamma_{\mathrm{opt}} \asymp C^{-\frac{s\beta}{1+(s+1)\beta}}, \ M_{\mathrm{opt}} \asymp C^{\frac{1}{1+(s+1)\beta}}, \ D_{\mathrm{opt}} \asymp C^{\frac{(s+1)\beta}{1+(s+1)\beta}}, \qquad \mathcal{E}_{\mathrm{opt}} \asymp C^{-\frac{s\beta}{1+s\beta+\beta}}.$$

This shows that the performance of the compute-optimal model improves with the budget C in a power law. For a fixed task ($\alpha = s\beta$ fixed), **increasing the model capacity ($\beta \downarrow$) can enhance the compute efficiency** – the extra β in the scaling exponent $\frac{s\beta}{1+s\beta+\beta}$ quantifies this gain. This provides an explanation of practical observations in LLM pre-training: Large models are more compute efficient than small models [29, 24]. In addition, the optimal learning rate γ_{opt} decreases as C grows and the optimal strategy allocates compute more aggressively to data than to model – again consistent with current LLM pre-training practices [9, 57, 24].

[11] also studied the compute-optimal scaling for a similar setting but assumed a fixed learning rate and an absence of label noise. In contrast, we consider a more realistic scenario where the learning rate is optimally tuned and the irreducible risk is present. This enables us to derive a compute-optimal scaling that matches empirical observations.

5.2 Exponential Decay LRS

We next consider the exponential decay (exp-decay) LRS for a given number of training steps K [19, 66]:

$$\varphi(\tau) = a\exp(-\lambda\tau), \qquad \varphi(Kh) = b,$$

where λ is chosen such that $\varphi(Kh) = b$. For brevity, we assume $h = 1$. Note that the two hyperparameters a and b denote the maximal and minimum learning rates, respectively.

Theorem 5.3 (Scaling law for exp-decay LRS). *Under Assumption 5.1, we have*

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

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

Let $a = b/K$. Then, the intrinsic time becomes $T = a(K - 1)/\log K$. In contrast, a constant LRS with $\eta = a$ achieves $T = aK$. Thus, exp-decay LRS incurs only a $\log K$ reduction in the intrinsic time, even if it drives the learning rate below a/K. Let $\gamma = a/B$ be the maximal effective learning rate.

Data-optimal scaling. In this regime, we have $M_{\text{opt}} = \infty$ and:

- If $s \geqslant 1 - 1/\beta$ (easy-learning), then $\gamma_{\text{opt}} \asymp (D/\log D)^{-\frac{1+s\beta-\beta}{1+s\beta}}$ and $\mathcal{E}_{\text{opt}} \asymp (D/\log D)^{-\frac{s\beta}{s\beta+1}}$.
- If $s < 1 - 1/\beta$ (hard-learning), then $\gamma_{\text{opt}} \asymp 1$ and $\mathcal{E}_{\text{opt}} \asymp (D/\log D)^{-s}$.

The derivation can be found in Appendix D.2. The hard-learning regime arises due to the stability constraint: learning rates must remain bounded when scaling D up. Relative to the constant LRS (see Eq. (8)), exp-decay LRS yields strictly faster excess-risk decay. This provides a formal justification for the importance of learning-rate decay in stochastic optimization.

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

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

In the easy-learning regime, we again observe that compute-optimal allocation consistently favors data over model. Notably, the optimal compute split depends solely on the task's intrinsic difficulty and the relative fraction of data over model: $D_{\text{opt}}/M_{\text{opt}} \asymp C^{\alpha/(2+\alpha)}$ *decreases* as the task becomes harder. The optimal γ_{opt} also decays with the compute budget C, and for fixed α, higher-capacity models ($\beta \downarrow$) require smaller γ_{opt}. Moreover, the excess-risk rate is governed solely by $\alpha = s\beta$, so increasing capacity alone does not yield asymptotic gains in this easy-learning regime.

In the hard-learning regime, data still dominates compute allocation, but now the optimal split depends only on the model capacity β, independent of the relative task difficulty. Moreover, the optimal maximal learning rate remains constant ($\gamma_{\text{opt}} \asymp 1$). 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 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 [70, 25], 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

$$\varphi(\tau) = \begin{cases} a & , \text{ if } \tau \leqslant K_1 h; \\ a\exp(-\lambda(\tau - K_1 h)) & , \text{ if } \tau > K_1 h. \end{cases} \tag{9}$$

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 largest learning rate a, the final learning rate b, and the decay proportion r controlling the decay-phase length. (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* (9):

$$\mathcal{E}_K \asymp M^{-s\beta} + (T_1 + T_2)^{-s} + \sigma^2 \left(\frac{b}{B} + (a-b)\frac{\min\{M, T_2^{1/\beta}\}}{BT_2} \right),$$

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 can see that WSD-like LRS can leverage the initial stable phase to boost the intrinsic training time. In particular, for any decay-phase fraction $r < 1$, we have $T = T_1 + T_2 \geqslant (1-r)Ka$, which far exceeds the the intrinsic time $T \asymp aK/\log K$ 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

- If $s \geqslant 1 - 1/\beta$, then $\gamma_{\text{opt}} \asymp D^{-\frac{1+s\beta-\beta}{1+s\beta}}(\log D)^{\frac{\beta-1}{1+s\beta}}$, $r_{\text{opt}} \in (0,1)$ and $\mathcal{E}_{\text{opt}} \asymp D^{-\frac{s\beta}{s\beta+1}}(\log D)^{\frac{s\beta-s}{1+s\beta}}$.

- If $s < 1 - 1/\beta$, then $\gamma_{\text{opt}} \asymp 1$, $r_{\text{opt}} \gtrsim D^{-\frac{\beta-1-s\beta}{\beta-1}}\log D$ and $\mathcal{E}_{\text{opt}} \asymp D^{-s}$.

Compared with the exp-decay LRS, both regimes enjoy an additional logarithmic improvement in excess-risk decay. Particularly, in the hard-learning regime, the logarithmic factor is completely eliminated. To achieve this, the **decay-phase duration only needs to scale sublinearly with D**, as indicated by $r_{\text{opt}} \to 0$ as $D \to \infty$. This observation aligns with the WSD practice in LLM pre-training – where the decay phase typically occupies only 10% of training. Moreover, our theory suggests that for harder tasks, the decay fraction could be reduced even further to enhance efficiency.

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

- If $s \geqslant 1 - 1/\beta$, then $\gamma_{\text{opt}} \asymp (C/\log C)^{-\frac{1+s\beta-\beta}{2+s\beta}}$, $r_{\text{opt}} \in (0,1)$, $M_{\text{opt}} \asymp (C/\log C)^{\frac{1}{2+s\beta}}$, $D_{\text{opt}} \asymp C^{\frac{1+s\beta}{2+s\beta}}(\log C)^{\frac{1}{2+s\beta}}$, and $\mathcal{E}_{\text{opt}} \asymp C^{-\frac{s\beta}{2+s\beta}}(\log C)^{\frac{s\beta-s}{2+s\beta}}$.

- If $s < 1 - 1/\beta$, then $\gamma_{\text{opt}} \asymp 1$, $r_{\text{opt}} \gtrsim D^{-\frac{\beta-1-s\beta}{\beta-1}}\log D$, $M_{\text{opt}} \asymp C^{\frac{1}{1+\beta}}$, $D_{\text{opt}} \asymp C^{\frac{\beta}{1+\beta}}$, and $\mathcal{E}_{\text{opt}} \asymp C^{-\frac{s\beta}{1+\beta}}$.

The proofs of this subsection can be found in Appendix D.3.

6 Experiments

6.1 Teacher–Student Kernel Regression

We begin by empirically assessing how LRS affects the scaling behavior of kernel-regression, thereby validating our theoretical predictions. The three – constant, exp-decay, and WSD-like – LRSs analyzed in Section 5 are evaluated on both easy ($s \geqslant 1 - 1/\beta$) and hard ($s < 1 - 1/\beta$) tasks. To keep the experiments lightweight, we confine our study to the data-limited regime; full implementation details are provided in the Appendix B.2.

Figure 2a compares the scaling laws obtained by numerical computation using the FSL (7) and the ones predicted by our analytical analyses in Section 5. We see that the scaling curves exhibit strong agreement across both easy and hard learning regimes for all three tested LRSs.

Figure 2b repeats the comparison using discrete SGD for a hard-learning task. Each configuration is run for 200 independent trials; we plot the mean risk together with the standard deviation. Figure 2b(left) confirms that the empirical scaling laws for SGD closely match the theoretical predictions.
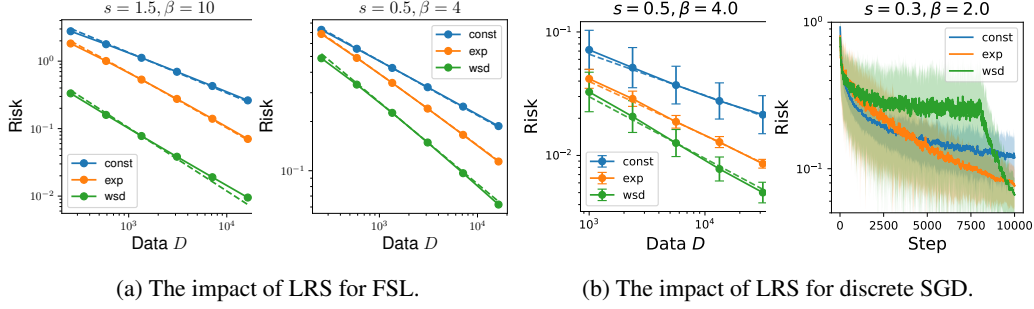
(a) The impact of LRS for FSL. (b) The impact of LRS for discrete SGD.

Figure 2: The impact of LRS on the scaling behavior of teacher–student kernel regression. In all panels, dashed curves represent the theoretical predictions from Section 5. **(a)** Solid curves show the risk predicted via numerical simulation using the FSL, where the maximal learning rate is scaled as $\eta_{\max} = 0.05 D^{-r}$, where the exponent r follows the theory in Section 5 for each LRS. **(b)** Discrete-SGD experiments: (left) empirical scaling laws; (right) risk trajectories under the three LRSs, each with well-tuned $\eta_{\max}$.

Among these LRSs, WSD indeed achieves the lowest risk, followed by exponential decay, with the constant schedule performing the worst. Figure 2b(right) further compares the risk trajectories of SGD under the three LRSs. While WSD decays much more slowly than exp-decay during the stable phase, it initiates a sharper drop once the decay phase begins, ultimately converging to a lower final risk. This counterintuitive pattern mirrors observations in LLM pre-training [25, 65].

6.2 LLM Pre-training

This subsection assesses the practical relevance of FSL – originally derived for teacher–student kernel regression – as a surrogate model for capturing the loss dynamics of LLM pre-training.

FSL can fit and predict loss curves. We first quantify the descriptive and predictive accuracy of FSL. Following the protocol of [60] and [42], we restrict attention to the post-warmup portion of the training trajectory. Two LLaMA [61] models (400 M and 1 B) are trained on 20 B tokens under three LRSs: cosine, WSD, and the 8-1-1 [9]. 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 rate.

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 dynamics for the cosine and WSD schedules. Figure 3a demonstrates that FSL not only fits the 8-1-1 trajectory accurately but also generalizes reliably to the unseen WSD and cosine LRSs, across both model sizes.
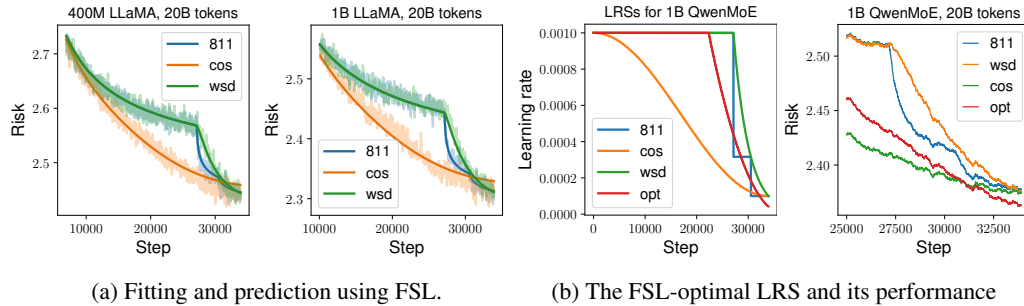

(a) Fitting and prediction using FSL. (b) The FSL-optimal LRS and its performance

Figure 3: **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.

The FSL-optimal LRS. We next ask whether the fitted FSL can be leveraged to *design* better LRSs. To this end, we numerically minimize the terminal risk over the space of LRSs using the fitted FSL (see Appendix B.3 for details). In this experiment, we consider a 1B-parameter mixture-of-experts QwenMoE model [69], trained on 20B tokens using the three previously discussed LRSs; see

Figure 3b (left). 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 schedule, using the same compute budget, and compared against the baseline LRSs. The results are summarized in Figure 3b.

Figure 3b(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 mirrors recent empirical recommendations by [7, 21]. Furthermore, Figure 3b(right) demonstrates that the FSL-optimal schedule yields a lower terminal loss than all baseline LRSs, substantiating its practical relevance.

Taken together, these results suggest that FSL offers a faithful analytical surrogate model for studying LLM training dynamics and holds promise as a principled tool for interpreting and designing LRSs in large-scale pre-training.

7 Concluding Remarks

In this paper, we present a systematic study of how LRS influences scaling laws by considering a teacher–student kernel regression setting. Our theoretical framework yields a novel functional scaling law, which explicitly characterizes the impact of both learning rate and batch size schedules via a convolution-type functional term. 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.

Looking ahead, several promising directions remain open. First, it is essential to develop a theoretical characterization of the optimal LRS under various resource constraints, which remains largely unresolved. Second, leveraging FSL studies the effect of batch size scheduling – a relatively under-explored topic despite its practical importance. Third, to make FSL truly practical, it is crucial to refine our FSL through extensive large-scale LLM pre-training experiments.

Acknowledgement

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.

[2] Stefan Ankirchner and Stefan Perko. A comparison of continuous-time approximations to stochastic gradient descent. *Journal of Machine Learning Research*, 25(13):1–55, 2024.

[3] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[4] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[5] 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.

[6] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade: Second edition*, pages 437–478. Springer, 2012.

[7] 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.

[8] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

[9] 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.

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

[11] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *arXiv preprint arXiv:2409.17858*, 2024.

[12] 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.

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

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

[15] Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm. 2005.

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

[17] 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.

[18] 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.

[19] 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.

[20] P Goyal. Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[21] Alex Hägele, Elie Bakouch, Atli Kosson, Leandro Von Werra, Martin Jaggi, et al. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

[27] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint arXiv:2402.04376*, 2024.

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

[29] 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.

[30] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[39] 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.

[40] 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.

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

[42] 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.

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

[44] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

[45] 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.

[46] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] 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.

[49] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

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

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

[52] 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.

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

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[55] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[56] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

[57] 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.

[58] 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.

[59] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

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

[61] 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.

[62] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

[63] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

Appendix

Table of Contents

A Additional Related Work

A.1 Empirical Findings on Scaling Laws

Scaling laws have emerged as a foundational principle in understanding the interplay between model size, dataset size, and performance in large-scale deep model pre-training. Initially proposed by [29], these laws have been further refined through a series of subsequent studies [22, 24, 70, 28, 1, 9, 57, 31, 60, 42].

Among these, [24] introduced the Chinchilla Law, which approximates the final-step risk $\mathcal{R}$ as follows:

$$\mathcal{R} = R_0 + A_1 D^{-\kappa_1} + A_2 M^{-\kappa_2},$$

where R_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. Based on this risk prediction, they formulate the following compute-optimal scaling problem:

$$\min_{DM=C} R_0 + A_1 D^{-\kappa_1} + A_2 M^{-\kappa_2},$$

where the compute budget $C = DM$ serves as the scaling bottleneck.

Furthermore, [60] and [42] empirically investigated scaling laws in the context of LRSs. [60] heuristically proposed the Momentum Law to characterize the impact of LRSs on scaling laws, where the population risk at the k-th step follows:

$$\mathcal{R}_k = R_0 + A S_1^{-\kappa} - C S_2,$$

where R_0, A, C, and κ are constant parameters, and S_1 and S_2 roughly denote the area under the learning rate curve and the learning rate annealing area, respectively, which are formally defined as follows:

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

where $\boldsymbol{\eta} := (\eta_1, \eta_2, \cdots, \eta_K)^\top \in \mathbb{R}_+^K$ denotes the learning rate schedule and λ is a hyper-parameter representing the decay factor for LR annealing momentum, which typically ranges from 0.99 to 0.999.

Subsequently, [42] further developed an empirical law, termed the Multi-Power Law, in which they replaced S_2 in the Momentum Law with additional power laws to account for the loss reduction effect as the learning rate decays. The Multi-Power Law says that the risk can be predicted as:

$$\mathcal{R}_k = R_0 + AS_1^{-\kappa} - \mathrm{LD}(k),$$

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

where $R_0, A, C, C', \kappa, \kappa', \kappa''$ are constants.

However, both the Momentum Law and the Multi-Power Law remain empirical, lacking a rigorous theoretical foundation. In our work, we build on a solvable model and theoretically derive a novel functional-form scaling law with arbitrary LRS, which also generalizes well for unseen larger model sizes.

A.2 SGD for Kernel Regression

Several works have analyzed the SGD algorithm within the context of high-dimensional kernel regression [16, 38, 53, 47, 8, 62]. [16] demonstrated that averaged-iteration SGD with a constant learning rate can achieve statistical optimality in the easy-task regime. [38] showed that multiple-pass, non-averaged SGD also attains statistical optimality in this regime, while [53] established that multiple-pass, averaged-iteration SGD can achieve optimality in the hard-task regime. Building on these results, [47] analyzed the convergence behavior of SGD in high-dimensional kernel regression, demonstrating that strategies such as averaging and multiple passes can enhance learning performance. [8] and [62] studied the noiseless case: [8] analyzed SGD for a noiseless linear model, deriving convergence results and extending the findings to online function estimation. [62] examined SGD in a noiseless least-squares setup, showing convergence for non-strongly convex problems and faster-than-usual convergence rates in over-parameterized settings.

B   Experiment Details and Additional Results

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

B.1   Popular Learning Rate Schedules

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

- **Cosine Schedule [41].** 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 $\eta_{\max}/10$ [61].

- **Warmup-Stable-Decay (WSD) Schedule [70, 25, 21].** 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 [9].** 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 $K_1 = 0.8K, K_2 = 0.9K$.

B.2 Teacher-Student Kernel Regression

Physical-time FSL. The FSL (7) 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 (7), 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 teacher-student model 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.

$$\mathcal{E}_k = c_1 T(k)^{-s} + c_2 \int_0^k \mathcal{K}(T(k) - T(u))e(T(u))\varphi(u)^2 \, \mathrm{d}u + c_3\sigma^2 \int_0^k \mathcal{K}(T(k) - T(u))\varphi(u)^2 \, \mathrm{d}u,$$

where c_1, c_2, c_3 are constants to fit.

Fitting the SGD trajectory, we minimize the mean squared error (MSE) between the empirical risk trajectory of SGD, denoted by $\mathcal{E}_{\mathrm{SGD}}(k)$, and the theoretical prediction from FSL, $\mathcal{E}_k$. Formally, we solve the following optimization problem:

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

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 4. Complementing Figure 1 (middle and right), additional experimental results for various values of s and β are presented in Figure 4.

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 (7), 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 risk 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
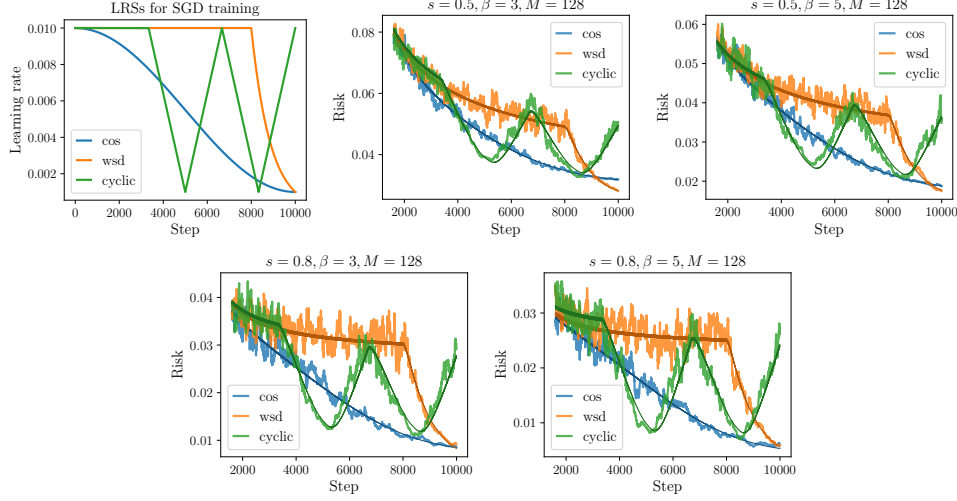
Figure 4: **Fitting results of FSL on SGD trajectories.** The shaded curves are the average over 200 independent SGD runs, while the solid curves show the predictions of FSL.

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 $\eta_{\max} = 0.05 D^{-r}$, 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.

B.3 LLM pre-training

Practical FSL Ansatz for LLM pre-training In this section, in order to fit the scaling laws on real LLM pre-training loss curves, we will derive an approximation form of the FSL in Theorem 4.1.

First, by the physical-time for of the FSL (7) 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

$$\int_0^k \mathcal{K}(T(k) - T(u))(\sigma^2 + e(T(k)) \cdot \varphi(u)\varphi(0) \, \mathrm{d}u = \varphi(0) \int_0^{T(k)} \mathcal{K}(T(k) - t)(\sigma^2 + e(t)) \, \mathrm{d}t. \quad (10)$$

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 D.1), we have

$$(10) \eqsim \varphi(0)(\sigma^2 + e(T(k))).$$

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

$$\mathcal{E}_k \eqsim \frac{1}{T(k)^s} + M^{-s\beta} - \mathrm{LRD}(k), \quad (11)$$

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 $r = 0, 1, \ldots, k$,

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

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

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

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

$$\int_{T(i)}^{T(k)} \mathcal{K}(T(k) - t) \, \mathrm{d}t = \int_{M^{-\beta}}^{1} u^{1-1/\beta} \int_{T(i)}^{T(k)} e^{-2u(T(k)-t)} \, \mathrm{d}t \, \mathrm{d}u = \frac{1}{2} - \frac{1}{2} g_{1-1/\beta}(T(k) - T(i)),$$

we then further simplify it as (see the definition of g function in Section C.3)

$$\text{LRD}(k) \approx \frac{1}{2B} \sum_{i=1}^{k} (\eta_{i-1} - \eta_i)(\sigma^2 + e(\xi_i))(1 - g_{1-1/\beta}(T(k) - T(i)))$$

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

Here, in the last line we approximate ξ_i as $T(i)$ and approximate $g_{1-1/\beta}(t)$ as $(1 + ct)^{-\gamma}$ with some constant $c > 0$ for simplicity.

Therefore, combining with (11), when the batch size B is fixed, after renaming some constants, the final discrete ansatz can be written as

$$\begin{aligned}
\mathcal{R}_k \approx c_0 + \frac{c_1}{T(k)^s} + c_2 M^{-s\beta} \\
- c_3 \sum_{i=1}^{k} (\eta_{i-1} - \eta_i) \left(c_4 + \frac{1}{T(i)^s} \right) \left(1 - (1 + c_5(T(k) - T(i)))^{-\gamma} \right),
\end{aligned} \tag{12}$$

where $c_0, c_1, c_2, c_3, c_4, c_5, s, \beta, \gamma$ 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 (12),

$$\mathcal{L}_\Theta(k) = L_0 + \frac{c_1}{T(k)^s} + c_2 M^{-s\beta} - \text{LRD}(k)$$

where

$$\text{LRD}(k) := c_3 \sum_{i=1}^{k} (\eta_{i-1} - \eta_i) \left(c_4 + \frac{1}{T(i)^s} \right) \left(1 - \frac{1}{(1 + c_5(T(k) - T(i)))^\gamma} \right),$$

and $\Theta = (L_0, c_1, c_2, c_3, c_4, c_5, s, \beta, \gamma)$.

Following [60], 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 3a.

FSL-optimal LRS via numerical variation. We propose to obtain a numerical optimal LRS by directly minimizing the final-step risk 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

$$\min_{\{\delta_i\}_{i=1}^K} \mathcal{L}_\Theta(\{\eta_i\}_{i=1}^K), \quad \text{subject to } \sum_{k=0}^{K-1} \delta_k \leqslant \eta_0, \, d\eta_i \geqslant 0, i = 0, 1, \ldots, K - 1. \tag{13}$$

To solve the above constraint optimization, we use the projected gradient descent (PGD) [13]. 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 3b (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 3b (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 3b.

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 [54], LLaMA [61] and QwenMoE [69]. 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 5 for the 1B LLaMA dense model, Figure 6 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 7 and 8, 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 configurations. As evidenced by Figures 9 and 10, 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.

C Proof of Theorem 4.1

C.1 Preliminaries

C.1.1 Connections to kernel regression

In this section we aim to explain the assumptions made in section 2.2, namely Assumption 2.2 and 2.3.

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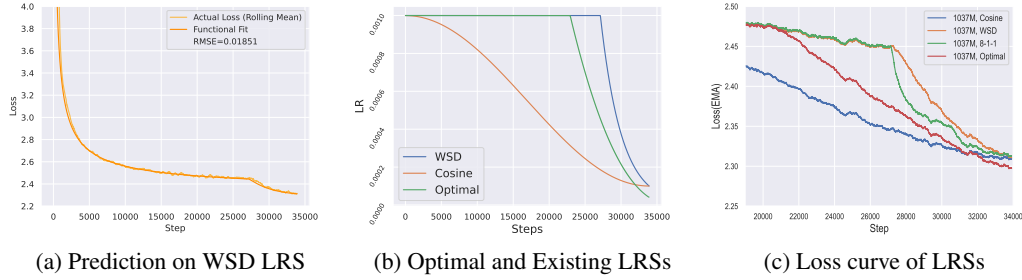
Figure 5: **Experiment on the 1B LLaMA (dense) model.** Figure (a): We fit our functional scaling law on the loss curve of 1B LLaMA (dense) model with 20B tokens training data and 8-1-1 LRS. Figures (b)(c): The comparison on the 1B model between the optimal LRS, cosine LRS, WSD LRS with exponential decay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

(a) Prediction on WSD LRS   (b) Optimal and Existing LRSs   (c) Loss curve of LRSs
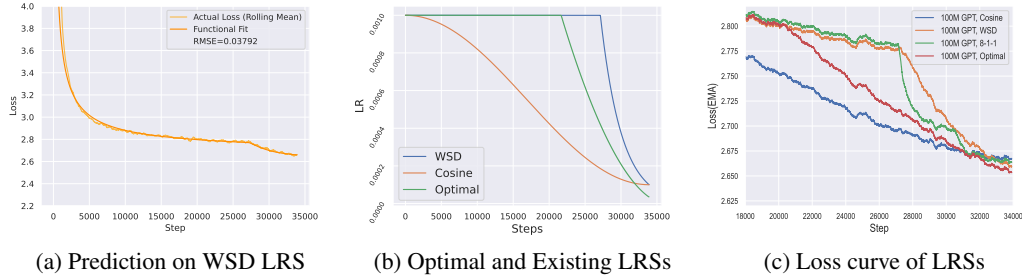
Figure 6: **Experiment on the 100M GPT2 (dense) model.** Figure (a): We fit our functional scaling law on the loss curve of 100M GPT2 (dense) model with 20B tokens training data and 8-1-1 LRS. Figures (b)(c): The comparison on the 100M model between the optimal LRS, cosine LRS, WSD LRS with exponential decay and 8-1-1 LRS.

(a) Prediction result   (b) Optimal LRS
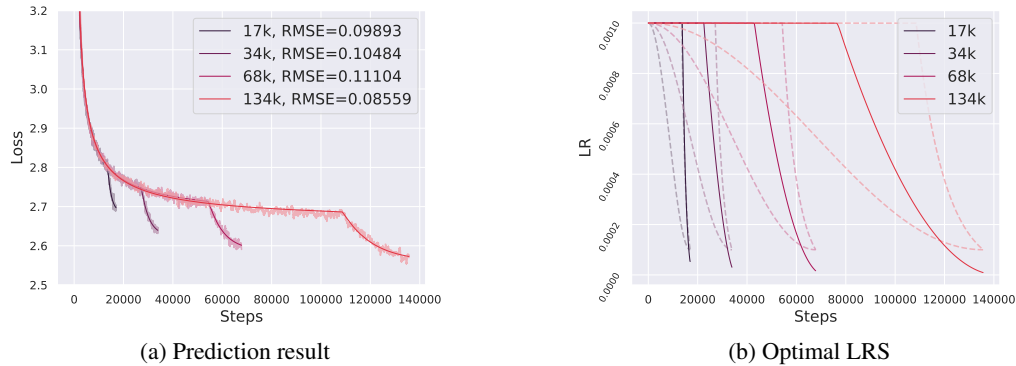
Figure 7: **Experiments with different total steps.** Figure (a): Fitted functional scaling laws on 100M LLaMA model with different total training steps 17k, 34k, 68k and 134k (corresponding to 10B, 20B, 40B and 80B tokens respectively). Figure (b): Optimal LRSs compared with cosine and WSD LRSs. The solid lines are optimal LRSs, and the dashed lines are cosine/WSD LRSs.

23

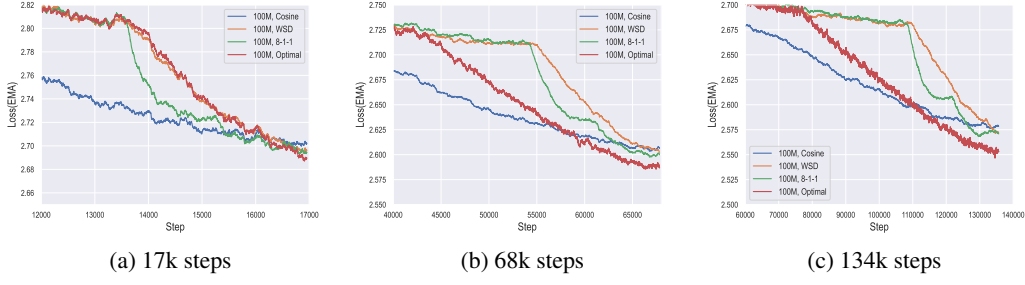| (a) 17k steps | (b) 68k steps | (c) 134k steps |

Figure 8: **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.
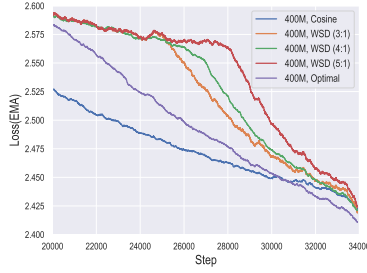

Figure 9: **WSD with different decay ratios:** We train a 400M LLaMA (dense) model with 20B tokens of training data and WSD LRSs with the ratios between stable time and decay time of 3:1, 4:1, and 5:1. All WSD LRSs exhibit a final loss similar to that of the Cosine LRS, and the optimal LRS derived from our functional scaling law outperforms all other LRSs by a loss gap of approximately 0.01.
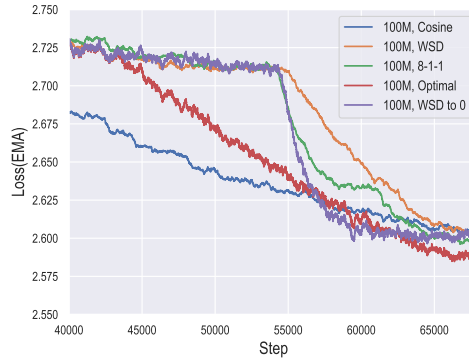
Figure 10: **Comparison between optimal LRS and WSD with a near-zero final learning rate:** We train a 100M LLaMA (dense) model with 40B tokens training data and various LRSs with the same $\eta_{\max} = 10^{-3}$, including WSD LRS with $\eta_{\min} = \frac{1}{10}\eta_{\max}$, WSD LRS with $\eta_{\min} = 10^{-7}$, 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.

24

Our setting is highly related to the classical kernel regression. In kernel regression, we aim to learn a function within a reproducing kernel Hilbert space (RKHS) from finite data samples. We begin by reviewing the foundational setup of kernel regression and subsequently show the connections between our assumptions and the classical ones.

Let $\mathcal{X}$ be a compact topological space. Let ρ be a probability measure on $\mathcal{X} \times \mathbb{R}$. We denote by ρ_x the marginal distribution on $\mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous PSD kernel such that $\mathbb{E}_{x \sim \rho_x}[K(x, x)] < \infty$. and we denote the corresponding RKHS by $\mathcal{H}_K$. For some $x \in \mathcal{X}$, define $K_x := K(x, \cdot) \in \mathcal{H}_K$.

For simplicity, we assume that there exists a function $f^* : \mathcal{X} \to \mathbb{R}$ such that the distribution ρ of (x, y) is given by

$$y = f^*(x) + \epsilon, \quad x \sim \rho_x,$$

where ϵ is a random variable independent of x and satisfies $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$. We call f^* the target function.

For any $f \in \mathcal{H}_K$, we define the excess risk as

$$\mathcal{E}(f) = \frac{1}{2} \|f - f^*\|_{L^2_{\rho_x}(\mathcal{X})}^2 = \frac{1}{2} \mathbb{E}_{x \sim \rho_x}[(f(x) - f^*(x))^2]$$

Our objective is to minimize the excess risk. The dataset is i.i.d. drawn from the distribution ρ.

Given a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, by Mercer's theorem there exists a nonnegative sequence $\{\lambda_j\}_{j=1}^\infty$ and an orthonormal basis $\{e_j\}_{j=1}^\infty$ of $L^2_\rho(\mathcal{X})$ such that

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

where the series converges absolutely and uniformly.

Define the feature functions $\phi_j(x) = \lambda_j^{\frac{1}{2}} e_j(x)$, then the RHKS $\mathcal{H}_K$ associated with K can be written as

$$K(x, x') = \sum_{j=1}^\infty \phi_j(x) \phi_j(x'), \quad \mathcal{H}_K = \left\{ \sum_{j=1}^\infty a_j \phi_j : \sum_{j=1}^\infty a_j^2 < \infty \right\},$$

with $\{\phi_j\}_{j=1}^\infty$ being an orthonormal basis of $\mathcal{H}_K$. Also we have $\mathbb{E}[\phi_j(x)^2] = \lambda_j \mathbb{E}[e_j(x)^2] = \lambda_j$, and $\{\phi_j\}_{j=1}^\infty$ is still orthogonal in $L^2_\rho(\mathcal{X})$.

The interpolation space $\mathcal{H}_K^s$ is

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

where $\mathcal{H}_K^0 = L^2_{\rho_x}$, $\mathcal{H}_K^1 = \mathcal{H}_K$. We have $\mathcal{H}_K^{s_1} \subset \mathcal{H}_K^{s_2}$ for any $s_1 > s_2$.

The source-capacity condition are usually given as

Assumption C.1 (Capacity condition). There exists some $\beta > 1$ such that $\lambda_j = O(j^{-\beta})$.

Assumption C.2 (Source condition). There exists some $s > 0$ such that $f^*(x) \in \mathcal{H}_K^s$.

In Assumption 2.2 and 2.3, the parameter θ_j corresponds to a_j here, therefore they are the power-law version of the standard source-capacity condition, which gives the following insights:

- A smaller s indicates that the target function f^* lies in a larger space, hence the learning problem is more difficult.

- A smaller β indicates that the RKHS $\mathcal{H}_K$ is larger, and in our setting $\mathcal{H}_K$ is the hypothesis space, meaning that the model capacity is larger.

C.1.2 The SDE modeling

In our setup, the SGD update can be written as

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

The term $\boldsymbol{\xi}_k$ is the gradient noise, whose covariance is $\frac{1}{B_k}\Sigma(\mathbf{v}_{k-1})$.

Therefore in order to approximate it as an Itô-type SDE, we assume the noise is Gaussian and write

$$\mathbf{v}_k - \mathbf{v}_{k-1} = -\varphi_k \nabla \mathcal{R}(\mathbf{v}_{k-1})h + \varphi_k \sqrt{\frac{h}{B_k}\Sigma(\mathbf{v}_{k-1})}\sqrt{h}\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}).$$

Hence the corresponding SDE (5) is

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

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

C.2 Analysis of the Intrinsic-time SDE

In this section, we will derive a Volterra integral equation of the excess risk from the intrinsic-time SDE. Recall the intrinsic-time SDE:

$$d\boldsymbol{\nu}_t = -\nabla \mathcal{R}(\boldsymbol{\nu}_t)\,dt + \sqrt{\gamma_t \Sigma(\boldsymbol{\nu}_t)}\,d\mathbf{B}_t, \tag{14}$$

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

By (2), we have $\nabla \mathcal{R}(\mathbf{v}) = \mathbf{W}\mathbf{H}(\mathbf{W}^\top \mathbf{v} - \theta^*)$. Let $\mathbf{u}_t = \mathbf{W}^\top \boldsymbol{\nu}_t - \theta^*$. 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 C.3. *We have*

$$d\mathbf{u}_t = -\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{u}_t\,dt + \sqrt{\gamma_t \mathbf{W}^\top \Sigma_t \mathbf{W}}\,d\mathbf{B}_t, \tag{15}$$

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

Proof. By Eq. (14),

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

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 $\mathbf{W}^\top \sqrt{\gamma_t \Sigma_t}\,d\tilde{\mathbf{B}}_t$ has the same distribution as $\sqrt{\gamma_t \mathbf{W}^\top \Sigma_t \mathbf{W}}\,d\mathbf{B}_t$, hence the SDE can be written in $\mathbf{B}_t$ as

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

$\square$

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 C.4 (Noise Structure). *For any $\mathbf{v} \in \mathbb{R}^M$, it holds that*

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

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 $\Sigma(\mathbf{v}) \approx \mathcal{R}(\mathbf{v})\nabla^2 \mathcal{R}(\mathbf{v})$. 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 [68, 67].

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

$$\nabla \ell(\mathbf{z}; \mathbf{v}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top \left(\mathbf{W}^\top \mathbf{v} - \boldsymbol{\theta}^*\right) - \mathbf{W}\boldsymbol{\phi}(\mathbf{x})\epsilon$$

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

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

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

Noting

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

then applying Assumption 2.1, we have

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

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. $\qquad\square$

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

Proposition C.5. *For the intrinsic-time SDE, we have*

$$2\mathbb{E}[\mathcal{E}_t] = \mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t \mathbf{u}_0 + \int_0^t \mathrm{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}) \cdot \gamma_\tau(c_\tau \mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau, \qquad (16)$$

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,

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

Integrating both sides, we get

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

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

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

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

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

By Lemma E.2 and Lemma C.4, we have

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

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

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

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

from which the lemma follows. $\qquad\square$

C.3 The *g* Functions

Our analysis will frequently use following family of functions

$$g_a(t) = g_a^{(M)}(t) := \int_{M^{-\beta}}^{1} u^{a-1} e^{-2ut} \, \mathrm{d}u$$

$$G_a(t) := \lim_{M \to \infty} g_a^{(M)}(t) = \int_0^1 u^{a-1} e^{-2ut} \, \mathrm{d}u \qquad (17)$$

Thus, the forgetting kernel and risk functions are special cases:

$$\mathcal{K}(t) = g_{2-1/\beta}(t), \quad e(t) = g_s(t). \qquad (18)$$

Lemma C.6. *It holds for any* $t > 0$ *that*

$$g_a(t) \leqslant \frac{1}{a}, \quad g_a(t) \leqslant \frac{\Gamma(a)}{(2t)^a}$$

Proof. The proof only involves calculating the integral. Notice that

$$g_a(t) \leqslant \int_0^1 u^{a-1} e^{-2ut} \, \mathrm{d}u \leqslant \int_0^1 u^{a-1} \, \mathrm{d}u = \frac{1}{a},$$

and

$$g_a(t) \leqslant \frac{1}{(2t)^a} \int_0^\infty u^{a-1} e^{-u} \, \mathrm{d}u = \frac{\Gamma(a)}{(2t)^a},$$

$\square$

Lemma C.7. *For* $a > 0$ *and* $t = \Omega(1)$, *if* $M^\beta > t$, *we have* $g_a(t) \eqsim \frac{1}{t^a}$. *If* $a > 1$, *we have* $\int_0^t g_a(\tau) \, \mathrm{d}\tau \eqsim 1$.

Proof. We have

$$g_a(t) = \int_{M^{-\beta}}^1 u^{a-1} e^{-2ut} \, \mathrm{d}u$$

$$= \frac{1}{(2t)^a} \int_{tM^{-\beta}}^t u^{a-1} e^{-u} \, \mathrm{d}u.$$

The integral part is a partial gamma function, hence when $tM^{-\beta} \lesssim 1$, we have $g_a(t) \eqsim \frac{1}{t^a}$.

When $a > 1$, we have

$$\int_0^t g_a(\tau) \, \mathrm{d}\tau = \int_{M^{-\beta}}^1 \int_0^t u^{a-1} e^{-2u\tau} \, \mathrm{d}\tau \, \mathrm{d}u$$

$$= \int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2ut}}{2u} \, \mathrm{d}u$$

$$= \frac{1 - M^{-(a-1)\beta}}{2(a-1)} - \frac{1}{2} g_{a-1}(t) \lesssim 1.$$

On the other hand, when $t = \Omega(1)$, $g_{a-1}(t) < \frac{1}{(a-1)}$, combining with $M \gg 1$ we have $\int_0^t g_a(\tau) \, \mathrm{d}\tau \eqsim 1$. $\square$

Lemma C.8. *For* $a, b \geqslant 0, t > 0$, *we have*

$$\int_1^{t-1} \frac{1}{(t-\tau)^a} \frac{1}{\tau^b} \, \mathrm{d}\tau \leqslant \frac{8}{t^{a \wedge b}},$$

where $a \wedge b := \min\{a, b\}$.

Proof. Without loss of generality, we can assume $a \geqslant b$. By Hölder's Inequality, we have

$$
\begin{aligned}
\int_1^{t-1} \frac{1}{(t-\tau)^a} \frac{1}{\tau^b} \, \mathrm{d}\tau &= \int_1^{t-1} \frac{1}{(\tau(t-\tau))^b} \frac{1}{(t-\tau)^{a-b}} \, \mathrm{d}\tau \\
&= \int_1^{t-1} \frac{1}{(\tau(t-\tau))^b} \frac{1}{(t-\tau)^{a-b}} \, \mathrm{d}\tau \\
&\leqslant \frac{2^{b-1}}{t^b} \int_1^{t-1} \left(\frac{1}{(t-\tau)^a} + \frac{1}{\tau^b(t-\tau)^{a-b}} \right) \mathrm{d}\tau \\
&\leqslant \frac{2^b}{t^b} \int_1^{t-1} \frac{1}{\tau^a} \, \mathrm{d}\tau \\
&\leqslant \frac{2^b a}{t^b} \leqslant \frac{8}{t^b}.
\end{aligned}
$$

$\square$

Lemma C.9. *Assume t, M is sufficiently large and $M^\beta > t$. For $a, b \geqslant 0, t > 0$, we have*

$$
\int_0^t g_a(t-\tau) g_b(\tau) \, \mathrm{d}\tau \eqsim g_{a \wedge b}(t),
$$

where $a \wedge b := \min\{a, b\}$.

Proof. Without loss of generality, we can assume $a \geqslant b$. Then, using Fubini's theorem gives

$$
\begin{aligned}
\int_0^t g_a(t-\tau) g_b(\tau) \, \mathrm{d}\tau &= \int_0^t \int_{M^{-\beta}}^1 \int_{M^{-\beta}}^1 u^{a-1} v^{b-1} e^{-2u(t-\tau)} e^{-2v\tau} \, \mathrm{d}u \, \mathrm{d}v \, \mathrm{d}\tau \\
&= \int_{M^{-\beta}}^1 \int_{M^{-\beta}}^1 u^{a-1} v^{b-1} \frac{e^{-2vt} - e^{-2ut}}{2(u-v)} \, \mathrm{d}u \, \mathrm{d}v \\
&= \int_{M^{-\beta}}^1 v^{b-1} e^{-2vt} \left(\int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u \right) \mathrm{d}v
\end{aligned}
$$

Lower bound: Let $\epsilon_1 = \frac{\log 2}{2t}$ and we have

$$
\int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u \geqslant \int_{M^{-\beta}}^{v-\epsilon} u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u + \int_{v+\epsilon}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u.
$$

Notice that when $|u - v| \geqslant \frac{\log 2}{2t}$

$$
\frac{1 - e^{-2(u-v)t}}{2(u-v)} \geqslant \frac{1}{4},
$$

and we have

$$
\int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u \geqslant \frac{1}{4} \left(\int_{M^{-\beta}}^{v-\epsilon} u^{a-1} \, \mathrm{d}u + \int_{v+\epsilon}^1 u^{a-1} \, \mathrm{d}u \right) = \frac{1}{4} \left(1 - M^{-a\beta} - \int_{v-\epsilon}^{v+\epsilon} u^{a-1} \, \mathrm{d}u \right).
$$

Notice that we have

$$
u^{a-1} \leqslant 1,
$$

and thus

$$
\int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, \mathrm{d}u \geqslant \frac{1}{4} \left(1 - M^{-a\beta} - \frac{\log 2}{t} \right).
$$

As $M, t = \Omega(1)$, we have

$$
M^{-a\beta} < \frac{1}{4},
$$

and

$$
\frac{\log 2}{t} < \frac{1}{4}.
$$

29

Combining them together, we have

$$\int_{M^{-\beta}}^1 u^{a-1} \frac{1 - e^{-2(u-v)t}}{2(u-v)} \, du \geqslant \frac{1}{8},$$

which implies that

$$\int_0^t g_a(t-\tau) g_b(\tau) \, d\tau \geqslant \frac{1}{8} g_M(a \wedge b, t).$$

Upper bound: By lemma C.6

$$\int_0^t g_a(t-\tau) g_b(\tau) \, d\tau \leqslant \frac{\Gamma(a)\Gamma(b)}{2^{a+b}} \int_1^{t-1} \frac{1}{(t-\tau)^a} \frac{1}{\tau^b} \, d\tau$$

$$+ \int_0^1 g_a(t-\tau) g_b(\tau) \, d\tau + \int_{t-1}^t g_a(t-\tau) g_b(\tau) \, d\tau$$

We observe that

$$\int_{t-1}^t g_a(t-\tau) g_b(\tau) \, d\tau \asymp \frac{1}{a} \int_{t-1}^t g_b(\tau) \, d\tau \asymp \frac{1}{a} \frac{1}{t^b}.$$

and for the same reason

$$\int_0^1 g_a(t-\tau) g_b(\tau) \, d\tau \asymp \frac{1}{b} \int_0^1 g_a(\tau) \, d\tau \asymp \frac{1}{b} \frac{1}{t^a} \leqslant \frac{1}{b} \frac{1}{t^b}.$$

On the other hand, by the Lemma C.8 we have

$$\int_1^{t-1} \frac{1}{(t-\tau)^a} \frac{1}{\tau^b} \, d\tau \leqslant \frac{8}{t^{a \wedge b}}.$$

Combine them together we have

$$\int_0^t g_a(t-\tau) g_b(\tau) \, d\tau \leqslant \left(\frac{1}{a} + \frac{1}{b} + 8 \right) \frac{1}{t^b} \leqslant \frac{10}{t^b}.$$

Observe that $M^{-\beta} t < 1$ and $t > t_0$, we have

$$g_b(t) = \frac{1}{t^b} \int_{M^{-\beta} t}^t u^{b-1} e^{-2u} \, du \geqslant \frac{1}{t^b} \int_1^{t_0} u^{b-1} e^{-2u} \, du = \frac{C_1}{t^b}$$

So we can give the upper bound

$$\int_0^t g_a(t-\tau) g_b(\tau) \, d\tau \leqslant 10 \frac{1}{t^b} \leqslant \frac{10}{C_1} g_{a \wedge b}(t).$$

By combining the lower and upper bounds, we complete the proof of the lemma. □

Recall the definition of $G_a(\cdot)$. Then, the following lemma follows from Lemma C.9 immediately by taking $M \to \infty$.

Lemma C.10. $\int_0^t G_a(t-\tau) G_b(\tau) \, d\tau \asymp G_{a \wedge b}(t)$.

C.4 The Case of Top-*M* Features

In this section, we prove the functional scaling law for the case of top-M features. Recall the choice of top-M features, where we choose $\mathbf{W} = [\mathbf{I}_M, \mathbf{0}_{M \times (N-M)}]$. In this case we can simplify the SDE and the expression of the $\mathcal{E}_t$. Then we will estimate each term of the Volterra equation to complete the proof.

Lemma C.11 (The intrinsic-time SDE of top-M case). *Let the projector* $\mathbf{W} = [\mathbf{I}_M, \mathbf{0}_{M \times (N-M)}]$. *Then, we have*

$$d\mathbf{u}_t = -\mathbf{H}_{0:M} \mathbf{u}_t \, dt + \sqrt{\gamma_t (c_t \mathcal{E}_t + \sigma^2)} \sqrt{\mathbf{H}_{0:M}} \, d\mathbf{B}_t. \tag{19}$$

where $\mathbf{H}_{0:M} = \text{diag}\{\underbrace{\lambda_1, ..., \lambda_M}_{M}, \underbrace{0, ..., 0}_{N-M}\}$.

Proof. Recall that from Eq. (15),

$$d\mathbf{u}_t = -\mathbf{W}^\top \mathbf{W} \mathbf{H} \mathbf{u}_t \, dt + \sqrt{\gamma_t (c_t \mathcal{E}_t + \sigma^2)} \sqrt{\mathbf{W}^\top \mathbf{W} \mathbf{H} \mathbf{W}^\top \mathbf{W}} \, d\mathbf{B}_t$$

Notice that

$$\mathbf{W}^\top \mathbf{W} = \mathrm{diag}\{\underbrace{1, ..., 1}_{M}, \underbrace{0, ..., 0}_{N-M}\},$$

and we have

$$\mathbf{W}^\top \mathbf{W} \mathbf{H} = \mathbf{H}_{0:M}, \qquad \mathbf{W}^\top \mathbf{W} \mathbf{H} \mathbf{W}^\top \mathbf{W} = \mathbf{H}_{0:M}.$$

Combine these results and we have

$$d\mathbf{u}_t = -\mathbf{H}_{0:M} \mathbf{u}_t \, dt + \sqrt{\gamma_t (c_t \mathcal{E}_t + \sigma^2)} \sqrt{\mathbf{H}_{0:M}} \, d\mathbf{B}_t.$$

$\square$

We have the following recursion for $\mathcal{E}_t$.

Lemma C.12 (Volterra equation of the top-M case). *We have*

$$\mathbb{E}[\mathcal{E}_t] \asymp g_s(t) + \int_0^t \mathcal{K}(t - \tau)\gamma_\tau \left(c_\tau \mathbb{E}[\mathcal{E}_\tau] + \sigma^2 \right) d\tau + M^{-s\beta}.$$

Proof. The proof is mainly based on the Lemma C.11. Similarly, we decompose the SDE into its individual components, solve the resulting equations. By addressing each component individually, we can derive precise solutions and expectations for $(\mathbf{u}_t^{(j)})^2$. For $1 \leqslant j \leqslant M$, we have

$$\mathbb{E}[(\mathbf{u}_t^{(j)})^2] = e^{-2\lambda_j t}(\mathbf{v}_0^{(j)})^2 + \lambda_j \int_0^t \gamma_\tau e^{-2\lambda_j(t-\tau)}(c_\tau \mathbb{E}[\mathcal{E}_\tau] + \sigma^2) \, d\tau,$$

We then calculate the risk:

$$2\mathbb{E}[\mathcal{E}_t] = \sum_{j=1}^M \lambda_j e^{-2\lambda_j t}(\mathbf{v}_0^{(j)})^2 + \sum_{j=1}^M \lambda_j^2 \int_0^t \gamma_\tau e^{-2\lambda_j(t-\tau)}(c_\tau \mathbb{E}[\mathcal{E}(\mathbf{v}_\tau)] + \sigma^2) \, d\tau + \sum_{j=M+1}^\infty \lambda_j (\mathbf{v}_0^{(j)})^2$$

$$= \underbrace{\sum_{j=1}^M \lambda_j e^{-2\lambda_j t}(\mathbf{v}_0^{(j)})^2}_{I_1} + \underbrace{\sum_{j=1}^M \lambda_j^2 \int_0^t \gamma_\tau e^{-2\lambda_j(t-\tau)}(c_\tau \mathbb{E}[\mathcal{E}(\mathbf{v}_\tau)]) \, d\tau}_{I_2} + \underbrace{\sum_{j=1}^M \lambda_j^2 \int_0^t \gamma_\tau e^{-2\lambda_j(t-\tau)}(\sigma^2) \, d\tau}_{I_3}$$

$$+ \underbrace{\sum_{j=M+1}^\infty \lambda_j (\mathbf{v}_0^{(j)})^2}_{I_4}.$$

We next shall bound the four terms separately.

- **The I_1 term.** We can substitute λ_j by $j^{-\beta}$ and approximate the sum by integral

$$I_1 \asymp \sum_{j=1}^M j^{-s\beta-1} e^{-2\lambda_j t} \asymp \int_1^M x^{-s\beta-1} e^{-2x^{-\beta} t} \, dx = \frac{1}{\beta} \int_{M^{-\beta}}^1 u^{s-1} e^{-2ut} \, du \asymp g_s(t).$$

- **The I_2 term.** Similarly, we substitute λ_j by $j^{-\beta}$ and approximate the sum by integral

$$I_2 = \int_0^t \gamma_\tau c_\tau \mathbb{E}[\mathcal{E}_\tau] \left(\sum_{j=1}^M \lambda_j^2 e^{-2\lambda_j(t-\tau)} \, d\tau \right) \asymp \int_0^t \gamma_\tau c_\tau \mathbb{E}[\mathcal{E}_\tau] \left(\int_1^M x^{2\beta}(e^{-2x^{-\beta}(t-\tau)}) dx \right)$$

$$= \frac{1}{\beta} \int_0^t \gamma_\tau c_\tau \mathbb{E}[\mathcal{E}_\tau] \left(\int_{M^{-\beta}}^1 u^{1-1/\beta} e^{-2u(t-\tau)} du \right) \asymp \int_0^t \mathcal{K}(t - \tau)\gamma_\tau c_\tau \mathbb{E}[\mathcal{E}_\tau] \, d\tau. \quad (20)$$

31

- **The I_3 term.** Similarly, we substitude λ_j by $j^{-\beta}$ and approximate the sum by integral

$$I_3 = \sigma^2 \int_0^t \gamma_\tau \left(\sum_{j=1}^M \lambda_j^2 e^{-2\lambda_j(t-\tau)} \, \mathrm{d}\tau \right) \approx \sigma^2 \int_0^t \gamma_\tau \left(\int_1^M x^{2\beta}(e^{-2x^{-\beta}(t-\tau)}) dx \right)$$

$$= \frac{1}{\beta}\sigma^2 \int_0^t \gamma_\tau \left(\int_{M^{-\beta}}^1 u^{1-1/\beta} e^{-2u(t-\tau)} du \right) \approx \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \sigma^2 \, \mathrm{d}\tau.$$

- **The I_4 term.** By approximating the sum with the integral, we have

$$I_4 = \sum_{j=M+1}^{\infty} j^{-1-s\beta} \approx \int_M^{\infty} x^{-1-s\beta} \, \mathrm{d}x = M^{-s\beta}.$$

Combining them together completes the proof. $\qquad\square$

Proposition C.13 (Functional Scaling Law of the top-M case). *Suppose that $\gamma_{\max} := \sup_t \gamma_t$ is sufficiently small. Then, for $s \in (0, 2 - 1/\beta]$ we have*

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

Proof. **The case of $M^\beta \geqslant t$:** The condition of Lemma C.9 holds.

From Lemma C.12 we have

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

First we define $B(t)$ as

$$B(t) = \int_0^t \mathcal{K}(t-\tau)\, c_\tau \mathbb{E}[\mathcal{E}_\tau]\gamma_\tau \, \mathrm{d}\tau$$

$$\eqsim \int_0^t \mathcal{K}(t-\tau)\, c_\tau\gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r \, \mathrm{d}r \right) \mathrm{d}\tau$$

$$+ \int_0^t \mathcal{K}(t-\tau)\, c_\tau\gamma_\tau \int_0^\tau \mathcal{K}(\tau-r)c_r\mathbb{E}[\mathcal{E}_r]\gamma_r \, \mathrm{d}r \, \mathrm{d}\tau. \tag{21}$$

Then we calculate the last term. By Fubini's theorem,

$$\int_0^t \mathcal{K}(t-\tau)\, c_\tau\gamma_\tau \int_0^\tau \mathcal{K}(\tau-r)c_r\mathbb{E}[\mathcal{E}_r]\gamma_r \, \mathrm{d}r \, \mathrm{d}\tau$$

$$= \int_0^t c_r\mathbb{E}[\mathcal{E}_r]\gamma_r \int_r^t \mathcal{K}(t-\tau)\mathcal{K}(\tau-r)\gamma_\tau \, \mathrm{d}\tau \, \mathrm{d}r$$

$$\leqslant \int_0^t c_r\mathbb{E}[\mathcal{E}_r]\gamma_r \int_r^t \mathcal{K}(t-\tau)\mathcal{K}(\tau-r)\gamma_{\max} \, \mathrm{d}\tau \, \mathrm{d}r$$

$$\leqslant C\gamma_{\max} \int_0^t c_r\mathbb{E}[\mathcal{E}_r]\gamma_r\mathcal{K}(t-r) \, \mathrm{d}r$$

$$= C\gamma_{\max} B(t),$$

where the last inequality follows from Lemma C.9. Therefore, we get

$$B(t) \leqslant \frac{1}{1 - C\gamma_{\max}} \int_0^t \mathcal{K}(t-\tau)\, c_\tau\gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r \, \mathrm{d}r \right) \mathrm{d}\tau$$

$$B(t) \geqslant \int_0^t \mathcal{K}(t-\tau)\, c_\tau\gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r \, \mathrm{d}r \right) \mathrm{d}\tau.$$

As $C\gamma_{\max} \leqslant \frac{1}{2}$, we have

$$B(t) \approx \int_0^t \mathcal{K}(t-\tau)\, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r\, dr \right) d\tau.$$

Let

$$p(t) := \int_0^t g_{2-1/\beta}(\tau)\gamma_\tau\, d\tau, \quad q(t) := \int_0^t g_{2-1/\beta}(t-\tau)g_s(\tau)\gamma_\tau\, d\tau. \tag{22}$$

Noting $\mathcal{K}(t) = g_{2-1/\beta}(t)$, we have

$$\mathbb{E}[\mathcal{E}_t] \approx g_s(t) + M^{-s\beta} + \sigma^2 \int_0^t \mathcal{K}(t-\tau)\gamma_\tau\, d\tau$$

$$+ \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)\gamma_r\, dr \right) d\tau$$

$$= g_s(t) + M^{-s\beta} + (\sigma^2 + M^{-s\beta})p(t) + q(t) + \sigma^2 \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \int_0^\tau \mathcal{K}(\tau-r)\gamma_r\, dr\, d\tau.$$

The last term can be bounded as follows

$$\sigma^2 \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \int_0^\tau \mathcal{K}(\tau-r)\gamma_r\, dr\, d\tau$$

$$= \sigma^2 \int_0^t \gamma_r \int_r^t \mathcal{K}(t-\tau)\mathcal{K}(\tau-r)\gamma_\tau\, d\tau\, dr$$

$$\leqslant \sigma^2 \int_0^t \gamma_r \int_r^t \mathcal{K}(t-\tau)\mathcal{K}(\tau-r)\gamma_{\max}\, d\tau\, dr$$

$$\overset{(i)}{\leqslant} C\gamma_{\max}\sigma^2 \int_0^t \gamma_r \mathcal{K}(t-r)\, dr = C\gamma_{\max}\sigma^2 p(t),$$

where (i) uses Lemma C.9. Suppose $C\gamma_{\max} < \frac{1}{2}$, we have

$$\mathbb{E}[\mathcal{E}_t] \approx g_s(t) + M^{-s\beta} + (\sigma^2 + M^{-s\beta})p(t) + q(t).$$

Suppose $\gamma_{\max}$ to be sufficiently small such that

$$p(t) \leqslant \gamma_{\max} \int_0^t g_{2-1/\beta}(\tau)\, d\tau \leqslant \gamma_{\max} \frac{\beta}{2(\beta-1)} < \frac{1}{2}.$$

Then, we have

$$\mathbb{E}[\mathcal{E}_t] \approx g_s(t) + M^{-s\beta} + \sigma^2 p(t) + q(t)$$

$$= g_s(t) + M^{-s\beta} + \int_0^t \mathcal{K}(t-\tau)(\sigma^2 + g_s(\tau))\gamma_\tau\, d\tau.$$

The case of $M^\beta \leqslant t$: The condition of Lemma C.9 does not hold. In this case, we first analyze the left-hand side of the equation. The key idea is to make a small modification to $B(t)$ and use similar methods to prove the result.

By (21),

$$B(t) = \int_0^t \mathcal{K}(t-\tau)\, c_\tau \mathbb{E}[\mathcal{E}_\tau]\gamma_\tau\, d\tau \leqslant \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \mathbb{E}[\mathcal{E}_\tau]\gamma_\tau\, d\tau =: \tilde{B}(t).$$

We next proceed to estimate $\tilde{B}(t)$ in an analogous manner.

$$\tilde{B}(t) = \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \mathbb{E}[\mathcal{E}_\tau]\gamma_\tau\, d\tau$$

$$= \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r\, dr \right) d\tau$$

$$+ \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \int_0^\tau \mathcal{K}(\tau-r) c_r \mathbb{E}[\mathcal{E}_r]\gamma_r \, dr \, d\tau.$$

Then we calculate the last term. By Fubini's theorem,

$$\int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \int_0^\tau \mathcal{K}(\tau-r) c_r \mathbb{E}[\mathcal{E}_r]\gamma_r \, dr \, d\tau$$

$$= \int_0^t c_r \mathbb{E}[\mathcal{E}_r]\gamma_r \int_r^t G_{2-1/\beta}(t-\tau)\mathcal{K}(\tau-r)\gamma_\tau \, d\tau \, dr$$

$$\leqslant \int_0^t c_r \mathbb{E}[\mathcal{E}_r]\gamma_r \int_r^t G_{2-1/\beta}(t-\tau)G_{2-1/\beta}(\tau-r)\gamma_{\max} \, d\tau \, dr$$

$$\leqslant C\gamma_{\max} \int_0^t c_r \mathbb{E}[\mathcal{E}_r]\gamma_r G_{2-1/\beta}(t-r) \, dr$$

$$= C\gamma_{\max}\tilde{B}(t),$$

where the last inequality is by Lemma C.10. Therefore, we get

$$\tilde{B}(t) \leqslant \frac{1}{1-C\gamma_{\max}} \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r) c_r \gamma_r \, dr \right) d\tau.$$

$$\tilde{B}(t) \geqslant \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r) c_r \gamma_r \, dr \right) d\tau.$$

As $C\gamma_{\max} \leqslant \frac{1}{2}$, we have

$$\tilde{B}(t) \asymp \int_0^t G_{2-1/\beta}(t-\tau)\, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r) c_r \gamma_r \, dr \right) d\tau.$$

Next we bound the last term

$$\int_0^t G_{2-1/\beta}(t-\tau)\gamma_\tau \int_0^\tau \mathcal{K}(\tau-r)\gamma_r \, dr \, d\tau$$

$$= \int_0^t \gamma_r \int_r^t G_{2-1/\beta}(t-\tau)\mathcal{K}(\tau-r)\gamma_\tau \, d\tau \, dr$$

$$\leqslant \int_0^t \gamma_r \int_r^t G_{2-1/\beta}(t-\tau)G_{2-1/\beta}(\tau-r)\gamma_{\max} \, d\tau \, dr$$

$$\asymp \gamma_{\max} \int_0^t G_{2-1/\beta}(t-r)\gamma_r \, dr.$$

where the last inequality is again by Lemma C.10.

Also we have

$$\int_0^t \mathcal{K}(\tau) \, d\tau = \int_0^t \int_{M^{-\beta}}^1 u^{1-1/\beta} e^{-2u\tau} \, du \, d\tau$$

$$= \int_{M^{-\beta}}^1 u^{1-1/\beta} \frac{1-e^{-2ut}}{2u} \, du$$

$$\leqslant \int_{M^{-\beta}}^1 u^{1-1/\beta} \frac{1}{2u} \asymp 1,$$

Since we assumed that t is sufficiently large, we also have $\int_0^t \mathcal{K}(\tau) \, d\tau \asymp 1$, and similarly

$$\int_0^t (G_{2-1/\beta}(\tau) - \mathcal{K}(\tau))\gamma_\tau \, d\tau \lesssim \gamma_{\max} \int_0^{M^{-\beta}} u^{-1/\beta} \, du \asymp \gamma_{\max} M^{1-\beta},$$

which implies $\int_0^t G_{2-1/\beta}(\tau)\gamma_\tau \, d\tau \asymp \int_0^t \mathcal{K}(\tau)\gamma_\tau \, d\tau$.

Therefore, continuing the computation of $\tilde{B}(t)$, since

$$M^\beta < t \implies M^{-2\beta+1} \geqslant t^{-(2-\frac{1}{\beta})} \gtrsim G_{2-1/\beta}(t) \geqslant \mathcal{K}(t),$$

we have $M^{-\beta} + G_{2-1/\beta}(t) \eqsim M^{-\beta} + \mathcal{K}(t)$, hence similar to the proof in the first case, let

$$p(t) := \int_0^t G_{2-1/\beta}(\tau)\gamma_\tau \, \mathrm{d}\tau \lesssim \gamma_{\max}, \quad q(t) := \int_0^t G_{2-1/\beta}(t-\tau)g_s(\tau)\gamma_\tau \, \mathrm{d}\tau.$$

Since $\int_0^t g_s(\tau) \, \mathrm{d}\tau \lesssim \max\{1, M^{-s\beta+\beta}\}$, by $s \leqslant 2 - 1/\beta$, we have $\min(M^{-s\beta}, M^{-\beta}) \geqslant M^{-2\beta+1}$, we have

$$q(t) + M^{-s\beta} \eqsim \int_0^t (G_{2-1/\beta}(t-\tau) + M^{-2\beta+1})g_s(\tau)\gamma_\tau \, \mathrm{d}\tau + M^{-s\beta}$$

$$\eqsim \int_0^t (g_{2-1/\beta}(t-\tau) + M^{-2\beta+1})g_s(\tau)\gamma_\tau \, \mathrm{d}\tau + M^{-s\beta}$$

$$\eqsim \int_0^t \mathcal{K}(t-\tau)g_s(\tau)\gamma_\tau \, \mathrm{d}\tau + M^{-s\beta}.$$

Therefore by $p(t) \eqsim \int_0^t \mathcal{K}(\tau) \, \mathrm{d}\tau \lesssim \gamma_{\max}$, we get

$$\tilde{B}(t) \eqsim \int_0^t G_{2-1/\beta}(t-\tau) \, c_\tau \gamma_\tau \left(g_s(\tau) + M^{-s\beta} + \sigma^2 \int_0^\tau \mathcal{K}(\tau-r)c_r\gamma_r \, \mathrm{d}r \right) \mathrm{d}\tau$$

$$\eqsim q(t) + M^{-s\beta}p(t) + \gamma_{\max}\sigma^2 p(t)$$

$$\lesssim q(t) + \gamma_{\max}M^{-s\beta} + \gamma_{\max}\sigma^2 \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \, \mathrm{d}\tau.$$

Substituting this into the expression of $\mathcal{E}_t$ (Lemma C.12), note that $\mathcal{K}(t) = g_{2-1/\beta}(t)$, we have

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

$$\lesssim g_s(t) + q(t) + M^{-s\beta} + \sigma^2 \int_0^t \mathcal{K}(t-\tau)\gamma_\tau \, \mathrm{d}\tau$$

$$\lesssim M^{-s\beta} + g_s(t) + \int_0^t \mathcal{K}(t-\tau)(\sigma^2 + g_s(t))\gamma_\tau \, \mathrm{d}\tau.$$

On the other hand, clearly $\mathbb{E}[\mathcal{E}_t] \gtrsim g_s(t)$, thus by Lemma C.12,

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

$$\gtrsim M^{-s\beta} + g_s(t) + \int_0^t \mathcal{K}(t-\tau)(\sigma^2 + g_s(t))\gamma_\tau \, \mathrm{d}\tau,$$

which completes the proof. $\qquad\square$

Corollary C.14. *Suppose that* $\gamma_{\max} := \sup_t \gamma_t$ *is sufficiently small and* $M = \infty$. *Then, for* $s \in (0, +\infty)$ *we have*

$$\mathbb{E}[\mathcal{E}_t] \eqsim G_s(t) + \int_0^t \mathcal{K}(t-\tau)\left(\sigma^2 + G_s(\tau)\right)\gamma_\tau \, \mathrm{d}\tau.$$

Proof. By the case $M^\beta > t$ in the proof of Proposition C.13, the result follows. Note that in the case $M^\beta > t$, we don't need the condition of $s \leqslant 2 - 1/\beta$. $\qquad\square$

C.5 The Case of Random Features

In this section, we prove Theorem 4.1 for the case of random features. The proof proceed similarly as the top-M case. First we introduce some concentration properties of the random martix $\mathbf{W}\mathbf{H}\mathbf{W}^\top$ from [39], then we estimate the upper and lower bound of each term of the Volterra equation in Proposition C.5.

C.5.1 Concentration Inequalities

Recall that we derived the following recursive equation in Eq. (16):

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

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 $b = N$),

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

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 [39].

Lemma C.15 (Lemma G.4 in [39]). *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 C.16 (Lemma G.5 in [39]). *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 $1 - e^{-\Omega(M)}$.

C.5.2 Upper and Lower Bounds

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

Lemma C.17. *With probability at least $1 - e^{-\Omega(M)}$, we have*

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

Proof. We can compute that

$$\text{tr}(\mathbf{S}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau}\mathbf{S}) = \text{tr}(\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{W}^\top \mathbf{W}\mathbf{A}_{t-\tau}^\top \mathbf{H}\mathbf{A}_{t-\tau})$$

$$= \text{tr}\left(\mathbf{W}^\top \mathbf{W}\mathbf{H}\mathbf{W}^\top \mathbf{W}\sum_{a,b=0}^\infty \frac{1}{a!b!}(-(t-\tau))^{a+b}(\mathbf{H}\mathbf{W}^\top \mathbf{W})^a\mathbf{H}(\mathbf{W}^\top \mathbf{W}\mathbf{H})^b\right)$$

$$= \text{tr}\left(\sum_{a,b=0}^\infty \frac{1}{a!b!}(-t+\tau)^{a+b} \cdot \mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b+1}\mathbf{W}\mathbf{H}\right)$$

$$= \text{tr}\left(\sum_{a,b=0}^\infty \frac{1}{a!b!}(-t+\tau)^{a+b} \cdot (\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{a+b+2}\right)$$

$$= \sum_{a,b=0}^\infty \frac{1}{a!b!}(-t+\tau)^{a+b}\sum_{j=1}^M \hat{\lambda}_j^{a+b+2}$$

$$= \sum_{j=1}^M e^{-2(t-\tau)\hat{\lambda}_j}\hat{\lambda}_j^2.$$

By Lemma C.15, $\hat{\lambda}_j \asymp \lambda_j \asymp j^{-\beta}$, hence the summation can be estimated by the integral

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

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

For the first term in Eq. (16), following [39], we have

Lemma C.18. *With probability at least $1 - e^{-\Omega(M)}$, we have*

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

Proof. Note

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

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

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

Then by Lemma C.21, we can derive an upper bound. Since $s < 1$,

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

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 $g_s(t) \asymp \frac{1}{t^s}$ by Lemma C.7, the conclusion follows. $\qquad\square$

Lemma C.19. *It holds with probability at least $1 - e^{-\Omega(M)}$ that*

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

Proof. Following the proof of Lemma C.18, 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 2.2 and 2.3,

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

$$\eqsim \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i} \hat{\lambda}_i^{-1} \mu_i (\mathbf{W} \mathbf{H}^{1+s+1/\beta} \mathbf{W}^\top)$$

$$\gtrsim \sum_{i=1}^{M} e^{-2t\hat{\lambda}_i} \hat{\lambda}_i^{s+1/\beta}$$

$$\eqsim \int_1^M e^{-2tx^{-\beta}} x^{-1-s\beta} \, \mathrm{d}x$$

$$= \int_{M^{-\beta}}^1 e^{-2tu} u^{s-1} \, \mathrm{d}u = g_s(t).$$

Here we used Lemma C.15.

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

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

which will be proved in the Lemma C.20

Notice that

$$\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{W}^\top (\mathbf{W} \mathbf{H} \mathbf{W}^\top)^{-1} \mathbf{W} \mathbf{H}^{\frac{1}{2}} = \begin{pmatrix} \mathbf{I}_k - \mathbf{H}_{0:k}^{1/2} \mathbf{W}_{0:k}^\top \mathbf{A}^{-1} \mathbf{W}_{0:k} \mathbf{H}_{0:k}^{1/2} - & -\mathbf{H}_{0:k}^{1/2} \mathbf{W}_{0:k}^\top \mathbf{A}^{-1} \mathbf{W}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \\ -\mathbf{H}_{k:\infty}^{1/2} \mathbf{W}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{W}_{0:k} \mathbf{H}_{0:k}^{1/2} & \mathbf{I}_{N-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{W}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{W}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \end{pmatrix}$$

$$=: \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Y}^\top & \mathbf{Z} \end{pmatrix},$$

where $\mathbf{A} = \mathbf{W} \mathbf{H} \mathbf{W}^\top$.

We have

$$\mathbb{E}_{\mathbf{u}_0} \| (\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{W}^\top (\mathbf{W} \mathbf{H} \mathbf{W}^\top)^{-1} \mathbf{W} \mathbf{H}^{\frac{1}{2}}) \mathbf{H}^{\frac{1}{2}} \mathbf{u}_0 \|^2$$

$$= \mathbb{E}_{\mathbf{u}_0} \left\langle \mathbf{X}^2 + \mathbf{Y} \mathbf{Y}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{u}_{0:k} \mathbf{u}_{0:k}^\top \mathbf{H}_{0:k}^{1/2} \right\rangle + \mathbb{E}_{\mathbf{u}_0} \left\langle \mathbf{Z}^2 + \mathbf{Y}^\top \mathbf{Y}, \mathbf{H}_{0:k}^{1/2} \mathbf{u}_{k:\infty} \mathbf{u}_{k:\infty}^\top \mathbf{H}_{k:\infty}^{1/2} \right\rangle$$

$$+ 2 \mathbb{E}_{\mathbf{u}_0} \left\langle \mathbf{X} \mathbf{Y} + \mathbf{Y} \mathbf{Z}, \mathbf{H}_{0:k}^{1/2} \mathbf{u}_{0:k} \mathbf{u}_{0:k}^\top \mathbf{H}_{k:\infty}^{1/2} \right\rangle$$

$$= \operatorname{Tr}((\mathbf{Z}^2 + \mathbf{Y} \mathbf{Y}^\top) \mathbf{H}_{0:k}) + \operatorname{Tr}((\mathbf{Z}^2 + \mathbf{Y}^\top \mathbf{Y}) \mathbf{H}_{k:\infty}),$$

We can calculate that

$$\mathbf{Z}^2 + \mathbf{Y}^\top \mathbf{Y} = \mathbf{Z}$$

and
$$\mathbf{X}^2 + \mathbf{Y}\mathbf{Y}^\top = \mathbf{H}_{0:k}^{-1/2} \left[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{W}_{0:k}\right]^{-1} \mathbf{H}_{0:k}^{-1/2},$$
where $\mathbf{A}_k = \mathbf{W}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{W}_{k:\infty}^\top$.

And then we have
$$\mathrm{Tr}((\mathbf{X}^2 + \mathbf{Y}\mathbf{Y}^\top)\mathbf{H}_{0:k}) + \mathrm{Tr}((\mathbf{Z}^2 + \mathbf{Y}^\top\mathbf{Y})\mathbf{H}_{k:\infty})$$
$$= \mathrm{Tr}\left(\mathbf{H}_{0:k}^{-1/2} \left[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{W}_{0:k}\right]^{-1} \mathbf{H}_{0:k}^{-1/2}\mathbf{H}_{0:k}\right) + \mathrm{Tr}(\mathbf{Z}\mathbf{H}_{k:\infty})$$
$$= \mathrm{Tr}\left(\left[\mathbf{H}_{0:k}^{-1} + \mathbf{W}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{W}_{0:k}\right]^{-1}\right) + \mathrm{Tr}(\mathbf{Z}\mathbf{H}_{k:\infty})$$
$$\geqslant \mathrm{Tr}(\mathbf{Z}\mathbf{H}_{k:\infty}) := T_3$$

We can then simplify T_3
$$T_3 = \mathrm{Tr}\left(\mathbf{H}_{k:\infty}^{1/2}\left[\mathbf{I}_{M-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\right]\right)$$
$$\geqslant \mathrm{Tr}\left(\mathbf{H}_{k:\infty}^{1/2}\left[\mathbf{I}_{M-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\right]\mathbf{H}_{k:\infty}^{1/2}\right)$$

where the second inequality follows from $A \succeq A_k$.

$$T_3 \geqslant \sum_{i=1}^{N-k} \mu_i\left(\mathbf{I}_{N-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}\right) \cdot \mu_{N+1-k-i}(\mathbf{H}_{k:\infty})$$
$$\geqslant \sum_{i=1}^{N-k} \mu_i(\mathbf{M}) \cdot \mu_{N+1-k-i}(\mathbf{H}_{k:\infty}) \geqslant \sum_{i=k+M}^{N} \lambda_i,$$

where the second line follows from $A \succeq A_k$ and the third line follows from Von-Neuman's Inequality. Since $\mathbf{M} = \mathbf{I}_{N-k} - \mathbf{H}_{k:\infty}^{1/2}\mathbf{W}_{k:\infty}^\top \mathbf{A}_k^{-1}\mathbf{W}_{k:\infty}\mathbf{H}_{k:\infty}^{1/2}$ is a projection matrix such that $\mathbf{M}^2 = \mathbf{M}$ and $\mathrm{Tr}(\mathbf{I}_{N-k} - \mathbf{M}) = N - k$, we have $\mathbf{M}$ has M eigenvalues 0 and $N - k - M$ eigenvalues 1.

Let k be zero and substitute $\lambda_i \approx i^{-\beta}$ into the expression above, we have
$$\sum_{i=M}^{N} \lambda_i \geqslant cM^{-s\beta}.$$

$\square$

Lemma C.20.
$$\mathbf{u}_0^\top \mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t \mathbf{u}_0 \geqslant \|(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})\mathbf{H}^{\frac{1}{2}}\mathbf{u}_0\|^2$$

Proof. By the definition of positive semi-definite, we only need to prove that
$$\mathbf{A}_t^\top \mathbf{H}\mathbf{A}_t \succeq \mathbf{H}^{\frac{1}{2}}\mathbf{W}(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{W}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}\mathbf{H}^{\frac{1}{2}})^2\mathbf{H}^{\frac{1}{2}}$$

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

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

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

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

$$= \mathbf{I} - \mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}\mathbf{P} + \mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}e^{-2\mathbf{P}\mathbf{P}^\top t}\mathbf{P}.$$

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 C.21. *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

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

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

$$T_1 \leqslant \|\mathbf{H}_{0:k}\mathbf{W}_{0:k}^\top(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{M}_t(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|\|\mathbf{u}_{0:k}\|_2^2$$

$$\leqslant \|\mathbf{M}_t\|_2\|(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|_2^2\|\mathbf{u}_{0:k}\|_2^2$$

$$\leqslant \frac{c}{t}\|(\mathbf{W}\mathbf{H}\mathbf{W}^\top)^{-1}\mathbf{W}_{0:k}\mathbf{H}_{0:k}\|_2^2\|\mathbf{u}_{0:k}\|_2^2.$$

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

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

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

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

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

where the second 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 [63]).

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

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

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

$\square$

Combining Lemma C.17, C.18 and C.19, we get with probability at least $1 - e^{-\Omega(M)}$,

$$\mathbb{E}[\mathcal{E}_t] \asymp g_s(t) + \int_0^t \mathcal{K}(t-\tau)\gamma_\tau(c_\tau\mathbb{E}[\mathcal{E}_\tau] + \sigma^2)\,\mathrm{d}\tau + M^{-s\beta}, \tag{23}$$

which is of the same form as in Lemma C.12. From here following the same proof as before, we get the functional scaling laws for random projection matrices.

D   Proofs for Section 5

D.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 D.1 (Restatement of Theorem 5.2). *When the learning rate $\eta(k) \equiv \eta$ and batch size B are constants, 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 4.1, when the learning rate $\eta(k) \equiv \eta$, denote $\gamma := \frac{\eta}{B}$, we have

$$\mathbb{E}[\mathcal{E}_K] \eqsim (t)^{-s} + \gamma \int_0^t \mathcal{K}(t-r)(e(r) + \sigma^2)\,\mathrm{d}\tau + M^{-s\beta}.$$

Hence by Lemma C.9 and Lemma C.7, note that when $M^\beta < t$ we have $M^{-s\beta} \gtrsim \frac{1}{t^s} \gtrsim e(t)$, therefore

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

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

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

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

$\square$

Theorem D.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:*

$$\gamma_{\mathrm{opt}} \eqsim D^{-\frac{s}{s+1}}, \quad M_{\mathrm{opt}} \gtrsim D^{\frac{1}{(1+s)\beta}}, \qquad \mathcal{E}_{\mathrm{opt}} \eqsim D^{-\frac{s}{s+1}}. \tag{24}$$

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

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

Substituting this into the error expression yields

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

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

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

Consequently, the optimal loss rate becomes

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

$\square$

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 D.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:*

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

Proof. Since we have

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

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

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

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

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

which gives

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

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

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

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

$$= \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2uT}\left(\int_0^T e^{2ut}(a - \lambda t)\,\mathrm{d}t\right)\mathrm{d}u$$

$$= \int_{M^{-\beta}}^1 u^{1-1/\beta}e^{-2uT}\left[\frac{a}{2u}\left(e^{2uT} - 1\right) - \frac{\lambda}{2u}\left(Te^{2uT} - \frac{e^{2uT} - 1}{2u}\right)\right]\mathrm{d}u$$

$$= \int_{M^{-\beta}}^1 \left[\frac{a}{2u^{1/\beta}} - \frac{ae^{-2uT}}{2u^{1/\beta}} - \frac{\lambda T}{2u^{1/\beta}} + \frac{\lambda(1 - e^{-2uT})}{4u^{1+1/\beta}}\right]\mathrm{d}u$$

43

$$= \int_{M^{-\beta}}^{1} \left[\frac{a - \lambda T}{2u^{1/\beta}} - \frac{(a - \lambda T + \lambda T)e^{-2uT}}{2u^{1/\beta}} + \frac{\lambda(1 - e^{-2uT})}{4u^{1+1/\beta}} \right] du$$

$$= (a - \lambda T) \int_{M^{-\beta}}^{1} \frac{1 - e^{-2uT}}{2u^{1/\beta}} \, du + \lambda T \int_{M^{-\beta}}^{1} \left(-\frac{e^{-2uT}}{2u^{1/\beta}} + \frac{1 - e^{-2uT}}{4Tu^{1+1/\beta}} \right) du.$$

Thus, we complete the proof. $\square$

We next bound I_1 and I_2 separately.

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

Proof. Note that

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

and

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

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

Lemma D.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

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

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 11 for an illustration of $q_\beta(\cdot)$.
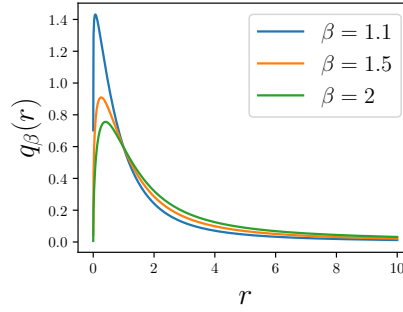

Figure 11: Illustration of the function $q_\beta(\cdot)$.

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

$$I_2 = C_{T,M} \frac{1}{4T^{1-1/\beta}} \int_{\frac{T}{M^\beta}}^{T} r^{-1-1/\beta} \, dr$$

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

Combining the two cases, we complete the proof. $\qquad\square$

Theorem D.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

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

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

Proof. By the functional scaling laws (7),

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

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

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

which gives

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

so we complete the proof. $\qquad\square$

Theorem D.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 $\gamma_{\text{opt}} \asymp (D/\log D)^{-\frac{1+s\beta-\beta}{1+s\beta}}$ and $\mathcal{E}_{\text{opt}} \asymp (D/\log D)^{-\frac{s\beta}{s\beta+1}}$.*

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

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

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

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

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

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

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

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

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

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

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.

45

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

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

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 $\gamma \asymp \tilde{D}^{\frac{\beta-1-s\beta}{1+s\beta}}$.

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

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

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

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

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 $\log K \asymp \log D$, $\tilde{D} \asymp \frac{D}{\log D}$.

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

Theorem D.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:*

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

  which leads to the following optimal final population risk:

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

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

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

  which leads to the following optimal final population risk:

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

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

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

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

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

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

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

and

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

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

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

and

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

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

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

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. $\qquad\square$

D.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 D.10. *For $T_2 > 0$, we have*

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

Proof. We have

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

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

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

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

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

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

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

$$\lesssim \frac{M}{T_2} \int_1^{M^\beta} u^{-\frac{1}{\beta}} e^{-2u}\, \mathrm{d}u \asymp \frac{M}{T_2}.$$

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

Theorem D.11 (Theorem 5.4 in the main paper). *Suppose the FSL (7) hold and M, K are sufficiently large. Then, we have*

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

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

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

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

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

Thus, we know

$$\int_0^{T(K)} \mathcal{K}(T(K) - t)\gamma_\varphi(t)\, \mathrm{d}t \asymp \int_0^{T_1} \mathcal{K}(T(K) - t)a\, \mathrm{d}t + b + \frac{(a - b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}$$

$$\asymp a \int_0^{T_1} \mathcal{K}(T_2 + t)\, \mathrm{d}t + b + \frac{(a - b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}$$

$$\asymp b + \frac{(a - b)\min\{M, T_2^{\frac{1}{\beta}}\}}{T_2}. \qquad \text{(by using Lemma D.10)}$$

Hence the loss is given by

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

$\qquad\square$

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

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

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

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

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

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

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

We see that in this case γ should be as large as possible, since $a \lesssim 1$, we set $\gamma \approx 1$ 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 $M \approx \tilde{D}_2^{\frac{1}{1+s\beta}}$.

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

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

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

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 $M^\beta > \gamma \tilde{D}_2$,

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

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 $\gamma \approx D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}}$.

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

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

When $s \leqslant 1 - \frac{1}{\beta}$, since $\gamma \lesssim 1$, we must have either $\gamma \approx 1$ or $\gamma \approx D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}} \lesssim 1$. 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

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

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 $\log K \approx \log D$, 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 D.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:*

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

    which leads to the following optimal final population risk:

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

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

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

    which leads to the following optimal final population risk:

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

49

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

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

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

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

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

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

and

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

Notice that

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

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

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

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

and

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

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

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

and

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

Notice that

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

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

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

and

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

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

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

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 $\gamma \asymp D^{-\frac{s\beta}{1+s\beta}} \tilde{D}_2^{\frac{\beta-1}{1+s\beta}}$.

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

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

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

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

which leads to the following optimal final population risk:

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

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

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

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

$$\gamma_{\mathrm{opt}} \asymp 1, M_{\mathrm{opt}} \asymp C^{\frac{1}{1+\beta}}, D_{\mathrm{opt}} \asymp C^{\frac{\beta}{1+\beta}},$$

which leads to the following optimal final population risk:

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

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

E   Auxiliary Lemmas

Lemma E.1. *For a semi-positive definite (SPD) matrix $\mathbf{A}$ and a random gaussian vector $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$,*

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

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 $i = j$

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

By the trace inequality we have
$$\mathbf{HA} \preceq \operatorname{tr}(\mathbf{HA}).$$

Multiplying $\mathbf{H}$ at both sides,
$$\mathbf{HAH} \preceq \operatorname{tr}(\mathbf{HA})\mathbf{H}.$$

Combining the results, we have
$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] = \operatorname{tr}(\mathbf{HA})\mathbf{H} + 2\mathbf{HAH} \preceq 2\operatorname{tr}(\mathbf{HA})\mathbf{H} + \mathbf{HAH}.$$

$\square$

Lemma E.2. *Let* $\mathbf{P} \preceq \mathbf{Q}$ *be two positive semi-definite (p.s.d.) matrices. Then for any p.s.d. matrix* $\mathbf{U}$*, we have*
$$\operatorname{tr}(\sqrt{\mathbf{P}}\mathbf{U}\sqrt{\mathbf{P}}) \leqslant \operatorname{tr}(\sqrt{\mathbf{Q}}\mathbf{U}\sqrt{\mathbf{Q}}).$$

Proof. It is clear that $\operatorname{tr}(\sqrt{\mathbf{P}}\mathbf{U}\sqrt{\mathbf{P}}) = \operatorname{tr}(\mathbf{UP})$ and
$$\operatorname{tr}(\mathbf{UQ}) - \operatorname{tr}(\mathbf{UP}) = \operatorname{tr}(\mathbf{U}(\mathbf{Q} - \mathbf{P})) \geqslant 0,$$

since $\mathbf{U}$ and $\mathbf{Q} - \mathbf{P}$ are both SPD matrices. $\square$