DEPTH IMPROVES TRAINING-TIME SCALING FOR DIFFICULT TASKS IN DEEP LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural scaling laws describe the predictable power-law dependence of model performance on training time, dataset size, and model scale, with the exponents quantifying how performance changes with resources. Recent theory has shown how scaling exponents are shaped by the data distribution and the structure of the learning task, but whether network depth can improve training-time scaling has remained unsettled. Through a refined analysis of deep linear networks, we clarify that the task spectrum determines when increasing network depth improves training-time scaling. For spectrally difficult tasks, increasing depth improves the asymptotic training-time scaling exponent, whereas for easier tasks, the exponent remains independent of depth, with depth affecting only a logarithmic correction at the boundary between these regimes. Remarkably, we find that a reparameterization of time governed by the predictor norm provides a unified description of the dynamics across task difficulties. Under this reparameterization, the loss follows a common scaling law whose exponent is independent of depth. Numerical experiments on a task derived from benchmark image data and with nonlinear networks exhibit qualitatively consistent task-dependent effects of depth on training-time scaling. Our findings advance the theoretical understanding of how task structure shapes the effects of network architecture on learning and offer a basis for studying scaling behavior across broader classes of neural networks.

1 INTRODUCTION

Performance in modern deep learning often improves predictably as training resources increase. Across a wide range of models and tasks, learning curves have often been observed to follow power laws with respect to training time, dataset size, and model scale (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). These empirical power-law relationships, commonly referred to as neural scaling laws, provide useful guidance for extrapolating performance to larger training scales and for allocating limited computational resources. Identifying what determines the exponents of these scaling laws is therefore an important step toward a theoretical understanding of learning in large neural networks.

Recent theory has clarified how scaling exponents are shaped by the structure of the data and the learning task (Berthier et al., 2020; Bordelon & Pehlevan, 2022; Maloney et al., 2022; Bordelon et al., 2024). For example, in kernel methods and random-feature models, scaling exponents depend on the spectral structure of the data and the target function (Spigler et al., 2020; Bordelon et al., 2020; Canatar et al., 2021; Defilippis et al., 2024). More recently, Bordelon et al. (2025) analyzed a two-layer model on power-law tasks and showed that training-time scaling depends on task difficulty. By comparison, the role of network architecture in determining these exponents remains less well characterized.

Depth provides a particularly direct way to study architectural effects on scaling. Deep linear networks are a fundamental model for studying the effects of depth (Baldi & Hornik, 1989; Fukumizu, 1998; Saxe et al., 2013). Although depth does not change the class of functions represented by the network, factorization across multiple layers induces nonlinear optimization dynamics, and this setting has enabled detailed analyses of convergence, optimization, and implicit bias (Arora et al., 2019; Du & Hu, 2019; Gidel et al., 2019; Gissin et al., 2020; Tarmoun et al., 2021). In particular, Arora et al. (2018) showed that network depth modifies the preconditioner governing the end-to-end

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

predictor dynamics and can thereby accelerate optimization. Their explicit acceleration result, however, is based on ℓ_p regression with $p > 2$, rather than the standard squared loss. More directly for scaling laws, Bordelon & Pehlevan (2025) derived scaling exponents for deep linear networks using a self-consistency argument inherited from the two-layer analysis of Bordelon et al. (2025), predicting no depth-induced acceleration in training-time scaling, with the same asymptotic exponent across depths. However, this argument directly applies the scaling relation obtained in the two-layer setting to general depth, without establishing its validity beyond two layers. Thus, even in deep linear networks, whether and when depth improves the training-time scaling exponent under standard squared loss has remained open.

In this work, we resolve this question through a refined analysis of deep linear networks on tasks with power-law spectral structure. By explicitly tracking the end-to-end predictor dynamics induced by the multilayer parameterization, we characterize the role of depth in training-time scaling. Our contributions are as follows:

- **Depth improves training-time scaling for spectrally difficult tasks.** We show that increasing depth can improve asymptotic training-time scaling under squared loss. For spectrally difficult tasks, where the target has relatively large components along low-eigenvalue directions of the data covariance, the decay exponent of the loss strictly increases with depth under sufficiently small random initialization (Theorem 4.5). We further show that the same asymptotic scaling extends to standard random initialization and characterize finite-dimensional effects that can delay its onset (Section 5).
- **Predictor-norm time reparameterization separates the effects of task spectrum and depth.** We introduce a reparameterized time defined through the predictor norm (Corollary 4.2). In this time coordinate, the loss has the same asymptotic scaling exponent across depths, determined solely by the task spectrum (Proposition 4.4).
- **Predictor-norm growth distinguishes three task regimes.** We prove that the predictor norm grows polynomially, logarithmically, or remains bounded depending on the task spectrum (Proposition 4.3). This characterization shows how predictor growth changes with task difficulty and, through the relation between reparameterized time and training time, determines the depth dependence of the training-time scaling law (Theorem 4.5).

Finally, experiments on a regression task derived from benchmark image data and on nonlinear ReLU networks exhibit qualitatively similar depth dependence across task regimes. These results provide a theoretical basis for studying how architecture and task structure jointly determine scaling laws in more complex neural networks.

2 Related Work

Theory of Scaling Laws and Spectral Learning Curves. Scaling laws have been derived analytically in a range of tractable learning models with structured spectra. Convergence rates have been characterized through spectral and regularity conditions on the data and target (Caponnetto & De Vito, 2007; Pillaud-Vivien et al., 2018; Cui et al., 2021). Sample-size learning curves have also been characterized in terms of the kernel eigenspectrum and the decomposition or alignment of the target in the associated eigenbasis (Spigler et al., 2020; Bordelon et al., 2020; Canatar et al., 2021).

Beyond sample-size scaling, theoretical work has characterized training-time scaling (Berthier et al., 2020; Bordelon & Pehlevan, 2022; Kramp et al., 2026). Scaling with dataset and model size (Maloney et al., 2022; Bahri et al., 2024; Lin et al., 2024; Defilippis et al., 2024; Atanasov et al., 2026) and compute-optimal scaling (Bordelon et al., 2024; Paquette et al., 2024) have also been studied.

Recent theoretical work has also begun to characterize scaling laws in nonlinear models with structured data or targets. These analyses include problems with structured or anisotropic data spectra (Worschech & Rosenow, 2025; Braun et al., 2026), as well as single- and multi-index models in which scaling laws emerge from the sequential acquisition of target components (Ren et al., 2025; Ben Arous et al., 2025; Defilippis et al., 2026). More directly related to our setting, Bordelon et al. (2025) analyzed a two-layer linear model on power-law tasks and showed that the training-time scaling exponent depends on task difficulty.

Deep Linear Networks and the Role of Depth. Deep linear networks provide a tractable setting for studying the role of depth in learning dynamics. Although the predictor remains linear in the input, its factorization across multiple layers induces nonlinear parameter dynamics that depend on depth. Early work characterized the loss landscapes and learning dynamics (Baldi & Hornik, 1989; Fukumizu, 1998; Saxe et al., 2013). Subsequent work studied convergence (Arora et al., 2019; Du & Hu, 2019), generalization (Lampinen & Ganguli, 2019; Advani et al., 2020), and implicit bias (Gidel et al., 2019; Gissin et al., 2020). Later work further characterized how initialization scale and layer balance shape the learning regime of deep linear networks (Woodworth et al., 2020; Tarmoun et al., 2021; Min et al., 2021; Jacot et al., 2021; Braun et al., 2022; Atanasov et al., 2022; Kunin et al., 2024; Domine et al., 2025).

An explicit mechanism by which depth can accelerate optimization was identified by Arora et al. (2018), who showed that additional layers modify the end-to-end predictor dynamics through a preconditioner. Their acceleration result, however, crucially exploits the structure of the ℓ_p loss for $p > 2$. For ℓ_p regression, the prediction-error dynamics involve a multiplicative interaction between the predictor-dependent preconditioning induced by depth and an error-dependent weighting arising from the ℓ_p loss. In their ill-conditioned two-coordinate example, the dominant coordinate constrains the global step size in the shallow parameterization, causing the other coordinate to converge much more slowly. In the deep parameterization, progress along the dominant coordinate increases the predictor norm, which raises the effective step size for the remaining coordinate and accelerates its convergence. At the same time, the shrinking error-dependent factor on the dominant coordinate suppresses its own update through the ℓ_p-dependent factor, preserving stability. For squared loss, this error-dependent factor is absent, so the mechanism underlying their acceleration result does not directly extend.

Most closely related to our setting, Bordelon & Pehlevan (2025) study learning on power-law structured tasks in deep linear networks and derive an asymptotic training-time scaling exponent that varies with task structure but not with network depth. Their scaling analysis builds on the self-consistency framework developed for the two-layer model in Bordelon et al. (2025). One component of this framework, the relation between the decay of the spectral error modes and the accumulated kernel scale, also applies beyond two layers. The depth dependence enters through the way layer growth determines the kernel scale. In the two-layer model, the kernel scale is determined by contributions from the hidden and readout layers, so its overall scale depends linearly on their growth. For deeper factorizations, however, each contribution to the kernel involves products across multiple layers, so the resulting kernel-growth relation depends on the number of layers. Their general-depth analysis nevertheless determined the scaling exponent using the two-layer kernel-growth relation without deriving its multilayer counterpart. Although the two analyses concern different settings and are therefore not directly comparable, they leave open whether and when depth changes training-time scaling under standard squared loss.

3 Problem Setup

Notation. For nonnegative quantities f and g, we write $f \lesssim g$ if $f \le Cg$, and $f \asymp g$ if $cg \le f \le Cg$, for positive constants c, C uniform in the parameters under consideration. We define $f \gtrsim g$ analogously. We take the maximum of an empty family of nonnegative quantities to be zero.

Data distribution. We consider a supervised regression problem with training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The inputs $\boldsymbol{x}_i$ are drawn i.i.d. from a distribution with zero mean and covariance

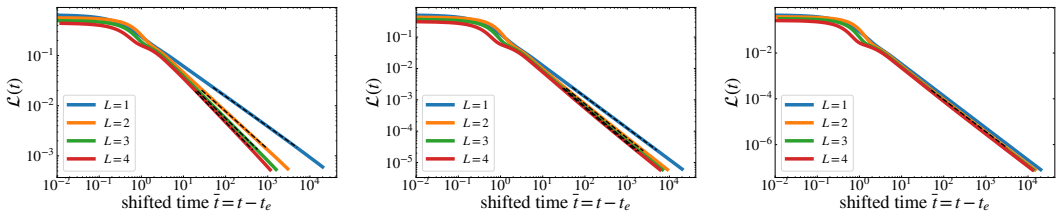
$$\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \qquad \lambda_i = c_\lambda i^{-a}. \tag{1}$$

Here, $c_\lambda > 0$ and $a > 0$. The labels are generated by the linear model $y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i$, with

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top, \qquad \beta_i = c_\beta i^{-b}, \tag{2}$$

where $c_\beta > 0$ and $b \in \mathbb{R}$.

We define the task exponent $p = (a + 2b - 1)/a$ and assume $p > 0$. This condition ensures that $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = c_\lambda c_\beta^2 \sum_{i=1}^d i^{-ap-1}$ remains bounded as $d \to \infty$. The exponent p characterizes the task spectrum and distinguishes the task regimes analyzed below. The case $p = 1$ (equivalently, $b = 1/2$) separates the polynomially divergent and bounded regimes of the target norm, with $\|\boldsymbol{\beta}\|^2$ growing logarithmically at the boundary.



(a) $p = 0.6$ (b) $p = 1.0$ (c) $p = 1.4$

Figure 1: Increasing depth improves the training-time scaling exponent for difficult tasks with $p < 1$, affects only the logarithmic correction at the boundary $p = 1$, and leaves the power-law exponent unchanged for easier tasks with $p > 1$. Dashed lines indicate the asymptotic training-time scaling laws given by Theorem 4.5.

Model. We parameterize the predictor using a linear network with $L \geq 1$ layers:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}, \qquad \boldsymbol{\theta} = (\boldsymbol{W}_L \cdots \boldsymbol{W}_1)^\top. \tag{3}$$

Here $\boldsymbol{W}_k \in \mathbb{R}^{n_k \times n_{k-1}}$, where n_k denotes the width of the k-th layer. We set $n_0 = d$ and $n_L = 1$ and assume $n_k \asymp d$ for $1 \leq k < L$. To track the scale associated with the depth-L factorization, we write $\sigma = \|\boldsymbol{\theta}\|^{1/L}$. Depth does not change the represented class of linear predictors. For $L \geq 2$, the factorized parameterization induces nonlinear dynamics.

Training algorithm and initialization. We train all layers jointly by population gradient flow for the squared loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\theta})\|^2$. For $j \geq k$, write $\boldsymbol{W}_{j:k} = \boldsymbol{W}_j \cdots \boldsymbol{W}_k$ with the conventions $\boldsymbol{W}_{0:1} = \boldsymbol{I}$ and $\boldsymbol{W}_{L:L+1} = 1$. For $1 \leq k \leq L$, the layer dynamics are

$$\frac{d}{dt}\boldsymbol{W}_k = -\nabla_{\boldsymbol{W}_k}\mathcal{L} = \boldsymbol{W}_{L:k+1}^\top (\boldsymbol{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\theta}))^\top \boldsymbol{W}_{k-1:1}^\top. \tag{4}$$

All entries are initialized independently according to $(\boldsymbol{W}_k(0))_{ij} \sim \mathcal{N}(0, s^2/n_{k-1})$ for $1 \leq k \leq L$, where $s > 0$ is the initialization scale.

4 Depth-Dependent Training-Time Scaling under Small Initialization

In this section, we derive the training-time loss scaling under sufficiently small random initialization and characterize when network depth improves the asymptotic decay exponent. The precise assumptions on the initialization scale s used in the analysis are stated in assumption B.1. To focus on the subsequent power-law scaling phase while accounting for the initial dynamics near the origin, we introduce a time t_e, defined precisely below, and use the shifted training time $\bar{t} = t - t_e$.

Numerical observation. Figure 1 shows the loss as a function of $\bar{t}$ for representative values of p below, at, and above 1. For $p < 1$, increasing depth produces a steeper power-law decay of the loss. At $p = 1$, the effect of depth is weaker, and for $p > 1$, the power-law exponent does not change.

Overview of the analysis. We proceed in several steps. We first analyze the end-to-end predictor dynamics and introduce the rescaled predictor $\boldsymbol{z} = \sigma^{-(L-1)}\boldsymbol{\theta}$, which cancels the leading depth-dependent anisotropy of the preconditioner (Section 4.1). We next analyze the initial escape phase and characterize the dependence of the time t_e on the initialization scale (Section 4.2). In the subsequent power-law scaling phase, we introduce the reparameterized time τ and analyze the dynamics in three steps. We first characterize the growth of the predictor norm as a function of τ (Section 4.3.1), then derive the loss scaling in τ and show that its exponent is independent of depth (Section 4.3.2), and finally translate this law back to the shifted training time $\bar{t}$ (Section 4.3.3).

4.1 Depth-Dependent Preconditioning and Rescaled Predictor Dynamics

In this subsection, we introduce the rescaled predictor $\boldsymbol{z} = \sigma^{-(L-1)}\boldsymbol{\theta}$ and derive its dynamics, for which the leading depth-dependent anisotropy of the end-to-end preconditioner is canceled. Since the loss depends on the network parameters only through the end-to-end predictor $\boldsymbol{\theta}$, we begin with

its dynamics, in which the effect of depth appears through a preconditioner. Specifically,

$$\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{K}_L \boldsymbol{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\theta}). \tag{5}$$

The preconditioner induced by the factorization is $\boldsymbol{K}_L = \sum_{k=1}^{L} \|\boldsymbol{W}_{L:k+1}\|^2 \boldsymbol{W}_{k-1:1}^\top \boldsymbol{W}_{k-1:1}$. From Equation (4), the adjacent-layer Gram differences are conserved: $\frac{d}{dt}(\boldsymbol{W}_k \boldsymbol{W}_k^\top) = \frac{d}{dt}(\boldsymbol{W}_{k+1}^\top \boldsymbol{W}_{k+1})$. We quantify this imbalance by $\varepsilon = \max_{1 \le k < L} \|\boldsymbol{W}_k \boldsymbol{W}_k^\top - \boldsymbol{W}_{k+1}^\top \boldsymbol{W}_{k+1}\|_{\mathrm{op}}$. When ε/σ^2 is small, this near-balance keeps all layers at the common scale σ, so each term in $\boldsymbol{K}_L$ has overall scale $\sigma^{2(L-1)}$. As shown in Proposition A.4, we get

$$\boldsymbol{K}_L = \sigma^{2(L-1)} \left(\boldsymbol{I} + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}} + \Delta \overline{\boldsymbol{K}}_L \right). \tag{6}$$

Here, for $\boldsymbol{\theta} \neq \boldsymbol{0}$, we define $\boldsymbol{P}_{\boldsymbol{\theta}} = \boldsymbol{\theta}\boldsymbol{\theta}^\top / \|\boldsymbol{\theta}\|^2$, the orthogonal projector onto $\mathrm{span}\{\boldsymbol{\theta}\}$, and $\|\Delta \overline{\boldsymbol{K}}_L\|_{\mathrm{op}} = O(\varepsilon/\sigma^2)$. Equation (6) shows that, for $L \ge 2$, the leading part of the preconditioner is anisotropic, with the depth dependence appearing through the rank-one term $(L-1)\boldsymbol{P}_{\boldsymbol{\theta}}$. This motivates the rescaling $\boldsymbol{z} = \sigma^{-(L-1)}\boldsymbol{\theta}$. Defining $\boldsymbol{J}_L = \sigma^{L-1}\partial_{\boldsymbol{\theta}}\boldsymbol{z} = \boldsymbol{I} - \frac{L-1}{L}\boldsymbol{P}_{\boldsymbol{\theta}}$, we have $\boldsymbol{J}_L(\boldsymbol{I} + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}}) = \boldsymbol{I}$, so the leading anisotropic factor is exactly canceled in the dynamics of $\boldsymbol{z}$.

Lemma 4.1 (Rescaled predictor dynamics). *For $\boldsymbol{z} = \sigma^{-(L-1)}\boldsymbol{\theta}$, the dynamics satisfy*

$$\frac{d\boldsymbol{z}}{dt} = (\boldsymbol{I} + \boldsymbol{J}_L \Delta \overline{\boldsymbol{K}}_L)\boldsymbol{\Sigma}(\sigma^{L-1}\boldsymbol{\beta} - \sigma^{2(L-1)}\boldsymbol{z}). \tag{7}$$

See Appendix A.3 for the derivation.

4.2 Escape from the origin

Before analyzing the power-law scaling phase, we characterize the initial escape from the origin. In particular, we determine how the time t_{e} at which the predictor scale σ has reached an $O(1)$ scale depends on the initialization scale s. From Equation (7), when $\sigma \ll 1$, the relaxation term is lower order and the forcing term has scale σ^{L-1}, suggesting $d\sigma/dt \approx \sigma^{L-1}$. Hence the time t_{e} satisfies

$$t_{\mathrm{e}} \asymp \begin{cases} 1, & L = 1, \\ \log(1/s), & L = 2, \\ s^{-(L-2)}, & L \ge 3. \end{cases} \tag{8}$$

A rigorous derivation is given in Corollary B.4. We therefore measure the subsequent dynamics using the shifted time $\bar{t} = t - t_{\mathrm{e}}$.

4.3 Predictor Norm and Loss Scaling

To characterize the loss scaling in the power-law scaling phase, we introduce a reparameterized time in which the leading spectral dynamics take a common form across depths. This separates the common spectral scaling in reparameterized time from the depth-dependent conversion back to training time.

Corollary 4.2 (Dynamics in reparameterized time). *We define the time reparameterization*

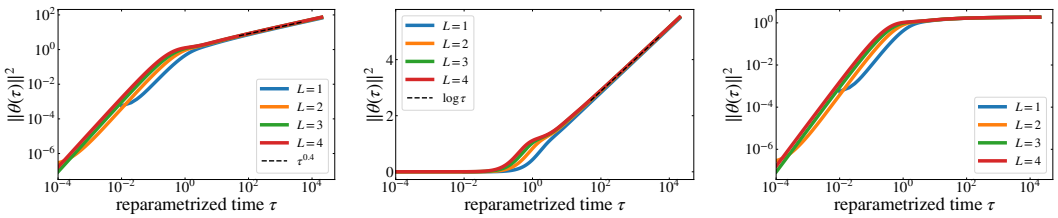
$$\tau(t) = \int_0^t \sigma(v)^{2(L-1)}\, dv. \tag{9}$$

Then the rescaled predictor satisfies:

$$\frac{d\boldsymbol{z}}{d\tau} = -\boldsymbol{\Sigma}\boldsymbol{z} + \sigma^{-(L-1)}\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{f}_K, \tag{10}$$

where $\boldsymbol{f}_K = \boldsymbol{J}_L \Delta \overline{\boldsymbol{K}}_L \boldsymbol{\Sigma} \left(\sigma^{-(L-1)}\boldsymbol{\beta} - \boldsymbol{z} \right)$.

See Appendix A.3 for the derivation. The relaxation term $-\boldsymbol{\Sigma}\boldsymbol{z}$ is independent of depth. Depth enters only through the predictor scale σ and the perturbation $\boldsymbol{f}_K$. The subsequent analysis therefore reduces to characterizing the growth of the predictor norm and the resulting spectral learning dynamics in τ.



(a) $p = 0.6$ (b) $p = 1.0$ (c) $p = 1.4$

Figure 2: Growth of the squared predictor norm distinguishes three task regimes: polynomial growth for difficult tasks with $p < 1$, logarithmic growth at the boundary $p = 1$, and bounded behavior for easier tasks with $p > 1$. Dashed lines indicate the asymptotic scaling laws given by Proposition 4.3. Panel (b) uses a linear vertical axis to make the logarithmic growth visible.

Since $\lambda_d \asymp d^{-a}$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$, the slowest spectral time scale is of order d^a. We therefore fix a sufficiently small constant $c_0 > 0$ and set $\tau_+ = c_0 d^a$. Let $t_+ > t_e$ be the corresponding time defined by $\tau(t_+) = \tau_+$, and set $\bar{t}_+ = t_+ - t_e$. We restrict the analysis in this subsection to $t_e \le t \le t_+$, equivalently $\tau_e \le \tau \le \tau_+$ with $\tau_e = \tau(t_e)$, before effects associated with the smallest eigenvalue become relevant. Beyond this window, the loss decays exponentially in reparameterized time (see Corollary C.4).

4.3.1 Scaling of the predictor norm

The relation between reparameterized time τ and training time is controlled by the predictor scale, since $d\tau/dt = \sigma^{2(L-1)}$. We therefore characterize $\sigma(\tau) = \|\boldsymbol{\theta}(\tau)\|^{1/L}$ through the spectral dynamics of the predictor. Applying the variation of constants formula to the dynamics above and using $\boldsymbol{\theta} = \sigma^{L-1}\boldsymbol{z}$, we obtain

$$\boldsymbol{\theta}(\tau) = \sigma(\tau)^{L-1} \int_0^\tau e^{-(\tau-v)\boldsymbol{\Sigma}} \boldsymbol{\Sigma}\boldsymbol{\beta}\, \sigma(v)^{-(L-1)}\, dv + \boldsymbol{R}(\tau), \tag{11}$$

where $\boldsymbol{R}(\tau)$ collects the effects of initialization and the conserved layer imbalance (see Corollary B.6 for its exact representation). Under sufficiently small initialization, $\boldsymbol{R}(\tau)$ does not change the predictor-norm scaling derived below, as shown in Appendix D.

In the i-th eigendirection, the exponential factor $e^{-(\tau-v)\lambda_i}$ suppresses contributions from times v much earlier than τ, over a time scale $1/\lambda_i$. The behavior therefore changes around $\tau\lambda_i \asymp 1$. For $\tau\lambda_i \gg 1$, the integral is dominated by v near τ, giving $\theta_i(\tau) \approx \beta_i$. For $\tau\lambda_i \ll 1$, we have $e^{-(\tau-v)\lambda_i} \approx 1$, so $\theta_i(\tau) \approx \beta_i\lambda_i\, \sigma(\tau)^{L-1} \int_0^\tau \sigma(v)^{-(L-1)}\, dv$. To obtain the scaling behavior, we make the self-consistent ansatz $\sigma(\tau)^{L-1} \int_0^\tau \sigma(v)^{-(L-1)}\, dv \asymp \tau$. This gives $\theta_i(\tau) \asymp \beta_i\tau\lambda_i$. Combining the two cases and summing over the eigendirections gives $\|\boldsymbol{\theta}(\tau)\|^2 \asymp \sum_{i=1}^d \beta_i^2 \min\{\tau\lambda_i, 1\}^2$. Substituting $\lambda_i = c_\lambda i^{-a}$ and $\beta_i = c_\beta i^{-b}$ into this expression, we obtain

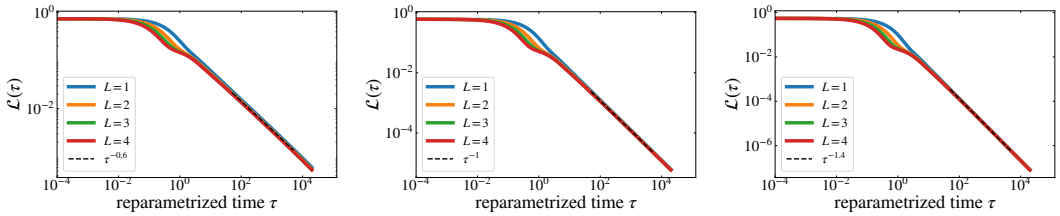
Proposition 4.3 (Predictor-norm scaling, informal)**.** *Under sufficiently small initialization, there exists a constant $\tau_0 > 0$, independent of d and s, such that, for all sufficiently large d, with high probability, uniformly for $\tau_0 \le \tau \le \tau_+$,*

$$\|\boldsymbol{\theta}(\tau)\|^2 \asymp \begin{cases} \tau^{1-p}, & p < 1, \\ \log \tau, & p = 1, \\ 1, & p > 1. \end{cases} \tag{12}$$

The resulting scaling of $\sigma(\tau) = \|\boldsymbol{\theta}(\tau)\|^{1/L}$ satisfies the assumed relation, confirming the self-consistency of the ansatz. A rigorous derivation is given in Corollary D.3 and does not rely on the self-consistency ansatz. Figure 2 illustrates the three regimes for the squared predictor norm: polynomial growth, logarithmic growth, and bounded behavior, respectively.

4.3.2 Common spectral law for the loss

We next derive the loss scaling in reparameterized time τ and show that its scaling exponent depends on the task spectrum but is independent of network depth. The same spectral threshold $\tau\lambda_i \asymp 1$ that



(a) $p = 0.6$ (b) $p = 1.0$ (c) $p = 1.4$

Figure 3: In reparameterized time, the loss follows a common depth-independent spectral scaling law, with exponent determined solely by the task spectrum. Dashed lines indicate the asymptotic reparameterized-time scaling laws given by Proposition 4.4.

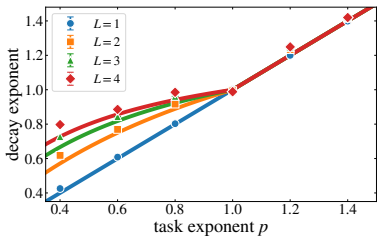


Figure 4: Task difficulty determines the magnitude of the depth-induced improvement in the training-time power-law exponent. Solid curves show the asymptotic exponents given by Theorem 4.5. Markers show numerical estimates averaged over five random seeds at dimension $d = 2048$, and error bars show one standard deviation across seeds. At $p = 1$, we account for the logarithmic correction by fitting $\mathcal{L}(\bar{t})(\log \bar{t})^{(L-1)/L}$ to a power law.

controls the predictor norm also determines which eigendirections contribute to the loss. Recall that $\mathcal{L}(\tau) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i(\beta_i - \theta_i(\tau))^2$. For $\tau\lambda_i \gg 1$, the corresponding mode is essentially learned, with $\theta_i(\tau) \approx \beta_i$, whereas for $\tau\lambda_i \ll 1$, it remains largely unlearned, with $\beta_i - \theta_i(\tau) \asymp \beta_i$. Thus, to leading order, the loss is governed by the unresolved spectral tail $\sum_{\lambda_i \lesssim 1/\tau} \lambda_i \beta_i^2$. Under $\lambda_i = c_\lambda i^{-a}$ and $\beta_i = c_\beta i^{-b}$, this tail is governed by the exponent $p = (a + 2b - 1)/a$.

Proposition 4.4 (Common spectral law for the loss, informal)**.** *Under sufficiently small initialization, there exists a constant $\tau_0 > 0$, independent of d and s, such that, for sufficiently large d, with high probability, uniformly for $\tau_0 \le \tau \le \tau_+$,*

$$\mathcal{L}(\tau) \asymp \tau^{-p}. \tag{13}$$

A rigorous proof is given in Proposition E.6. Thus, in reparameterized time, the loss decays with the same power-law exponent p across all depths. Figure 3 illustrates this common scaling numerically.

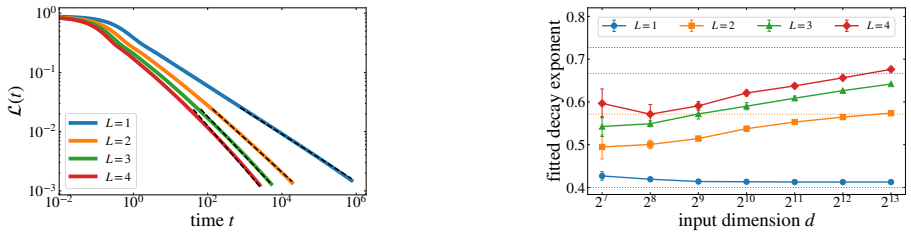
4.3.3 Training-time loss scaling

We finally translate the loss scaling in reparameterized time τ back to the shifted training time $\bar{t} = t - t_e$ and determine how the resulting training-time scaling depends on network depth. Since $d\tau/d\bar{t} = \sigma^{2(L-1)}$, the predictor-norm scaling in Proposition 4.3 determines the relation between τ and $\bar{t}$. Combining this relation with the loss scaling in Proposition 4.4 gives the following result.

Theorem 4.5 (Training-time loss scaling, informal)**.** *Under sufficiently small initialization, there exists a constant $\bar{t}_0 > 0$, independent of d and s, such that, for sufficiently large d, with high probability, uniformly for $\bar{t}_0 \le \bar{t} \le \bar{t}_+$,*

$$\mathcal{L}(t_e + \bar{t}) \asymp \begin{cases} \bar{t}^{-pL/[1+p(L-1)]}, & p < 1, \\ \bar{t}^{-1}(\log \bar{t})^{-(L-1)/L}, & p = 1, \\ \bar{t}^{-p}, & p > 1. \end{cases} \tag{14}$$

A rigorous proof is given in Theorem F.1. For $p < 1$, the decay exponent $pL/[1 + p(L - 1)]$ increases with depth. For $p > 1$, the power-law exponent is independent of depth, while at $p = 1$, depth affects only the logarithmic correction. Figure 4 summarizes the resulting dependence of the decay exponent on the task exponent p and network depth.



(a) Loss dynamics

(b) Finite-size effects on the fitted decay exponent

Figure 5: The depth-dependent asymptotic scaling under standard initialization ($s = 1$) emerges with increasing dimension, while finite-size effects make this scaling harder to observe in deeper networks. (a) Loss dynamics for the $p = 0.4$ task at dimension $d = 8192$. Dashed lines show the asymptotic power-law decay predicted by Corollary 5.1, with exponents corresponding to the asymptotic values shown by the dotted horizontal lines in panel (b). (b) As dimension increases, the fitted decay exponents approach these asymptotic values. Markers show the mean over five random seeds, and error bars show one standard deviation across seeds.

5 Standard Initialization and Finite-Size Effects for Difficult Tasks

Having shown that depth improves the training-time scaling exponent for difficult tasks under sufficiently small initialization, we next ask whether this improvement persists under standard initialization. We focus on the difficult-task regime $p < 1$, where the predictor norm grows polynomially during training. This growth makes the effect of the $O(1)$ initialization scale asymptotically negligible.

Corollary 5.1 (Training-time loss scaling under standard initialization, informal)**.** *Suppose $p < 1$ and consider standard initialization with $s = 1$. There exists a constant $t_0 > 0$, independent of d, such that, for sufficiently large d, with high probability, uniformly for $t_0 \le t \le t_+$,*

$$\mathcal{L}(t) \asymp t^{-pL/[1+p(L-1)]}. \tag{15}$$

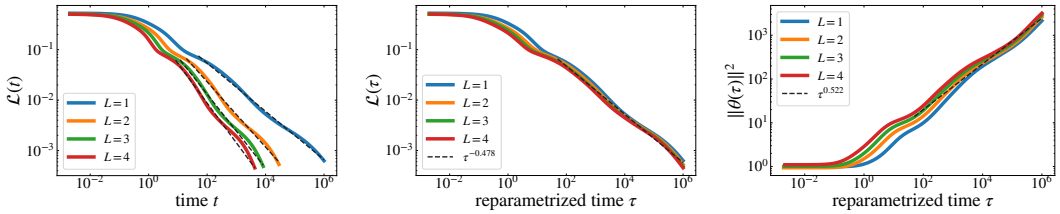
A rigorous proof is given in Corollary F.2. Figure 5a shows that the loss decays faster with increasing depth, consistent with the predicted depth dependence.

At finite input dimension, however, the predicted asymptotic scaling may not be clearly observable. Figure 5b shows the fitted decay exponent as a function of the input dimension d. As d increases, the fitted exponents approach their asymptotic predictions, but this convergence becomes slower with increasing depth. To understand this finite-size effect, recall that the power-law scaling window is restricted to $\tau \le \tau_+ = c_0 d^a$. Within this window, the largest asymptotic scale available to the squared predictor norm is $\tau_+^{1-p} \asymp d^{a(1-p)} = d^{1-2b}$. Although this quantity diverges as $d \to \infty$ for $p < 1$, the divergence becomes slower as $p \uparrow 1$. Moreover, the corresponding per-layer scale grows only as $\sigma(\tau_+) \asymp d^{a(1-p)/(2L)} = d^{(1-2b)/(2L)}$. This suggests that, as depth L increases, the per-layer scale separates more slowly from its $O(1)$ initialization scale. Consequently, especially for large L or for p close to 1, increasingly large dimensions may be required for the asymptotic power-law scaling to emerge clearly, making the ideal depth-dependent exponent difficult to observe at finite input dimension.

6 Experiments on Real Data and Nonlinear Networks

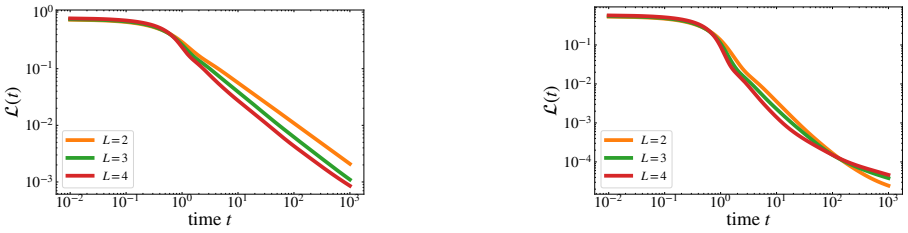
Finally, we numerically examine whether the task-dependent effect of depth extends beyond the analytically tractable setting to real data and nonlinear networks.

First, we train deep linear networks with standard initialization ($s = 1$) on a regression task constructed from CIFAR-5M (Nakkiran et al., 2021), whose empirical task spectrum is approximately power-law with $p_{\text{eff}} \approx 0.478$, placing it in the $p < 1$ regime. As shown in Figure 6, the loss decay becomes steeper with increasing depth in training time. The loss trajectories nearly collapse under the predictor-norm time reparameterization, while the squared predictor norm exhibits approximately the predicted polynomial growth.



(a) Loss in training time t (b) Loss in reparameterized time τ (c) Growth of squared predictor norm

Figure 6: The depth-dependent acceleration persists on a task derived from benchmark image data. (a) In training time, the loss decays faster with increasing depth. (b) In reparameterized time, the loss trajectories nearly collapse onto a common spectral scaling law. (c) The squared predictor norm exhibits the predicted polynomial growth. Dashed lines show the theoretical slopes obtained from $p_{\text{eff}} \approx 0.478$: $p_{\text{eff}} L/[1 + p_{\text{eff}}(L-1)]$ in (a), p_{eff} in (b), and $1 - p_{\text{eff}}$ in (c).



(a) $p = 0.6$ (b) $p = 1.4$

Figure 7: The task-dependent effect of depth persists qualitatively in nonlinear ReLU networks. (a) For the difficult task with $p = 0.6$, deeper networks achieve lower loss at late training times. (b) For the easier task with $p = 1.4$, the loss trajectories converge and cross, with no consistent late-time advantage from increasing depth.

We also test ReLU MLPs on the synthetic power-law tasks used in our theoretical analysis. As shown in Figure 7, for the difficult $p = 0.6$ task, the loss trajectories separate at late training times, with deeper networks attaining lower loss, whereas for the easier $p = 1.4$ task the trajectories remain much closer and show no systematic advantage with increasing depth. These observations are qualitatively consistent with the task-dependent trend predicted by our theory, although the distinction is less pronounced than for the deep linear networks.

7 Conclusion

In this work, we showed for the first time that task difficulty explains the previously unresolved depth dependence of convergence in the most fundamental setting of deep linear networks trained with MSE loss. That is, increasing network depth improves the asymptotic training-time scaling exponent for spectrally difficult tasks. A time reparameterization defined through the predictor norm reveals that the loss follows a depth-independent spectral scaling law in reparameterized time, while depth changes its relation to training time. This also explains why depth does not improve the power-law exponent for easier tasks.

Our analysis has several limitations. The theory assumes a linear teacher and student and population gradient flow, while our results for standard initialization are limited to difficult tasks. It remains to understand how the mechanisms identified here extend to more general task structures, nonlinear networks, and finite-sample training with discrete-time stochastic optimization.

Even in deep linear networks, explicitly accounting for task structure can substantially complicate a sharp characterization of the training dynamics. Our results show that this structure can qualitatively change how network depth affects training-time scaling. These results provide a basis for studying such architectural effects in broader classes of neural networks.

AI USE STATEMENT

In this work, we used generative AI tools to help refine and formalize aspects of the theoretical framework and presentation, to refine the formulation of some mathematical statements and technical assumptions, and to assist with writing and formalizing proofs. The central research hypotheses, core mathematical ideas, and main theoretical mechanisms were developed by the authors. We also used generative AI tools to generate and edit code implementing numerical experiments, fitting procedures, and plotting routines, and for translation. The experimental settings and comparisons were specified by the authors. Additionally, we used generative AI tools to suggest some experimental parameters, to draft and edit parts of the manuscript, and to identify relevant literature and examine specific mathematical arguments in prior work. The authors reviewed and verified all AI-assisted text, mathematical claims and proofs, code, experimental results, analyses, and citations, and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The assumptions, technical derivations, and complete proofs for the theoretical results are provided in Appendices A to F. Details for reproducing the numerical experiments are provided in Appendix G, including experimental setups, data processing, optimization settings, and fitting procedures.

REFERENCES

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pp. 244–253. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4): 043404, 2026.

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.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Gerard Ben Arous, Murat Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146725–146812, 2025.

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.

Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In *International Conference on Learning Representations*, 2022.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4968–4997. PMLR, 2025.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. In *International Conference on Learning Representations*, 2025.

Guillaume Braun, Bruno Loureiro, Minh Ha Quang, and Masaaki Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data. In *International Conference on Learning Representations*, 2026.

Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, 2022.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. Optimal scaling laws in learning hierarchical multi-index models. *arXiv preprint arXiv:2602.05846*, 2026.

Clementine Domine, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro Mediano, and Andrew Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025.

Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pp. 1655–1664. PMLR, 2019.

Kenji Fukumizu. Effect of batch learning in multilayer neural networks. In *Proceedings of the 5th International Conference on Neural Information Processing*, pp. 67–70, 1998.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

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

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

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

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.

Jakob Kramp, Javed Lindner, and Moritz Helias. Dynamics of neural scaling laws in random feature regression with powerlaw-distributed kernel eigenvalues. *arXiv preprint arXiv:2602.23039*, 2026.

Daniel Kunin, Allan Raventos, Clémentine Carla Juliette Dominé, Feng Chen, David Klindt, Andrew M Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2019.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

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

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pp. 7760–7768. PMLR, 2021.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations*, 2021.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

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.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38227–38309, 2025.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

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.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pp. 10153–10161. PMLR, 2021.

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*, pp. 23549–23588. PMLR, 2022.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. In *International Conference on Learning Representations*, 2025.

A Gradient-Flow Dynamics, Preconditioner Geometry, and Initialization Bounds

This section establishes the gradient-flow identities, preconditioner estimates, and Gaussian initialization bounds used throughout the analysis.

A.1 Exact Gradient-Flow Dynamics

Throughout the appendices, $c, C > 0$ denote generic constants that may change from line to line. Unless otherwise specified, they do not depend on d or on the variables quantified in the corresponding estimate.

Throughout the appendices, write $\boldsymbol{g} := \boldsymbol{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\theta})$ and $r := \|\boldsymbol{\theta}\| = \sigma^L$. On any interval on which $r > 0$, write $\widehat{\boldsymbol{\theta}} := \boldsymbol{\theta}/r$.

The preconditioner induced by the factorization is

$$\boldsymbol{K}_L := \sum_{k=1}^{L} \|\boldsymbol{W}_{L:k+1}\|^2 \boldsymbol{W}_{k-1:1}^\top \boldsymbol{W}_{k-1:1}. \tag{16}$$

For every $1 \le k \le L$, using $\nabla_{\boldsymbol{\theta}} \mathcal{L} = -\boldsymbol{g}$ and applying the chain rule to the factorization $\boldsymbol{\theta} = (\boldsymbol{W}_{L:k+1} \boldsymbol{W}_k \boldsymbol{W}_{k-1:1})^\top$, we obtain

$$\frac{d}{dt} \boldsymbol{W}_k = \boldsymbol{W}_{L:k+1}^\top \boldsymbol{g}^\top \boldsymbol{W}_{k-1:1}^\top. \tag{17}$$

Differentiating the layer product and using Equation (17) yields

$$\frac{d}{dt} \boldsymbol{\theta} = \sum_{k=1}^{L} \|\boldsymbol{W}_{L:k+1}\|^2 \boldsymbol{W}_{k-1:1}^\top \boldsymbol{W}_{k-1:1} \boldsymbol{g} = \boldsymbol{K}_L \boldsymbol{g}. \tag{18}$$

Together with $\nabla_{\boldsymbol{\theta}} \mathcal{L} = -\boldsymbol{g}$, this gives

$$\frac{d}{dt} \mathcal{L} = -\boldsymbol{g}^\top \boldsymbol{K}_L \boldsymbol{g} = -\sum_{k=1}^{L} \left\| \frac{d}{dt} \boldsymbol{W}_k \right\|_{\mathrm{F}}^2. \tag{19}$$

The last equality follows by taking the squared Frobenius norm of both sides of Equation (17) and summing over the layers.

On such an interval, define the normalized preconditioner by

$$\overline{\boldsymbol{K}}_L := \sigma^{-2(L-1)} \boldsymbol{K}_L. \tag{20}$$

Since $d\tau/dt = \sigma^{2(L-1)} > 0$ on this interval, reparameterizing the predictor and loss identities in terms of τ gives

$$\frac{d}{d\tau} \boldsymbol{\theta} = \overline{\boldsymbol{K}}_L \boldsymbol{g}, \tag{21}$$

$$\frac{d}{d\tau} \mathcal{L} = -\boldsymbol{g}^\top \overline{\boldsymbol{K}}_L \boldsymbol{g}. \tag{22}$$

Lemma A.1. *For every finite d and any initialization, the gradient-flow system in Equation (17) admits a unique solution for all $t \ge 0$.*

Proof. The vector field is polynomial in the weights and hence locally Lipschitz. Therefore, the system admits a unique local solution. Let $[0, T_{\max})$ be the maximal interval of existence of this solution. For every $T < T_{\max}$, Cauchy–Schwarz and Equation (19) give

$$\begin{aligned} \left(\sum_{k=1}^{L} \|\boldsymbol{W}_k(T) - \boldsymbol{W}_k(0)\|_{\mathrm{F}}^2 \right)^{1/2} &\le \int_0^T \left(\sum_{k=1}^{L} \left\| \frac{d}{dt} \boldsymbol{W}_k \right\|_{\mathrm{F}}^2 \right)^{1/2} dt \\ &\le \sqrt{T} \left(\mathcal{L}(0) - \mathcal{L}(T) \right)^{1/2} \\ &\le \sqrt{T \mathcal{L}(0)}. \end{aligned} \tag{23}$$

If $T_{\max} < \infty$, then Equation (23) shows that every weight matrix remains bounded on $[0, T_{\max})$, contradicting the ODE continuation theorem. Therefore $T_{\max} = \infty$. □

A.2 Layer Balance and Preconditioner Geometry

To quantify the deviation from layer balance at initialization, define

$$\varepsilon := \max_{1\le k<L} \left\| \boldsymbol{W}_k(0)\boldsymbol{W}_k(0)^\top - \boldsymbol{W}_{k+1}(0)^\top \boldsymbol{W}_{k+1}(0) \right\|_{\mathrm{op}}. \tag{24}$$

Lemma A.2. *For every $1 \le k < L$ and every $t \ge 0$,*

$$\boldsymbol{W}_k(t)\boldsymbol{W}_k(t)^\top - \boldsymbol{W}_{k+1}(t)^\top \boldsymbol{W}_{k+1}(t) = \boldsymbol{W}_k(0)\boldsymbol{W}_k(0)^\top - \boldsymbol{W}_{k+1}(0)^\top \boldsymbol{W}_{k+1}(0). \tag{25}$$

Consequently, at every $t \ge 0$,

$$\max_{1\le k<L} \left\| \boldsymbol{W}_k(t)\boldsymbol{W}_k(t)^\top - \boldsymbol{W}_{k+1}(t)^\top \boldsymbol{W}_{k+1}(t) \right\|_{\mathrm{op}} = \varepsilon. \tag{26}$$

Proof. Using Equation (17) together with $\boldsymbol{W}_{k:1} = \boldsymbol{W}_k \boldsymbol{W}_{k-1:1}$ and $\boldsymbol{W}_{L:k+1} = \boldsymbol{W}_{L:k+2}\boldsymbol{W}_{k+1}$ gives

$$\frac{d}{dt}(\boldsymbol{W}_k \boldsymbol{W}_k^\top) = \boldsymbol{W}_{L:k+1}^\top \boldsymbol{g}^\top \boldsymbol{W}_{k:1}^\top + \boldsymbol{W}_{k:1}\boldsymbol{g}\boldsymbol{W}_{L:k+1} = \frac{d}{dt}(\boldsymbol{W}_{k+1}^\top \boldsymbol{W}_{k+1}). \tag{27}$$

Hence the difference in Equation (25) has zero derivative. Its value at time zero proves both conclusions. □

Define

$$\Delta\overline{\boldsymbol{K}}_L := \overline{\boldsymbol{K}}_L - \boldsymbol{I} - (L-1)\boldsymbol{P}_{\boldsymbol{\theta}}. \tag{28}$$

Suppose that $r > 0$. Set $\boldsymbol{\rho}_L = 1$ and, for $k = L, L-1, \dots, 1$, recursively define

$$\sigma_k := \|\boldsymbol{W}_k^\top \boldsymbol{\rho}_k\|, \quad \boldsymbol{\rho}_{k-1} := \frac{\boldsymbol{W}_k^\top \boldsymbol{\rho}_k}{\sigma_k}, \quad \boldsymbol{W}_k^\perp := (\boldsymbol{I} - \boldsymbol{\rho}_k \boldsymbol{\rho}_k^\top)\boldsymbol{W}_k. \tag{29}$$

A backward induction shows simultaneously that this recursion is well defined and that

$$\boldsymbol{W}_{L:k+1} = \left(\prod_{j=k+1}^{L} \sigma_j \right) \boldsymbol{\rho}_k^\top, \tag{30}$$

for $0 \le k \le L$. Indeed, the identity holds at $k = L$ by the convention $\boldsymbol{W}_{L:L+1} = 1$. If it holds at k and $\sigma_k = 0$, then $\boldsymbol{W}_{L:k} = 0$ and hence $\boldsymbol{W}_{L:1} = 0$, contradicting $r > 0$. Thus $\sigma_k > 0$, and the defining relation $\boldsymbol{W}_k^\top \boldsymbol{\rho}_k = \sigma_k \boldsymbol{\rho}_{k-1}$ yields the identity at $k-1$.

By construction, each $\boldsymbol{\rho}_k$ is a unit vector and

$$\boldsymbol{W}_k = \sigma_k \boldsymbol{\rho}_k \boldsymbol{\rho}_{k-1}^\top + \boldsymbol{W}_k^\perp, \qquad \boldsymbol{\rho}_k^\top \boldsymbol{W}_k^\perp = 0. \tag{31}$$

Taking $k = 0$ in Equation (30) and comparing with $\boldsymbol{W}_{L:1} = \boldsymbol{\theta}^\top = r\widehat{\boldsymbol{\theta}}^\top$ gives

$$\prod_{k=1}^{L} \sigma_k = r = \sigma^L, \qquad \boldsymbol{\rho}_0 = \widehat{\boldsymbol{\theta}}. \tag{32}$$

Lemma A.3. *There exists a constant $C > 0$, depending only on L, such that at every time for which $r > 0$, the decomposition in Equation (31) satisfies*

$$\max_{1\le k\le L} \|\boldsymbol{W}_k^\perp\|_{\mathrm{op}}^2 \le C\varepsilon, \qquad \max_{1\le k\le L} |\sigma_k^2 - \sigma^2| \le C\varepsilon. \tag{33}$$

Consequently, for any fixed $M > 0$, if $\varepsilon \le M\sigma^2$, then

$$\sigma_k \asymp \sigma, \tag{34}$$

for $1 \le k \le L$, where the comparison constants depend only on L and M.

Proof. The orthogonality in the decomposition gives

$$\boldsymbol{W}_{k+1}^{\top}\boldsymbol{W}_{k+1} = \sigma_{k+1}^2\boldsymbol{\rho}_k\boldsymbol{\rho}_k^{\top} + (\boldsymbol{W}_{k+1}^{\perp})^{\top}\boldsymbol{W}_{k+1}^{\perp}. \tag{35}$$

Taking the quadratic form of the adjacent Gram difference along $\boldsymbol{\rho}_k$ and using Equation (26), we obtain

$$|\sigma_k^2 - \sigma_{k+1}^2| \le \|\boldsymbol{W}_{k+1}^{\perp}\|_{\mathrm{op}}^2 + \varepsilon. \tag{36}$$

Projecting the same Gram difference onto the orthogonal complement of $\boldsymbol{\rho}_k$ gives

$$\|\boldsymbol{W}_k^{\perp}\|_{\mathrm{op}}^2 \le \|\boldsymbol{W}_{k+1}^{\perp}\|_{\mathrm{op}}^2 + \varepsilon. \tag{37}$$

Since $\boldsymbol{W}_L^{\perp} = 0$, backward induction shows that

$$\|\boldsymbol{W}_k^{\perp}\|_{\mathrm{op}}^2 \le C\varepsilon, \tag{38}$$

for $1 \le k \le L$. Combining $|\sigma_k^2 - \sigma_{k+1}^2| \le \|\boldsymbol{W}_{k+1}^{\perp}\|_{\mathrm{op}}^2 + \varepsilon$ with the bounds on $\|\boldsymbol{W}_k^{\perp}\|_{\mathrm{op}}^2$ above shows that the values $\sigma_1^2, \dots, \sigma_L^2$ differ pairwise by at most $C\varepsilon$. Their geometric mean is σ^2 by Equation (32), so

$$|\sigma_k^2 - \sigma^2| \le C\varepsilon, \tag{39}$$

for $1 \le k \le L$.

If $\varepsilon \le M\sigma^2$, this bound gives

$$\sigma_k^2 \le \sigma^2 + C\varepsilon \le (1 + CM)\sigma^2. \tag{40}$$

Using Equation (32) and applying this upper bound to all factors other than σ_k^2 gives

$$\sigma_k^2 = \frac{\sigma^{2L}}{\prod_{j\neq k}\sigma_j^2} \ge (1 + CM)^{-(L-1)}\sigma^2. \tag{41}$$

These two bounds prove the claimed comparison. □

Proposition A.4. *Fix $M > 0$, and suppose that $r > 0$ and $\varepsilon \le M\sigma^2$. Then there exist constants $c, C > 0$, depending only on L and M, such that*

$$\|\Delta\overline{\boldsymbol{K}}_L\|_{\mathrm{op}} \le C\frac{\varepsilon}{\sigma^2} \tag{42}$$

and

$$c\boldsymbol{I} \preceq \overline{\boldsymbol{K}}_L \preceq C\boldsymbol{I} \tag{43}$$

hold.

Proof. Fix a time satisfying the assumptions of the proposition. By Lemma A.3, for every $1 \le k \le L$, the layer amplitudes are comparable to σ, while $|\sigma_k^2 - \sigma^2| \le C\varepsilon$ and $\|\boldsymbol{W}_k^{\perp}\|_{\mathrm{op}}^2 \le C\varepsilon$. Together with the decomposition in Equation (31), these bounds imply

$$\|\boldsymbol{W}_k\|_{\mathrm{op}}^2 \le C\sigma^2, \tag{44}$$

for $1 \le k \le L$.

For $1 \le j < L$, define

$$\boldsymbol{E}_j := \boldsymbol{W}_{j-1:1}^{\top}(\boldsymbol{W}_j^{\perp})^{\top}\boldsymbol{W}_j^{\perp}\boldsymbol{W}_{j-1:1} \succeq 0. \tag{45}$$

By Lemma A.3 and Equation (44),

$$\|\boldsymbol{E}_j\|_{\mathrm{op}} \le \|\boldsymbol{W}_{j-1:1}\|_{\mathrm{op}}^2\|\boldsymbol{W}_j^{\perp}\|_{\mathrm{op}}^2 \le C\sigma^{2(j-1)}\varepsilon. \tag{46}$$

The backward recursion gives

$$\boldsymbol{W}_{j-1:1}^{\top}\boldsymbol{\rho}_{j-1} = \left(\prod_{k=1}^{j-1}\sigma_k\right)\boldsymbol{\rho}_0 \tag{47}$$

Combining this identity with the orthogonality of the transverse component yields

$$\boldsymbol{W}_{j:1}^{\top}\boldsymbol{W}_{j:1} = \left(\prod_{k=1}^{j}\sigma_k^2\right)\boldsymbol{\rho}_0\boldsymbol{\rho}_0^{\top} + \boldsymbol{E}_j. \tag{48}$$

Substituting Equation (30) and Equation (48) into Equation (16), and then using Equation (32), yields

$$\overline{\boldsymbol{K}}_L = \frac{\sigma^2}{\sigma_1^2}\boldsymbol{I}_d + \sum_{k=2}^{L}\frac{\sigma^2}{\sigma_k^2}\boldsymbol{P}_{\boldsymbol{\theta}} + \sum_{k=2}^{L}\sigma^{-2(L-1)}\left(\prod_{j=k+1}^{L}\sigma_j^2\right)\boldsymbol{E}_{k-1}. \tag{49}$$

For every $1 \le k \le L$, the amplitude bounds imply

$$\left|\frac{\sigma^2}{\sigma_k^2} - 1\right| = \frac{|\sigma^2 - \sigma_k^2|}{\sigma_k^2} \le C\frac{\varepsilon}{\sigma^2}. \tag{50}$$

For every $2 \le k \le L$, Lemma A.3 and Equation (46) give

$$\left\|\sigma^{-2(L-1)}\left(\prod_{j=k+1}^{L}\sigma_j^2\right)\boldsymbol{E}_{k-1}\right\|_{\mathrm{op}} \le C\frac{\varepsilon}{\sigma^2}. \tag{51}$$

Subtracting $\boldsymbol{I}_d + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}}$ from Equation (49), and applying the triangle inequality to the fixed number of terms, proves

$$\|\Delta\overline{\boldsymbol{K}}_L\|_{\mathrm{op}} \le C\frac{\varepsilon}{\sigma^2}. \tag{52}$$

This establishes Equation (42).

Finally, all terms on the right-hand side of Equation (49) are positive semidefinite. Thus the amplitude comparison in Lemma A.3, together with Equation (49), gives $\overline{\boldsymbol{K}}_L \succeq c\boldsymbol{I}$. The same comparison, Equations (49) and (51), and $\varepsilon/\sigma^2 \le M$ give $\overline{\boldsymbol{K}}_L \preceq C\boldsymbol{I}$. □

A.3 Rescaled Predictor Dynamics and Time Reparameterization

We prove Lemma 4.1 and Corollary 4.2 and record the resulting variation of constants formula. Throughout this subsection, we work on an interval on which $r > 0$.

Proof of Lemma 4.1. Recall that $r = \sigma^L$,

$$\boldsymbol{z} = \sigma^{-(L-1)}\boldsymbol{\theta}, \tag{53}$$

and

$$\boldsymbol{J}_L = \sigma^{L-1}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}} = \boldsymbol{I} - \frac{L-1}{L}\boldsymbol{P}_{\boldsymbol{\theta}}. \tag{54}$$

By Equation (28), $\overline{\boldsymbol{K}}_L = \boldsymbol{I} + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}} + \Delta\overline{\boldsymbol{K}}_L$. Since $\boldsymbol{P}_{\boldsymbol{\theta}}^2 = \boldsymbol{P}_{\boldsymbol{\theta}}$, we have $\boldsymbol{J}_L(\boldsymbol{I} + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}}) = \boldsymbol{I}$. Moreover, $\partial \boldsymbol{z}/\partial \boldsymbol{\theta} = \sigma^{-(L-1)}\boldsymbol{J}_L$. By Equations (18) and (20), we have

$$\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{K}_L\boldsymbol{g} = \sigma^{2(L-1)}\overline{\boldsymbol{K}}_L\boldsymbol{g}. \tag{55}$$

We therefore obtain

$$\begin{aligned}\frac{d\boldsymbol{z}}{dt} &= \sigma^{-(L-1)}\boldsymbol{J}_L\frac{d\boldsymbol{\theta}}{dt}\\ &= \sigma^{L-1}\boldsymbol{J}_L\big(\boldsymbol{I} + (L-1)\boldsymbol{P}_{\boldsymbol{\theta}} + \Delta\overline{\boldsymbol{K}}_L\big)\boldsymbol{g}\\ &= \sigma^{L-1}(\boldsymbol{I} + \boldsymbol{J}_L\Delta\overline{\boldsymbol{K}}_L)\boldsymbol{g}\\ &= (\boldsymbol{I} + \boldsymbol{J}_L\Delta\overline{\boldsymbol{K}}_L)\boldsymbol{\Sigma}\big(\sigma^{L-1}\boldsymbol{\beta} - \sigma^{2(L-1)}\boldsymbol{z}\big).\end{aligned} \tag{56}$$

□

Proof of Corollary 4.2. Differentiating Equation (9) gives $d\tau/dt = \sigma^{2(L-1)}$. Combining this identity with Lemma 4.1, we obtain

$$\frac{d\boldsymbol{z}}{d\tau} = (\boldsymbol{I} + \boldsymbol{J}_L\Delta\overline{\boldsymbol{K}}_L)\boldsymbol{\Sigma}\big(\sigma^{-(L-1)}\boldsymbol{\beta} - \boldsymbol{z}\big). \tag{57}$$

Hence,

$$\frac{d\boldsymbol{z}}{d\tau} = -\boldsymbol{\Sigma}\boldsymbol{z} + \sigma^{-(L-1)}\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{f}_K, \tag{58}$$

where the forcing term in the corollary statement is given by

$$\boldsymbol{f}_K = \boldsymbol{J}_L \Delta\overline{\boldsymbol{K}}_L \boldsymbol{\Sigma}\left(\sigma^{-(L-1)}\boldsymbol{\beta} - \boldsymbol{z}\right). \tag{59}$$

□

For any $0 \le \tau_0 < \tau$, the corresponding variation of constants formula is

$$\boldsymbol{z}(\tau) = e^{-(\tau-\tau_0)\boldsymbol{\Sigma}}\boldsymbol{z}(\tau_0) + \int_{\tau_0}^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\sigma(v)^{-(L-1)}\boldsymbol{\Sigma}\boldsymbol{\beta}\, dv + \int_{\tau_0}^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{f}_K(v)\, dv. \tag{60}$$

The perturbation term also satisfies the following pointwise bound. For any fixed $M > 0$, at every time at which $\varepsilon \le M\sigma^2$, Proposition A.4 and $\|\boldsymbol{J}_L\|_{\mathrm{op}} = 1$ imply

$$\|\boldsymbol{f}_K\| \le \sigma^{-(L-1)}\|\boldsymbol{J}_L\|_{\mathrm{op}}\|\Delta\overline{\boldsymbol{K}}_L\|_{\mathrm{op}}\|\boldsymbol{g}\| \le C\varepsilon\sigma^{-(L+1)}\|\boldsymbol{g}\|. \tag{61}$$

A.4 Gaussian initialization

Recall that the initial weight matrices are independent across layers, with i.i.d. entries within each layer distributed as $(\boldsymbol{W}_k(0))_{ij} \sim \mathcal{N}(0, s^2/n_{k-1})$ for $1 \le k \le L$.

Lemma A.5. *Fix L, and suppose that $n_k \asymp d$ for $1 \le k < L$ and $s > 0$. There exist constants $c, C > 0$, independent of d, s, and η_0, such that, for every fixed $\eta_0 > 0$ and all sufficiently large d, the following bounds hold with probability at least $1 - C\exp(-c\min\{1, \eta_0^2\}d)$:*

$$\sigma(0)^2 \asymp s^2, \tag{62}$$

$$\varepsilon \lesssim s^2, \tag{63}$$

$$\max_{1\le k\le L}\|\boldsymbol{W}_k(0)\|_{\mathrm{op}} \lesssim s, \qquad \|\boldsymbol{W}_{L:2}(0)\| \asymp s^{L-1}, \tag{64}$$

and

$$\left|\mathcal{L}(0) - \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2\right| \le C(\eta_0 r(0) + \eta_0^2 r(0)^2). \tag{65}$$

Moreover, on the same event, if $s = 1$ and η_0 is sufficiently small, then

$$\|\boldsymbol{g}(0)\| \ge c. \tag{66}$$

Proof. Write $\boldsymbol{W}_k(0) = s\boldsymbol{G}_k$, where the entries of $\boldsymbol{G}_k$ are i.i.d. $\mathcal{N}(0, 1/n_{k-1})$. Since L is fixed and $n_k \asymp d$ for $k < L$, the standard Gaussian operator-norm bound and a union bound over the layers give

$$\max_{1\le k\le L}\|\boldsymbol{W}_k(0)\|_{\mathrm{op}} \le Cs \tag{67}$$

with probability at least $1 - Ce^{-cd}$. Consequently,

$$\varepsilon \le \max_{1\le k<L}\left(\|\boldsymbol{W}_k(0)\|_{\mathrm{op}}^2 + \|\boldsymbol{W}_{k+1}(0)\|_{\mathrm{op}}^2\right) \le Cs^2, \tag{68}$$

proving Equation (63) and the first bound in Equation (64).

For $1 \le k \le L$, set $q_k := \|\boldsymbol{W}_{L:k}(0)\|$, with $q_{L+1} := 1$. Conditional on $\boldsymbol{W}_{L:k+1}(0)$, the coordinates of $\boldsymbol{W}_{L:k+1}(0)\boldsymbol{W}_k(0)$ are independent centered Gaussians with variance $s^2 q_{k+1}^2/n_{k-1}$. Hence, conditionally,

$$q_k^2 \overset{\mathrm{d}}{=} s^2 q_{k+1}^2 \frac{\chi^2_{n_{k-1}}}{n_{k-1}}. \tag{69}$$

For $1 \le k \le L$, standard chi-square concentration, iterated over the fixed number of layers, yields

$$q_k \asymp s^{L-k+1}, \tag{70}$$

with probability at least $1 - Ce^{-cd}$. In particular, $\|\boldsymbol{W}_{L:2}(0)\| = q_2 \asymp s^{L-1}$ and $r(0) = q_1 \asymp s^L$. Since $\sigma(0)^2 = r(0)^{2/L}$, this proves Equation (62) and the remaining bound in Equation (64).

It remains to control the loss. Conditioned on $\boldsymbol{W}_{L:2}(0)$, the end-to-end predictor is a centered isotropic Gaussian vector:

$$\boldsymbol{\theta}(0) \overset{\mathrm{d}}{=} \frac{s\|\boldsymbol{W}_{L:2}(0)\|}{\sqrt{d}}\boldsymbol{h}, \qquad \boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \tag{71}$$

where $\boldsymbol{h}$ is independent of $\boldsymbol{W}_{L:2}(0)$. Thus, $\widehat{\boldsymbol{\theta}}(0)$ is uniformly distributed on $\mathbb{S}^{d-1}$. Since $a > 0$ and $p > 0$, $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = c_\lambda$, $\|\boldsymbol{\Sigma}\boldsymbol{\beta}\| \lesssim 1$, and $\operatorname{tr}\boldsymbol{\Sigma} = o(d)$. Since $\widehat{\boldsymbol{\theta}}(0) = \boldsymbol{h}/\|\boldsymbol{h}\|$, Gaussian norm concentration gives $\|\boldsymbol{h}\|^2 \geq d/2$ except with probability Ce^{-cd}. The map $\boldsymbol{h} \mapsto \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|$ is $\sqrt{c_\lambda}$-Lipschitz, and $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|] \leq \sqrt{\operatorname{tr}\boldsymbol{\Sigma}}$. Gaussian concentration therefore gives

$$\Pr\left(\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\| > \sqrt{\operatorname{tr}\boldsymbol{\Sigma}} + \eta_0\sqrt{d}\right) \leq \exp\left(-\frac{\eta_0^2 d}{2c_\lambda}\right). \tag{72}$$

For every fixed $\eta_0 > 0$, we have $\operatorname{tr}\boldsymbol{\Sigma} \leq \eta_0^2 d$ for all sufficiently large d. Combining these estimates with Gaussian concentration for $\langle\boldsymbol{\Sigma}\boldsymbol{\beta}, \boldsymbol{h}\rangle$ and h_1, and taking a union bound, gives

$$|\langle\boldsymbol{\Sigma}\boldsymbol{\beta}, \boldsymbol{h}\rangle| \leq C\eta_0\sqrt{d}, \qquad \boldsymbol{h}^\top\boldsymbol{\Sigma}\boldsymbol{h} \leq 4\eta_0^2 d, \qquad |h_1| \leq \eta_0\sqrt{d/2} \tag{73}$$

with failure probability at most $C\exp(-c\min\{1, \eta_0^2\}d)$, for every fixed $\eta_0 > 0$ and all sufficiently large d. Dividing the first and third inequalities by $\|\boldsymbol{h}\|$ and the second by $\|\boldsymbol{h}\|^2$ therefore yields

$$|\langle\boldsymbol{\Sigma}\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}(0)\rangle| \leq C\eta_0, \qquad \widehat{\boldsymbol{\theta}}(0)^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}(0) \leq C\eta_0^2, \qquad |\widehat{\theta}_1(0)| \leq \eta_0, \tag{74}$$

with probability at least $1 - C\exp(-c\min\{1, \eta_0^2\}d)$. It follows that

$$\begin{aligned}\left|\mathcal{L}(0) - \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2\right| &\leq r(0)|\langle\boldsymbol{\Sigma}\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}(0)\rangle| + \frac{1}{2}r(0)^2\widehat{\boldsymbol{\theta}}(0)^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}(0) \\ &\leq C\big(\eta_0 r(0) + \eta_0^2 r(0)^2\big),\end{aligned} \tag{75}$$

which proves Equation (65).

Finally, suppose that $s = 1$. On the intersection of the high-probability events used to establish Equations (70) and (74), we have $r(0) \leq C$ and $|\widehat{\theta}_1(0)| \leq \eta_0$. Since $\lambda_1 = c_\lambda$ and $\beta_1 = c_\beta > 0$,

$$|g_1(0)| = \lambda_1|\beta_1 - r(0)\widehat{\theta}_1(0)| \geq c_\lambda(c_\beta - C\eta_0). \tag{76}$$

Choosing $\eta_0 > 0$ sufficiently small gives $\|\boldsymbol{g}(0)\| \geq |g_1(0)| \geq c$, proving Equation (66). □

B From Initialization to the Onset of the Power-Law Scaling Phase

Throughout this section, fix $\eta_0 > 0$ sufficiently small for all arguments below and work on the high-probability event in Lemma A.5.

B.1 Escape from the Origin under Small Initialization

We now state the precise small-initialization assumption used in the main text.

Assumption B.1 (Small initialization)**.** Let $0 < \delta_0 \leq 1$ be sufficiently small. For any fixed $0 < \delta \leq \delta_0$, assume that d and the initialization scale $s > 0$ satisfy

$$s^2 \leq \begin{cases}\delta, & p < 1, \\ \delta(\log d)^{-(L-1)/L}, & p = 1, \\ \delta d^{-a(p-1)}, & p > 1.\end{cases} \tag{77}$$

The escape estimates below require only a dimension-independent upper bound on s. The full condition in assumption B.1 is first used in the proof of Corollary D.3 to control initialization effects uniformly up to $\tau_+ = c_0 d^a$.

In addition, we define

$$\Phi(\boldsymbol{\theta}) := \langle\boldsymbol{\Sigma}\boldsymbol{\beta}, \boldsymbol{\theta}\rangle - \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 - \mathcal{L}(\boldsymbol{\theta}). \tag{78}$$

Proposition B.2. *There exist fixed constants $s_0 > 0$, $0 < \tau_e < \min\{1, c_\lambda^{-1}\}$, and $c, C > 0$, independent of d and s, such that, for all sufficiently large d, if $0 < s \le s_0$, then for $0 \le \tau \le \tau_e$,*

$$c(s^L + \tau) \le r(\tau) \le C(s^L + \tau), \tag{79}$$

and

$$\mathcal{L}(\tau_e) \le \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 - c. \tag{80}$$

Proof. Let $s_0 > 0$ be chosen sufficiently small below, and suppose that $0 < s \le s_0$. By Equations (62), (63) and (65), $\sigma(0) \asymp s$ and $\varepsilon \lesssim \sigma(0)^2$ for all sufficiently large d. After decreasing s_0 if necessary, we also have $\sigma(0) \le 1$, so Equation (65) gives

$$\Phi(\boldsymbol{\theta}(0)) \ge -C\eta_0\sigma(0)^L. \tag{81}$$

We define

$$u(t) := \int_0^t \sigma(v)^{L-1}\, dv. \tag{82}$$

On any interval on which $\sigma > 0$, we have $du/dt = \sigma^{L-1} > 0$. Hence u is strictly increasing and admits an inverse $t(u)$ on its image. We therefore write $\boldsymbol{\theta}(u) = \boldsymbol{\theta}(t(u))$. Then

$$\frac{d\boldsymbol{z}}{du} = \boldsymbol{J}_L\overline{\boldsymbol{K}}_L\boldsymbol{g}, \tag{83}$$

and

$$\frac{d}{du}\Phi(\boldsymbol{\theta}(u)) = \sigma^{L-1}\boldsymbol{g}^\top\overline{\boldsymbol{K}}_L\boldsymbol{g}. \tag{84}$$

Step 1: Growth in the u-parameterization. We split the argument into two substeps. On $[0, u_1]$, we keep σ comparable to $\sigma(0)$ and prove that $\Phi(\boldsymbol{\theta}(u_1)) > 0$. On $[u_1, u_*]$, we use this positivity to derive a lower bound for $\Phi(\boldsymbol{\theta}(u))^{1/L}$. We first record estimates common to both substeps.

Since $\|\boldsymbol{g} - \boldsymbol{\Sigma}\boldsymbol{\beta}\| \le Cr = C\sigma^L$ and $\|\boldsymbol{\Sigma}\boldsymbol{\beta}\| \asymp 1$, we may fix $\sigma_* > 0$ sufficiently small so that $\|\boldsymbol{g}\| \asymp 1$ whenever $\sigma \le \sigma_*$.

For any fixed $\gamma \in (0, 1]$, as long as $\gamma\sigma(0) \le \sigma \le \sigma_*$, the bound $\varepsilon \lesssim \sigma(0)^2$ implies $\varepsilon \le C_\gamma\sigma^2$. Hence Equation (43) gives

$$\frac{d}{du}\Phi(\boldsymbol{\theta}(u)) \ge c_\gamma\sigma^{L-1}. \tag{85}$$

Moreover, combining Equation (83) with $\|\boldsymbol{z}\| = \sigma$, $\|\boldsymbol{J}_L\|_{\mathrm{op}} = 1$, and the same kernel bound yields

$$\left|\frac{d\sigma}{du}\right| \le \left\|\frac{d\boldsymbol{z}}{du}\right\| \le C_\gamma. \tag{86}$$

Step 1-1: The interval $0 \le u \le u_1$. We first apply the preceding bounds with $\gamma = 1/2$. Choose a fixed $h \in (0, 1]$ such that $C_{1/2}h \le 1/4$, and set $u_1 := h\sigma(0)$. Then, as long as $\frac{1}{2}\sigma(0) \le \sigma \le \sigma_*$,

$$|\sigma(u) - \sigma(0)| \le C_{1/2}u \le C_{1/2}h\sigma(0) \le \frac{1}{4}\sigma(0). \tag{87}$$

This implies

$$\frac{3}{4}\sigma(0) \le \sigma(u) \le \frac{5}{4}\sigma(0). \tag{88}$$

Using $\sigma(0) \lesssim s$, decrease s_0 if necessary so that $5\sigma(0)/4 < \sigma_*$ for all sufficiently large d whenever $0 < s \le s_0$. Thus neither $\sigma = \sigma(0)/2$ nor $\sigma = \sigma_*$ can be reached before u_1. By continuity, the bounds hold throughout $0 \le u \le u_1$.

Integrating Equation (85) over this interval gives

$$\begin{aligned}\Phi(\boldsymbol{\theta}(u_1)) &\ge \Phi(\boldsymbol{\theta}(0)) + c_{1/2}\int_0^{u_1}\sigma(v)^{L-1}\, dv \\ &\ge -C\eta_0\sigma(0)^L + c_{1/2}h\left(\frac{3}{4}\right)^{L-1}\sigma(0)^L.\end{aligned} \tag{89}$$

Since η_0 is sufficiently small,

$$\Phi(\boldsymbol{\theta}(u_1)) \geq c\sigma(0)^L > 0. \tag{90}$$

For every $\boldsymbol{\theta}$,

$$\Phi(\boldsymbol{\theta}) \leq \langle \boldsymbol{\Sigma}\boldsymbol{\beta}, \boldsymbol{\theta} \rangle \leq C\sigma^L. \tag{91}$$

Therefore $\sigma(u_1) \geq c\sigma(0)$. We may thus fix $\gamma_* \in (0, 1/2)$ sufficiently small so that

$$\sigma(u_1) \geq 2\gamma_*\sigma(0). \tag{92}$$

Step 1-2: The interval $u_1 \leq u \leq u_*$. We now apply the preceding bounds with $\gamma = \gamma_*$. As long as $\gamma_*\sigma(0) \leq \sigma \leq \sigma_*$ after u_1, Equation (85) shows that $\Phi(\boldsymbol{\theta}(u))$ remains positive. Moreover, Equations (85) and (91) give

$$\frac{d}{du}\Phi(\boldsymbol{\theta}(u))^{1/L} = \frac{1}{L}\Phi(\boldsymbol{\theta}(u))^{-(L-1)/L}\frac{d}{du}\Phi(\boldsymbol{\theta}(u)) \geq c. \tag{93}$$

Integrating this inequality from u_1 to u gives

$$\begin{aligned} \Phi(\boldsymbol{\theta}(u))^{1/L} &\geq \Phi(\boldsymbol{\theta}(u_1))^{1/L} + c(u - u_1) \\ &\geq c\sigma(0) + c(u - u_1) \\ &\geq c(\sigma(0) + u). \end{aligned} \tag{94}$$

Combining this with Equation (91) gives $\sigma(u) \geq c(\sigma(0) + u)$.

On the other hand, integrating Equation (86) gives

$$\begin{aligned} \sigma(u) &\leq \sigma(u_1) + \int_{u_1}^{u} \left|\frac{d\sigma}{dv}\right| dv \\ &\leq \frac{5}{4}\sigma(0) + C_{\gamma_*}(u - u_1) \\ &\leq C(\sigma(0) + u). \end{aligned} \tag{95}$$

Choose a fixed $u_* > 0$ sufficiently small that $C_{\gamma_*}u_* \leq \sigma_*/4$. After decreasing s_0 if necessary, we have $u_1 < u_*$ and $5\sigma(0)/4 + C_{\gamma_*}u_* < \sigma_*$ for all sufficiently large d whenever $0 < s \leq s_0$. Since $u \mapsto \Phi(\boldsymbol{\theta}(u))$ is nondecreasing, Equations (90) and (91) and the choice of γ_* give $\sigma(u) \geq 2\gamma_*\sigma(0)$. Moreover, $\sigma(u) \leq 5\sigma(0)/4 + C_{\gamma_*}u_* < \sigma_*$. Hence neither boundary $\sigma = \gamma_*\sigma(0)$ nor $\sigma = \sigma_*$ can be reached before u_*. By continuity, the estimates hold throughout $u_1 \leq u \leq u_*$. Together with step 1-1, we obtain

$$\sigma(u) \asymp \sigma(0) + u \tag{96}$$

for $0 \leq u \leq u_*$.

Step 2: Conversion to the τ-parameterization. Since $d\tau/du = \sigma^{L-1} > 0$, the map $u \mapsto \tau(u)$ is strictly increasing on $[0, u_*]$. Combining $d\tau/du = \sigma^{L-1}$ with Equation (96) and integrating from 0 to u yields

$$\tau(u) \asymp (\sigma(0) + u)^L - \sigma(0)^L. \tag{97}$$

Since $u_* > 0$ is fixed, Equation (97) gives $\tau(u_*) \gtrsim u_*^L > 0$. On the other hand, since $u_1 = h\sigma(0)$, the same estimate gives $\tau(u_1) \lesssim \sigma(0)^L \lesssim s^L$. We may therefore fix $0 < \tau_{\mathrm{e}} < \min\{1, c_\lambda^{-1}\}$ sufficiently small and then decrease s_0 if necessary so that $\tau(u_1) < \tau_{\mathrm{e}} < \tau(u_*)$ for all sufficiently large d whenever $0 < s \leq s_0$.

For such d, the monotonicity of $\tau(u)$ implies that, for every $0 \leq \tau \leq \tau_{\mathrm{e}}$, there exists a unique $u \in [0, u_*]$ such that $\tau = \tau(u)$. Combining Equations (96) and (97), we obtain

$$r(\tau) = \sigma(u)^L \asymp (\sigma(0) + u)^L \asymp \sigma(0)^L + \tau \asymp s^L + \tau, \tag{98}$$

which proves Equation (79).

In particular, the unique u corresponding to τ_{e} satisfies $u > u_1$. Applying Equations (94) and (97) at this value of u gives

$$\Phi(\boldsymbol{\theta}(\tau_{\mathrm{e}})) \geq c(\sigma(0)^L + \tau_{\mathrm{e}}) \geq c. \tag{99}$$

The definition of Φ now proves Equation (80).

□

Fix $0 < s_0 < 1$ and τ_e as in Proposition B.2, and let t_e denote the time such that $\tau(t_e) = \tau_e$.

Corollary B.3. *Suppose $0 < s \le s_0$. Then there exist constants $c, c', C > 0$, independent of d, s, and τ, such that, for all sufficiently large d and every $\tau \ge \tau_e$,*

$$\sigma(\tau)^2 \ge c', \tag{100}$$

and

$$c\boldsymbol{I} \preceq \overline{\boldsymbol{K}}_L(\tau) \preceq C\boldsymbol{I}. \tag{101}$$

Proof. For every $t \ge t_e$, monotonicity of the loss and Equation (80) give

$$c' \le \Phi(\boldsymbol{\theta}(t_e)) \le \Phi(\boldsymbol{\theta}(t)) \le \langle \boldsymbol{\Sigma}\boldsymbol{\beta}, \boldsymbol{\theta}(t)\rangle \le C\sigma(t)^L. \tag{102}$$

Hence $\sigma(t)^2 \ge c'$ for every $t \ge t_e$.

Since $\varepsilon \lesssim s^2 \le s_0^2$ and $\sigma(t)^2 \ge c'$, there exists a fixed $M > 0$, independent of d, s, and t, such that $\varepsilon \le M\sigma(t)^2$. Applying Proposition A.4 gives $c\boldsymbol{I} \preceq \overline{\boldsymbol{K}}_L(t) \preceq C\boldsymbol{I}$.

To return to the τ-parameterization, note that $d\tau/dt \ge (c')^{L-1}$, so $\tau(t) \to \infty$. Since τ is strictly increasing and $\tau(t_e) = \tau_e$, it defines a bijection from $[t_e, \infty)$ onto $[\tau_e, \infty)$. Reparameterizing the preceding bounds by τ proves Equations (100) and (101). □

Corollary B.4. *Suppose $0 < s \le s_0$. Then, for all sufficiently large d,*

$$t_e \asymp \begin{cases} 1, & L = 1, \\ \log(1/s), & L = 2, \\ s^{-(L-2)}, & L > 2. \end{cases} \tag{103}$$

Proof. For $0 \le \tau \le \tau_e$,

$$\frac{dt}{d\tau} = \sigma(\tau)^{-2(L-1)} = r(\tau)^{-2(L-1)/L} \asymp (s^L + \tau)^{-2(L-1)/L} \asymp (s^{2L} + \tau^2)^{-(L-1)/L}. \tag{104}$$

Integrating from 0 to τ_e and then setting $\tau = s^L x$, we obtain

$$t_e \asymp \int_0^{\tau_e} (s^{2L} + \tau^2)^{-(L-1)/L}\, d\tau = s^{2-L} \int_0^{\tau_e s^{-L}} (1 + x^2)^{-(L-1)/L}\, dx. \tag{105}$$

For $L = 1$, the time reparameterization satisfies $\tau = t$, so $t_e = \tau_e \asymp 1$. If $L = 2$, then $(L-1)/L = 1/2$, and the integral is comparable to $\log(1/s)$. If $L > 2$, then $(L-1)/L > 1/2$, so the integral is bounded above and below by positive constants for all sufficiently small s. This proves Equation (103). □

To isolate the contribution accumulated up to τ_e, define

$$\boldsymbol{\xi}_e := \boldsymbol{z}(\tau_e) - \int_0^{\tau_e} e^{-(\tau_e - v)\boldsymbol{\Sigma}} \sigma(v)^{-(L-1)} \boldsymbol{\Sigma}\boldsymbol{\beta}\, dv. \tag{106}$$

Corollary B.5. *Suppose $0 < s \le s_0$. Then, for all sufficiently large d,*

$$\|\boldsymbol{\xi}_e\| \le Cs. \tag{107}$$

Proof. By Equation (79) and $\varepsilon \lesssim s^2$, we have $\varepsilon \le C\sigma(\tau)^2$ for $0 \le \tau \le \tau_e$, so Equation (61) applies throughout this interval. Moreover, by the monotonicity of the loss, $\|\boldsymbol{g}\|^2 \le 2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\mathcal{L} \le C$. Therefore,

$$\int_0^{\tau_e} \|\boldsymbol{f}_K(\tau)\|\, d\tau \le C\varepsilon \int_0^{\tau_e} (s^L + \tau)^{-(L+1)/L}\, d\tau \le CL\frac{\varepsilon}{s} \le Cs. \tag{108}$$

Set $\tau_0 = 0$ and $\tau = \tau_e$ in Equation (60), and substitute the resulting expression into Equation (106). This yields

$$\boldsymbol{\xi}_e = e^{-\tau_e\boldsymbol{\Sigma}}\boldsymbol{z}(0) + \int_0^{\tau_e} e^{-(\tau_e - v)\boldsymbol{\Sigma}} \boldsymbol{f}_K(v)\, dv. \tag{109}$$

Using $\|\boldsymbol{z}(0)\| \le \sigma(0) \le Cs$ and Equation (108), we conclude that $\|\boldsymbol{\xi}_e\| \le Cs$. □

Corollary B.6. *Suppose $0 < s \le s_0$, and let d be sufficiently large. For each $\tau \ge \tau_e$, define $\boldsymbol{R}(\tau)$ by*

$$\boldsymbol{R}(\tau) := \sigma(\tau)^{L-1}\left[e^{-(\tau-\tau_e)\boldsymbol{\Sigma}}\boldsymbol{\xi}_e + \int_{\tau_e}^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{f}_K(v)\,dv\right]. \tag{110}$$

Then

$$\boldsymbol{\theta}(\tau) = \sigma(\tau)^{L-1}\int_0^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{\Sigma}\boldsymbol{\beta}\,\sigma(v)^{-(L-1)}\,dv + \boldsymbol{R}(\tau). \tag{111}$$

Proof. Applying Equation (60) with $\tau_0 = \tau_e$, substituting the definition of $\boldsymbol{\xi}_e$ from Equation (106), and using the semigroup property of $e^{-\tau\boldsymbol{\Sigma}}$, we obtain

$$\boldsymbol{z}(\tau) = e^{-(\tau-\tau_e)\boldsymbol{\Sigma}}\boldsymbol{\xi}_e + \int_0^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\sigma(v)^{-(L-1)}\boldsymbol{\Sigma}\boldsymbol{\beta}\,dv + \int_{\tau_e}^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{f}_K(v)\,dv. \tag{112}$$

Multiplying Equation (112) by $\sigma(\tau)^{L-1}$ and using Equation (53) together with the definition of $\boldsymbol{R}$ in Equation (110) gives Equation (111). □

The first term in $\boldsymbol{R}(\tau)$ propagates the initialization and pre-entry perturbation encoded by $\boldsymbol{\xi}_e$, whereas the second accumulates the post-entry kernel perturbation.

B.2 Standard Initialization

For $s = 1$, the initialization estimates give $\sigma(0)^2 \asymp 1$. The next lemma shows that $\sigma(\tau)^2$ remains bounded below by a positive constant and that $\overline{\boldsymbol{K}}_L(\tau)$ remains uniformly comparable to the identity for all $\tau \ge 0$.

Lemma B.7. *Suppose that $s = 1$. Then there exist constants $c, C > 0$, independent of d and τ, such that the following bounds hold for all sufficiently large d and all $\tau \ge 0$:*

$$\sigma(\tau)^2 \ge c, \tag{113}$$

and

$$c\boldsymbol{I} \preceq \overline{\boldsymbol{K}}_L(\tau) \preceq C\boldsymbol{I}. \tag{114}$$

Proof. By Equation (64), $\max_k \|\boldsymbol{W}_k(0)\|_{\mathrm{op}} \le C$ and $\|\boldsymbol{W}_{L:2}(0)\| \ge c$. Moreover, Equations (62) and (66) give $r(0) \ge c$ and $\|\boldsymbol{g}(0)\| \ge c$. Finally, Equations (62) and (65) imply $\mathcal{L}(0) \le \frac{1}{2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 + C\eta_0$. By the Cauchy–Schwarz inequality and Equation (19),

$$\left(\sum_{k=1}^{L}\|\boldsymbol{W}_k(t) - \boldsymbol{W}_k(0)\|_{\mathrm{F}}^2\right)^{1/2} \le \sqrt{t\mathcal{L}(0)} \le C\sqrt{t} \tag{115}$$

for all $t \ge 0$. In particular, $\|\boldsymbol{W}_k(t)\|_{\mathrm{op}} \le C$ for all $1 \le k \le L$ and $0 \le t \le 1$. Thus

$$\begin{aligned} &|r(t) - r(0)| + |\|\boldsymbol{W}_{L:2}(t)\| - \|\boldsymbol{W}_{L:2}(0)\|| + \|\boldsymbol{g}(t) - \boldsymbol{g}(0)\| \\ &\le C\sum_{k=1}^{L}\|\boldsymbol{W}_k(t) - \boldsymbol{W}_k(0)\|_{\mathrm{op}} \le C\sqrt{t}. \end{aligned} \tag{116}$$

Choose a sufficiently small fixed $t_s > 0$. Then

$$\min\{r(t), \|\boldsymbol{W}_{L:2}(t)\|, \|\boldsymbol{g}(t)\|\} \ge c \tag{117}$$

for all $0 \le t \le t_s$. By Equations (19) and (117),

$$\begin{aligned} \frac{d\mathcal{L}}{dt} &= -\boldsymbol{g}(t)^{\top}\boldsymbol{K}_L(t)\boldsymbol{g}(t) \\ &\le -\|\boldsymbol{W}_{L:2}(t)\|^2\|\boldsymbol{g}(t)\|^2 \\ &\le -c. \end{aligned} \tag{118}$$

Since η_0 is sufficiently small, integrating the preceding inequality over $[0, t_s]$ gives

$$\begin{aligned}\mathcal{L}(t_s) &\le \mathcal{L}(0) - ct_s \\ &\le \frac{1}{2}\|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 + C\eta_0 - ct_s \\ &\le \frac{1}{2}\|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 - c.\end{aligned} \tag{119}$$

Since the loss is nonincreasing, for all $t \ge t_s$, we have

$$c \le \frac{1}{2}\|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 - \mathcal{L}(t) \le \langle \mathbf{\Sigma}\boldsymbol{\beta}, \boldsymbol{\theta}(t)\rangle \le Cr(t). \tag{120}$$

Equation (117) covers $0 \le t \le t_s$, whereas Equation (120) applies for $t \ge t_s$. Hence $\sigma(t)^2 \ge c$ for all $t \ge 0$.

Finally, Equation (63) and the preceding lower bound give $\varepsilon/\sigma(t)^2 \le C$. Applying Proposition A.4, we obtain $c\boldsymbol{I} \preceq \overline{\boldsymbol{K}}_L(t) \preceq C\boldsymbol{I}$. Moreover, $\frac{d\tau}{dt} = \sigma(t)^{2(L-1)} \ge c$, so $\tau(t) \to \infty$. Hence these estimates give the stated bounds in the τ-parameterization. □

C Spectral Estimates and Auxiliary Loss Bounds

We first characterize the scaling of $S_d(\tau)$ and estimate the components of the target vector $\boldsymbol{\beta}$ above and below a spectral cutoff. We then use these estimates to bound the loss $\mathcal{L}(\tau)$.

C.1 Spectral Scaling and Tail Estimates

Fix a constant $c_0 > 0$ sufficiently small that $(c_\lambda c_0)^{1/a} < 1/2$, and set

$$\tau_+ := c_0 d^a. \tag{121}$$

Define

$$S_d(\tau) := \sum_{i=1}^{d} \beta_i^2 \min\{\tau\lambda_i, 1\}^2. \tag{122}$$

The function S_d is continuous and nondecreasing.

Lemma C.1. *For all sufficiently large d, we have*

$$S_d(\tau) \asymp \tau^2 \tag{123}$$

for $0 \le \tau \le c_\lambda^{-1}$, *and*

$$S_d(\tau) \asymp \begin{cases} (1+\tau)^{1-p}, & p < 1, \\ 1 + \log(1+\tau), & p = 1, \\ 1, & p > 1 \end{cases} \tag{124}$$

for $c_\lambda^{-1} \le \tau \le \tau_+$.

Proof. By the definitions of β_i and λ_i,

$$S_d(\tau) = c_\beta^2 \sum_{i=1}^{d} i^{-2b} \min\{c_\lambda \tau i^{-a}, 1\}^2. \tag{125}$$

When $0 \le \tau \le c_\lambda^{-1}$, we have $c_\lambda \tau i^{-a} \le 1$ for all i, and since $2a + 2b = 1 + a(1+p) > 1$,

$$S_d(\tau) = c_\lambda^2 c_\beta^2 \tau^2 \sum_{i=1}^{d} i^{-(2a+2b)} \asymp \tau^2. \tag{126}$$

To prove Equation (124), we first consider $2^a c_\lambda^{-1} \le \tau \le \tau_+$ and set $k := \lfloor (c_\lambda \tau)^{1/a} \rfloor$. The remaining fixed interval $[c_\lambda^{-1}, 2^a c_\lambda^{-1}]$ will be handled by monotonicity. The lower bound on τ implies $k \ge 2$, and the choice of c_0 ensures $k < d/2$. Moreover, $k \asymp \tau^{1/a}$. By the definition of k,

$$S_d(\tau) = c_\beta^2 \sum_{i=1}^{k} i^{-2b} + c_\lambda^2 c_\beta^2 \tau^2 \sum_{i=k+1}^{d} i^{-(2a+2b)}. \tag{127}$$

Integral comparisons, together with $k \asymp \tau^{1/a}$ and $(1-2b)/a = 1-p$, give

$$\sum_{i=1}^{k} i^{-2b} \asymp 1 + \int_1^k x^{-2b}\,dx \asymp \begin{cases} (1+\tau)^{1-p}, & p<1, \\ 1+\log(1+\tau), & p=1, \\ 1, & p>1, \end{cases} \tag{128}$$

and

$$\tau^2 \sum_{i=k+1}^{d} i^{-(2a+2b)} \asymp \tau^2 \int_k^d x^{-(2a+2b)}\,dx \asymp \tau^2 k^{1-(2a+2b)} \asymp \tau^{1-p}. \tag{129}$$

Thus, for $2^a c_\lambda^{-1} \le \tau \le \tau_+$,

$$S_d(\tau) \asymp \begin{cases} (1+\tau)^{1-p}, & p<1, \\ 1+\log(1+\tau), & p=1, \\ 1, & p>1. \end{cases} \tag{130}$$

It remains to consider $c_\lambda^{-1} \le \tau \le 2^a c_\lambda^{-1}$. By Equation (123), $S_d(c_\lambda^{-1}) \asymp 1$, while the preceding calculation gives $S_d(2^a c_\lambda^{-1}) \asymp 1$. Since S_d is nondecreasing, it follows that $S_d(\tau) \asymp 1$ throughout this fixed interval. For τ in this interval, each case on the right-hand side of Equation (124) is also comparable to one, which completes the proof. □

For $\mu > 0$, define

$$\boldsymbol{P}_{\le\mu} := \operatorname{diag}\big(\mathbf{1}_{\{\lambda_i \le \mu\}}\big)_{i=1}^{d}, \qquad \boldsymbol{P}_{>\mu} := \boldsymbol{I}_d - \boldsymbol{P}_{\le\mu}. \tag{131}$$

Fix constants $c_\mu > 2^a c_\lambda$ and $0 < \mu_\star \le 2^{-a} c_\lambda$.

Lemma C.2. *For all sufficiently large d and every $\mu > 0$, we have*

$$\big\|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\big\|^2 \lesssim \mu^p. \tag{132}$$

If $c_\mu d^{-a} \le \mu \le \mu_\star$, we also have

$$\big\|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\big\|^2 \asymp \mu^p, \tag{133}$$

and

$$\big\|\boldsymbol{P}_{>\mu}\boldsymbol{\beta}\big\|^2 \asymp \begin{cases} \mu^{-(1-p)}, & p<1, \\ 1+\log(c_\lambda/\mu), & p=1, \\ 1, & p>1. \end{cases} \tag{134}$$

Proof. For $\mu > 0$, define $N_\mu := |\{i : \lambda_i > \mu\}|$. If $\lambda_d \le \mu < c_\lambda$, then $N_\mu + 1 = \lceil (c_\lambda/\mu)^{1/a} \rceil \asymp \mu^{-1/a}$. Since $a + 2b = 1 + ap > 1$,

$$\sum_{\lambda_i \le \mu} \lambda_i \beta_i^2 = c_\lambda c_\beta^2 \sum_{i=N_\mu+1}^{d} i^{-(a+2b)} \lesssim (N_\mu+1)^{-(a+2b-1)} \lesssim \mu^p. \tag{135}$$

If $\mu < \lambda_d$, the left-hand side is zero. If $\mu \ge c_\lambda$, the sum is bounded above by a convergent series. Thus, this upper bound holds for all $\mu > 0$, establishing Equation (132).

Next, suppose $c_\mu d^{-a} \le \mu \le \mu_\star$. For sufficiently large d, we have $2 \le N_\mu + 1 \le d/2$ and $N_\mu + 1 \asymp \mu^{-1/a}$. In particular, $2(N_\mu + 1) \le d$, so

$$\sum_{i=N_\mu+1}^{d} i^{-(a+2b)} \ge \sum_{i=N_\mu+1}^{2(N_\mu+1)} i^{-(a+2b)} \ge (N_\mu+1)\{2(N_\mu+1)\}^{-(a+2b)} \asymp (N_\mu+1)^{-(a+2b-1)}. \tag{136}$$

Therefore, by $N_\mu + 1 \asymp \mu^{-1/a}$ and $(a+2b-1)/a = p$,

$$\sum_{\lambda_i \le \mu} \lambda_i \beta_i^2 = c_\lambda c_\beta^2 \sum_{i=N_\mu+1}^{d} i^{-(a+2b)} \asymp \mu^p. \tag{137}$$

This proves Equation (133).

Finally, under the same bounds on μ, we have $N_\mu \asymp \mu^{-1/a}$ and $1 \le N_\mu \le d/2$. Therefore, since $(1-2b)/a = 1-p$,

$$\|\boldsymbol{P}_{>\mu}\boldsymbol{\beta}\|^2 = c_\beta^2 \sum_{i=1}^{N_\mu} i^{-2b} \asymp \begin{cases} \mu^{-(1-p)}, & p<1, \\ 1+\log(c_\lambda/\mu), & p=1, \\ 1, & p>1. \end{cases} \tag{138}$$

□

C.2 Loss Bounds via Spectral Cutoffs

We now combine the preceding spectral bound for $\boldsymbol{\beta}$ with a bound on $r(\tau)$ to control the low-eigenvalue contribution to the loss. For $\mu > 0$, define

$$\mathcal{L}_{\le\mu}(\tau) := \frac{1}{2}\|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\theta}(\tau))\|^2. \tag{139}$$

The triangle inequality, the bound $\lambda_i \le \mu$ on the range of $\boldsymbol{P}_{\le\mu}$, and $r(\tau) = \|\boldsymbol{\theta}(\tau)\|$ give

$$\begin{aligned} \mathcal{L}_{\le\mu}(\tau) &\le \|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 + \|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}(\tau)\|^2 \\ &\le \|\boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 + \mu r(\tau)^2. \end{aligned} \tag{140}$$

Lemma C.3. *Let $c_K > 0$ and $0 \le \tau_0 \le \tau_1$. Assume that $\overline{\boldsymbol{K}}_L(\tau) \succeq c_K \boldsymbol{I}$ throughout $[\tau_0, \tau_1]$. Then, for any positive continuous function $\mu\colon [\tau_0,\tau_1] \to (0,\infty)$ and every $\tau \in [\tau_0,\tau_1]$,*

$$\mathcal{L}(\tau) \le \exp\!\left(-2c_K\int_{\tau_0}^{\tau}\mu(w)\,dw\right)\mathcal{L}(\tau_0) + 2c_K\int_{\tau_0}^{\tau}\exp\!\left(-2c_K\int_{v}^{\tau}\mu(w)\,dw\right)\mu(v)\mathcal{L}_{\le\mu(v)}(v)\,dv. \tag{141}$$

In particular, for any fixed $\alpha > 0$, setting $\mu(v) = \alpha/\{2c_K(1+v)\}$ gives

$$\mathcal{L}(\tau) \le \left(\frac{1+\tau_0}{1+\tau}\right)^{\alpha}\mathcal{L}(\tau_0) + \alpha(1+\tau)^{-\alpha}\int_{\tau_0}^{\tau}(1+v)^{\alpha-1}\mathcal{L}_{\le\mu(v)}(v)\,dv. \tag{142}$$

Proof. For $\tau_0 \le v \le \tau_1$, the identity $\boldsymbol{g}(v) = \boldsymbol{\Sigma}(\boldsymbol{\beta}-\boldsymbol{\theta}(v))$ gives

$$\begin{aligned} \|\boldsymbol{g}(v)\|^2 &= \|\boldsymbol{\Sigma}(\boldsymbol{\beta}-\boldsymbol{\theta}(v))\|^2 \\ &\ge \mu(v)\|\boldsymbol{P}_{>\mu(v)}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\theta}(v))\|^2 \\ &= 2\mu(v)\left(\mathcal{L}(v) - \mathcal{L}_{\le\mu(v)}(v)\right). \end{aligned} \tag{143}$$

Therefore,

$$\begin{aligned} \frac{d}{dv}\mathcal{L}(v) &= -\boldsymbol{g}(v)^\top\overline{\boldsymbol{K}}_L(v)\boldsymbol{g}(v) \\ &\le -c_K\|\boldsymbol{g}(v)\|^2 \\ &\le -2c_K\mu(v)\left(\mathcal{L}(v) - \mathcal{L}_{\le\mu(v)}(v)\right). \end{aligned} \tag{144}$$

Multiplying the resulting differential inequality by the integrating factor $\exp\left(2c_K\int_{\tau_0}^{v}\mu(w)\,dw\right)$ and integrating from τ_0 to τ yields Equation (141). The polynomial form Equation (142) follows by substituting the specified $\mu(v)$. □

If $\overline{\boldsymbol{K}}_L$ remains uniformly positive definite, the smallest eigenvalue gives the following exponential bound beyond τ_+.

Corollary C.4. *Suppose that there exist constants $\tau_0 \ge 0$ and $c_K > 0$, independent of d, such that, for all sufficiently large d, $\overline{\boldsymbol{K}}_L(\tau) \succeq c_K\boldsymbol{I}$ for every $\tau \ge \tau_0$. Then, for all sufficiently large d and every $\tau \ge \tau_+$,*

$$\mathcal{L}(\tau) \le \mathcal{L}(\tau_+)\exp\!\left(-2c_K c_\lambda d^{-a}(\tau-\tau_+)\right). \tag{145}$$

Proof. By Equation (22) and the assumed lower bound on $\overline{\boldsymbol{K}}_L$, for every $v \geq \tau_0$,

$$\begin{aligned}\frac{d}{dv}\mathcal{L}(v) &= -\boldsymbol{g}(v)^\top \overline{\boldsymbol{K}}_L(v)\boldsymbol{g}(v) \\ &\leq -c_K\|\boldsymbol{g}(v)\|^2 \\ &\leq -c_K\lambda_d\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\theta}(v))\|^2 \\ &= -2c_K\lambda_d\mathcal{L}(v).\end{aligned} \tag{146}$$

For all sufficiently large d, we have $\tau_+ \geq \tau_0$. Applying Grönwall's inequality on $[\tau_+, \tau]$ and using $\lambda_d = c_\lambda d^{-a}$ proves the claim. □

For sufficiently small initialization, Corollary B.3 shows that the preceding corollary applies with $\tau_0 = \tau_e$. Hence the loss decays exponentially beyond $\tau_+ = c_0 d^a$. For $s = 1$, Lemma B.7 shows that it applies with $\tau_0 = 0$, so the same conclusion holds.

D Predictor-Norm Scaling in Reparameterized Time

To treat small and standard initialization in a unified way, we establish predictor-norm and loss bounds under the following conditions. Fix a constant $\tau_\circ > 0$, independent of d and s. For all sufficiently large d, we have $\tau_\circ \leq \tau_+$. Since $\tau_\circ$ is fixed, Lemma C.1 gives $S_d(\tau_\circ) \asymp 1$. Suppose that the initial predictor scale satisfies

$$\sigma(\tau_\circ) \lesssim S_d(\tau_\circ)^{1/(2L)}. \tag{147}$$

Assume also that the initial loss satisfies

$$\mathcal{L}(\tau_\circ) \lesssim \frac{S_d(\tau_\circ)}{1+\tau_\circ}. \tag{148}$$

For $\tau_\circ \leq \tau \leq \tau_+$, we also assume the uniform lower bounds

$$\sigma(\tau) \gtrsim 1 \tag{149}$$

and

$$\overline{\boldsymbol{K}}_L(\tau) \succeq c_K \boldsymbol{I}. \tag{150}$$

These conditions are verified in the proofs of Corollaries D.3 and D.4. The following lemma derives a loss bound from an upper bound on σ.

Lemma D.1. *Let $C_+ \geq 1$ and $\tau_1 \in [\tau_\circ, \tau_+]$. Suppose that Equation (148) holds and that, for $\tau_\circ \leq \tau \leq \tau_1$, $\overline{\boldsymbol{K}}_L(\tau) \succeq c_K \boldsymbol{I}$ and*

$$\sigma(\tau) \leq C_+ S_d(\tau)^{1/(2L)}. \tag{151}$$

Then, for $\tau_\circ \leq \tau \leq \tau_1$,

$$\mathcal{L}(\tau) \leq C(1+C_+^{2L})\frac{S_d(\tau)}{1+\tau}. \tag{152}$$

Here C is independent of C_+ and τ_1.

Proof. For $\tau_\circ \leq v \leq \tau_1$, set

$$\mu(v) := \frac{p+1}{2c_K(1+v)}. \tag{153}$$

Since $\tau_\circ > 0$ is fixed, the monotonicity of S_d and Equations (123) and (124) give, for $\tau_\circ \leq v \leq \tau_1$,

$$(1+v)^{-p} \lesssim \frac{S_d(v)}{1+v}. \tag{154}$$

Using the upper bound in Lemma C.2 together with Equations (140) and (151), we obtain

$$\mathcal{L}_{\leq\mu(v)}(v) \leq C(1+v)^{-p} + CC_+^{2L}\frac{S_d(v)}{1+v} \leq C(1+C_+^{2L})\frac{S_d(v)}{1+v}. \tag{155}$$

Using Equation (142) with $\tau_0 = \tau_\circ$ and $\alpha = p+1$ gives

$$\mathcal{L}(\tau) \le \left(\frac{1+\tau_\circ}{1+\tau}\right)^{p+1} \mathcal{L}(\tau_\circ) + C(1+C_+^{2L})(1+\tau)^{-(p+1)} \int_{\tau_\circ}^{\tau} (1+v)^{p-1} S_d(v)\, dv. \tag{156}$$

By Equation (148) and the monotonicity of S_d,

$$\left(\frac{1+\tau_\circ}{1+\tau}\right)^{p+1} \mathcal{L}(\tau_\circ) \lesssim \left(\frac{1+\tau_\circ}{1+\tau}\right)^{p} \frac{S_d(\tau_\circ)}{1+\tau} \le \frac{S_d(\tau)}{1+\tau}. \tag{157}$$

The same monotonicity and $p > 0$ give

$$(1+\tau)^{-(p+1)} \int_{\tau_\circ}^{\tau} (1+v)^{p-1} S_d(v)\, dv \le S_d(\tau)(1+\tau)^{-(p+1)} \int_{\tau_\circ}^{\tau} (1+v)^{p-1}\, dv \lesssim \frac{S_d(\tau)}{1+\tau}. \tag{158}$$

Combining these bounds proves Equation (152). □

Write $\psi(t) := 1 + \log(1+t)$ and define

$$\psi_p(t) := \begin{cases} 1, & p < 1, \\ \psi(t)^{(L-1)/L}, & p = 1, \\ \psi(t), & p > 1. \end{cases} \tag{159}$$

Proposition D.2. *Suppose that Equations (147) to (150) hold. Assume in addition that, for some fixed $\bar{\varepsilon} > 0$,*

$$\varepsilon \psi_p(\tau_+) \le \bar{\varepsilon}. \tag{160}$$

Then, for $\tau_\circ \le \tau \le \tau_+$,

$$\sigma(\tau) \asymp S_d(\tau)^{1/(2L)}. \tag{161}$$

Moreover, throughout the same interval,

$$\mathcal{L}(\tau) \lesssim \frac{S_d(\tau)}{1+\tau}. \tag{162}$$

In particular,

$$\mathcal{L}(\tau) \lesssim \begin{cases} (1+\tau)^{-p}, & p < 1, \\ \psi(\tau)(1+\tau)^{-1}, & p = 1, \\ (1+\tau)^{-1}, & p > 1. \end{cases} \tag{163}$$

The implicit constants may depend on $\bar{\varepsilon}$, $\tau_\circ$, and the constants in Equations (147) to (150), but are independent of d, s, and τ.

Proof.

Step 1: Upper bound on $\sigma(\tau)$. By Equation (147), choose $A \ge 1$ large enough that $\sigma(\tau_\circ) < AS_d(\tau_\circ)^{1/(2L)}$. We will impose further lower bounds on A below. Suppose for contradiction that $\sigma(\tau) \ge 2AS_d(\tau)^{1/(2L)}$ for some $\tau \in [\tau_\circ, \tau_+]$, and let

$$\tau^\dagger := \min\left\{\tau \in [\tau_\circ, \tau_+] : \sigma(\tau) \ge 2AS_d(\tau)^{1/(2L)}\right\}. \tag{164}$$

Then $\tau^\dagger > \tau_\circ$ and, by continuity, $\sigma(\tau^\dagger) = 2AS_d(\tau^\dagger)^{1/(2L)}$. Define

$$\tau_- := \max\left\{\tau \in [\tau_\circ, \tau^\dagger] : \sigma(\tau) \le AS_d(\tau)^{1/(2L)}\right\}. \tag{165}$$

The defining set is nonempty and closed. By maximality and continuity, $\sigma(\tau_-) = AS_d(\tau_-)^{1/(2L)}$. The definitions of τ_- and $\tau^\dagger$ imply that, for $\tau_- \le \tau \le \tau^\dagger$,

$$AS_d(\tau)^{1/(2L)} \le \sigma(\tau) \le 2AS_d(\tau)^{1/(2L)}. \tag{166}$$

By Equation (160), $\varepsilon \le \bar{\varepsilon}$. Moreover, the monotonicity of S_d and Lemma C.1 give $S_d(\tau) \ge S_d(\tau_\circ) \gtrsim 1$ on this interval. Thus, by increasing A depending only on $\bar{\varepsilon}$ and this fixed lower bound, we have, throughout the same interval,

$$\frac{\varepsilon}{\sigma(\tau)^2} \le \frac{\varepsilon}{A^2 S_d(\tau)^{1/L}} \le 1. \tag{167}$$

Hence Proposition A.4 applies on this interval with $M = 1$, with constants independent of A.

By the definition of $\tau^\dagger$, Equation (151) holds with $C_+ = 2A$ up to this time. Therefore, Lemma D.1 gives, for $\tau_\circ \le \tau \le \tau^\dagger$,

$$\mathcal{L}(\tau) \le CA^{2L}\frac{S_d(\tau)}{1+\tau}. \tag{168}$$

Applying Equation (60) from τ_-, we obtain

$$\sigma(\tau^\dagger) \le \left\|e^{-(\tau^\dagger-\tau_-)\mathbf{\Sigma}}\boldsymbol{z}(\tau_-)\right\| + \left\|\int_{\tau_-}^{\tau^\dagger} e^{-(\tau^\dagger-v)\mathbf{\Sigma}}\sigma(v)^{-(L-1)}\mathbf{\Sigma}\boldsymbol{\beta}\,dv\right\| + \left\|\int_{\tau_-}^{\tau^\dagger} e^{-(\tau^\dagger-v)\mathbf{\Sigma}}\boldsymbol{f}_K(v)\,dv\right\|. \tag{169}$$

We now estimate each term separately.

Step 1-1: The term involving $\boldsymbol{z}(\tau_-)$.

$$\|e^{-(\tau^\dagger-\tau_-)\mathbf{\Sigma}}\boldsymbol{z}(\tau_-)\| \le \sigma(\tau_-) \le AS_d(\tau^\dagger)^{1/(2L)}. \tag{170}$$

Step 1-2: The term involving $\boldsymbol{\beta}$. From Equation (166), for $\tau_- \le \tau \le \tau^\dagger$ we have

$$\sigma(\tau)^{-(L-1)} \le A^{-(L-1)}S_d(\tau)^{-(L-1)/(2L)}. \tag{171}$$

If $\tau^\dagger \le 4$, then $S_d(\tau) \gtrsim 1$ for $\tau_- \le \tau \le \tau^\dagger$. The contraction of the semigroup and $\|\mathbf{\Sigma}\boldsymbol{\beta}\| \lesssim 1$ therefore give

$$\left\|\int_{\tau_-}^{\tau^\dagger} e^{-(\tau^\dagger-v)\mathbf{\Sigma}}\sigma(v)^{-(L-1)}\mathbf{\Sigma}\boldsymbol{\beta}\,dv\right\| \le CA^{-(L-1)}S_d(\tau^\dagger)^{1/(2L)}. \tag{172}$$

Next, we consider the case $\tau^\dagger > 4$.

First, consider $v \in [\max\{\tau_-, \tau^\dagger/2\}, \tau^\dagger]$. In this interval, $v \asymp \tau^\dagger$, so by Equation (124), $S_d(v) \asymp S_d(\tau^\dagger)$. Therefore,

$$\begin{aligned}\left\|\int_{\max\{\tau_-,\tau^\dagger/2\}}^{\tau^\dagger} e^{-(\tau^\dagger-v)\mathbf{\Sigma}}\sigma(v)^{-(L-1)}\mathbf{\Sigma}\boldsymbol{\beta}\,dv\right\|^2 &\le CA^{-2(L-1)}S_d(\tau^\dagger)^{-(L-1)/L}\sum_{i=1}^{d}\beta_i^2\min\{\tau^\dagger\lambda_i, 1\}^2 \\ &\le CA^{-2(L-1)}S_d(\tau^\dagger)^{1/L}.\end{aligned} \tag{173}$$

Next, suppose $\tau_- \le \tau^\dagger/2$ and consider $v \in [\tau_-, \tau^\dagger/2]$. In this range, $e^{-(\tau^\dagger-v)\lambda_i} \le e^{-\tau^\dagger\lambda_i/2}$.

By Equation (124), for $p < 1$,

$$\int_{\tau_\circ}^{\tau^\dagger/2} S_d(v)^{-(L-1)/(2L)}\,dv \lesssim \int_{\tau_\circ}^{\tau^\dagger/2}(1+v)^{-(L-1)(1-p)/(2L)}\,dv \lesssim \tau^\dagger(1+\tau^\dagger)^{-(L-1)(1-p)/(2L)} \tag{174}$$

For $p = 1$,

$$\begin{aligned}\int_{\tau_\circ}^{\tau^\dagger/2} S_d(v)^{-(L-1)/(2L)}\,dv &\lesssim \int_{\tau_\circ}^{\sqrt{\tau^\dagger}}\psi(v)^{-(L-1)/(2L)}\,dv + \int_{\sqrt{\tau^\dagger}}^{\tau^\dagger/2}\psi(v)^{-(L-1)/(2L)}\,dv \\ &\lesssim \sqrt{\tau^\dagger} + \tau^\dagger\psi(\tau^\dagger)^{-(L-1)/(2L)} \\ &\lesssim \tau^\dagger S_d(\tau^\dagger)^{-(L-1)/(2L)}\end{aligned} \tag{175}$$

For $p > 1$,

$$\int_{\tau_\circ}^{\tau^\dagger/2} S_d(v)^{-(L-1)/(2L)}\,dv \lesssim \tau^\dagger \tag{176}$$

Therefore,

$$\int_{\tau_\circ}^{\tau^\dagger/2} S_d(v)^{-(L-1)/(2L)}\,dv \lesssim \tau^\dagger S_d(\tau^\dagger)^{-(L-1)/(2L)}. \tag{177}$$

Moreover,

$$(\tau^\dagger)^2 \sum_{i=1}^{d} \beta_i^2 \lambda_i^2 e^{-\tau^\dagger \lambda_i} \lesssim \sum_{i=1}^{d} \beta_i^2 \min\{\tau^\dagger \lambda_i, 1\}^2 \lesssim S_d(\tau^\dagger). \tag{178}$$

Hence,

$$\begin{aligned}\left\| \int_{\tau_-}^{\tau^\dagger/2} e^{-(\tau^\dagger - v)\boldsymbol{\Sigma}} \sigma(v)^{-(L-1)} \boldsymbol{\Sigma}\boldsymbol{\beta}\, dv \right\|^2 &\le CA^{-2(L-1)} \left(\tau^\dagger S_d(\tau^\dagger)^{-(L-1)/(2L)}\right)^2 \sum_{i=1}^{d} \beta_i^2 \lambda_i^2 e^{-\tau^\dagger \lambda_i} \\ &\le CA^{-2(L-1)} S_d(\tau^\dagger)^{1/L}.\end{aligned} \tag{179}$$

Combining Equations (173) and (179) when $\tau_- \le \tau^\dagger/2$, and using Equation (173) alone otherwise, we obtain

$$\left\| \int_{\tau_-}^{\tau^\dagger} e^{-(\tau^\dagger - v)\boldsymbol{\Sigma}} \sigma(v)^{-(L-1)} \boldsymbol{\Sigma}\boldsymbol{\beta}\, dv \right\| \le CA^{-(L-1)} S_d(\tau^\dagger)^{1/(2L)}. \tag{180}$$

Step 1-3: The term involving $\boldsymbol{f}_K$. By Equations (61) and (167), for $\tau_- \le \tau \le \tau^\dagger$,

$$\|\boldsymbol{f}_K(\tau)\| \le C\varepsilon \sigma(\tau)^{-(L+1)} \|\boldsymbol{g}(\tau)\|. \tag{181}$$

For $j \ge 0$, set $x_j := 2^j(1+\tau_-) - 1$, and let $J := \max\{j \ge 0 : x_j \le \tau^\dagger\}$. For $0 \le j \le J$, Cauchy–Schwarz and Equations (22), (150) and (168) give

$$\begin{aligned}\int_{x_j}^{\min\{x_{j+1}, \tau^\dagger\}} \|\boldsymbol{g}(v)\|\, dv &\le (1+x_j)^{1/2} \left\{ \int_{x_j}^{\min\{x_{j+1}, \tau^\dagger\}} \|\boldsymbol{g}(v)\|^2\, dv \right\}^{1/2} \\ &\le c_K^{-1/2} (1+x_j)^{1/2} \left\{ \mathcal{L}(x_j) - \mathcal{L}(\min\{x_{j+1}, \tau^\dagger\}) \right\}^{1/2} \\ &\le CA^L S_d(x_j)^{1/2}.\end{aligned} \tag{182}$$

Since S_d is nondecreasing, Equations (166) and (181) now imply

$$\begin{aligned}\int_{x_j}^{\min\{x_{j+1}, \tau^\dagger\}} \|\boldsymbol{f}_K(v)\|\, dv &\le C\varepsilon A^{-(L+1)} S_d(x_j)^{-(L+1)/(2L)} \int_{x_j}^{\min\{x_{j+1}, \tau^\dagger\}} \|\boldsymbol{g}(v)\|\, dv \\ &\le C \frac{\varepsilon}{A} S_d(x_j)^{-1/(2L)}.\end{aligned} \tag{183}$$

These intervals cover $[\tau_-, \tau^\dagger]$. We next use Lemma C.1 and $1 + x_j = 2^j(1+\tau_-)$.

For $p < 1$,

$$\begin{aligned}\sum_{j=0}^{J} S_d(x_j)^{-1/(2L)} &\lesssim \sum_{j=0}^{J} (1+x_j)^{-(1-p)/(2L)} \\ &\lesssim \sum_{j=0}^{J} 2^{-j(1-p)/(2L)} \\ &\lesssim 1 \lesssim S_d(\tau^\dagger)^{1/(2L)}.\end{aligned} \tag{184}$$

For $p = 1$, we have $S_d(x_j) \gtrsim 1 + j$ and $J + 1 \lesssim \psi(\tau^\dagger)$. Therefore,

$$\begin{aligned}\sum_{j=0}^{J} S_d(x_j)^{-1/(2L)} &\lesssim \sum_{j=0}^{J} (1+j)^{-1/(2L)} \\ &\lesssim (1+J)^{1-1/(2L)} \\ &\lesssim \psi(\tau^\dagger)^{1-1/(2L)} \\ &\lesssim S_d(\tau^\dagger)^{1/(2L)} \psi(\tau^\dagger)^{(L-1)/L}.\end{aligned} \tag{185}$$

For $p > 1$, we have $S_d(x_j) \asymp 1$ and $J + 1 \lesssim \psi(\tau^\dagger)$. Hence,

$$\sum_{j=0}^{J} S_d(x_j)^{-1/(2L)} \lesssim J + 1 \lesssim \psi(\tau^\dagger) \lesssim S_d(\tau^\dagger)^{1/(2L)} \psi(\tau^\dagger). \tag{186}$$

Combining the three cases gives

$$\sum_{j=0}^{J} S_d(x_j)^{-1/(2L)} \le C S_d(\tau^\dagger)^{1/(2L)} \psi_p(\tau^\dagger). \tag{187}$$

Since ψ_p is nondecreasing and $\tau^\dagger \le \tau_+$, Equation (160) gives $\varepsilon \psi_p(\tau^\dagger) \le \bar{\varepsilon}$. Summing Equation (183) over j and using Equation (187), we obtain

$$\int_{\tau_-}^{\tau^\dagger} \|\boldsymbol{f}_K(v)\|\, dv \le C \frac{\varepsilon}{A} \psi_p(\tau^\dagger) S_d(\tau^\dagger)^{1/(2L)} \le C \frac{\bar{\varepsilon}}{A} S_d(\tau^\dagger)^{1/(2L)}. \tag{188}$$

Step 1-4: Completion of the upper bound. Substituting Equations (170), (180) and (188) into Equation (169) gives

$$\sigma(\tau^\dagger) = 2A S_d(\tau^\dagger)^{1/(2L)} \le \left(1 + CA^{-L} + C\bar{\varepsilon}A^{-2}\right) A S_d(\tau^\dagger)^{1/(2L)}. \tag{189}$$

Fixing A sufficiently large so that $CA^{-L} + C\bar{\varepsilon}A^{-2} < 1/2$ makes the right-hand side strictly smaller than $2AS_d(\tau^\dagger)^{1/(2L)}$, a contradiction. Thus the set defining $\tau^\dagger$ is empty. Hence, for $\tau_\circ \le \tau \le \tau_+$,

$$\sigma(\tau) \le 2A S_d(\tau)^{1/(2L)}. \tag{190}$$

Applying Lemma D.1 with $C_+ = 2A$ and $\tau_1 = \tau_+$ to Equation (190) gives

$$\mathcal{L}(\tau) \le C \frac{S_d(\tau)}{1 + \tau}. \tag{191}$$

This proves Equation (162).

Step 2: Lower bound on $\sigma(\tau)$. Let B be a fixed constant chosen large enough so that $B > 2c_\mu c_0$, and set $\mu(\tau) := B/(1 + \tau)$. Then,

$$\|\boldsymbol{P}_{>\mu(\tau)}(\boldsymbol{\beta} - \boldsymbol{\theta}(\tau))\|^2 \le \frac{2}{\mu(\tau)} \mathcal{L}(\tau) = \frac{2(1+\tau)}{B} \mathcal{L}(\tau) \le \frac{C}{B} S_d(\tau). \tag{192}$$

Here, define

$$\tau_B := \begin{cases} \max\left\{\tau_\circ, c_\lambda^{-1}, \dfrac{B}{\mu_\star} - 1, \left(\dfrac{B}{c_\lambda}\right)^2 - 1\right\}, & p = 1, \\ \max\left\{\tau_\circ, c_\lambda^{-1}, \dfrac{B}{\mu_\star} - 1\right\}, & p \ne 1 \end{cases} \tag{193}$$

For $\tau_B \le \tau \le \tau_+$, the definition of τ_B gives $\mu(\tau) \le \mu_\star$. Moreover, since $\tau \le \tau_+ = c_0 d^a$ and $B > 2c_\mu c_0$, for all sufficiently large d we have $\mu(\tau) \ge B/(1 + \tau_+) \ge c_\mu d^{-a}$. Therefore, Lemma C.2 applies with $\mu(\tau) = B/(1 + \tau)$ and gives

$$\|\boldsymbol{P}_{>\mu(\tau)}\boldsymbol{\beta}\|^2 \gtrsim \begin{cases} B^{-(1-p)}(1+\tau)^{1-p}, & p < 1, \\ 1 + \log \dfrac{c_\lambda(1+\tau)}{B}, & p = 1, \\ 1, & p > 1. \end{cases} \tag{194}$$

In the case $p = 1$, since $\tau \ge \tau_B$, we have $1 + \tau \ge (B/c_\lambda)^2$ and thus

$$1 + \log \frac{c_\lambda(1+\tau)}{B} \ge \frac{1}{2}\{1 + \log(1+\tau)\}. \tag{195}$$

Therefore, by Equation (124),

$$\|\boldsymbol{P}_{>\mu(\tau)}\boldsymbol{\beta}\|^2 \geq \begin{cases} cB^{-(1-p)}S_d(\tau), & p<1, \\ cS_d(\tau), & p \geq 1. \end{cases} \tag{196}$$

Thus, by choosing B sufficiently large,

$$\begin{aligned} \sigma(\tau)^L &= \|\boldsymbol{\theta}(\tau)\| \\ &\geq \|\boldsymbol{P}_{>\mu(\tau)}\boldsymbol{\theta}(\tau)\| \\ &\geq \|\boldsymbol{P}_{>\mu(\tau)}\boldsymbol{\beta}\| - \|\boldsymbol{P}_{>\mu(\tau)}(\boldsymbol{\beta}-\boldsymbol{\theta}(\tau))\| \\ &\geq \begin{cases} (cB^{-(1-p)/2} - CB^{-1/2})S_d(\tau)^{1/2}, & p<1, \\ (c-CB^{-1/2})S_d(\tau)^{1/2}, & p \geq 1 \end{cases} \\ &\geq cS_d(\tau)^{1/2}. \end{aligned} \tag{197}$$

For $p<1$, choose B so that $CB^{-p/2} \leq c/2$; for $p \geq 1$, choose B so that $CB^{-1/2} \leq c/2$. With B thus fixed, the last inequality follows after adjusting c. Hence $\sigma(\tau) \geq cS_d(\tau)^{1/(2L)}$ for $\tau_B \leq \tau \leq \tau_+$.

On the remaining fixed interval $\tau_\circ \leq \tau \leq \min\{\tau_B, \tau_+\}$, the monotonicity of S_d and Equation (124) give $S_d(\tau) \lesssim 1$ because τ_B is fixed, while Equation (149) gives $\sigma(\tau) \gtrsim 1$. After decreasing c if necessary, the lower bound holds throughout $[\tau_\circ, \tau_+]$. Together, this lower bound and Equation (190) prove Equation (161). Finally, combining Equations (124) and (162) proves Equation (163). □

The following corollary gives a precise version of Proposition 4.3.

Corollary D.3 (Predictor-norm scaling under small initialization)**.** *Consider Gaussian initialization satisfying assumption B.1. Then there exist constants $c, C, \tau_0 > 0$, independent of d, s, and τ, such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $\tau_0 \leq \tau \leq \tau_+$,*

$$\|\boldsymbol{\theta}(\tau)\|^2 \asymp \begin{cases} \tau^{1-p}, & p<1, \\ \log\tau, & p=1, \\ 1, & p>1. \end{cases} \tag{198}$$

Proof. Work on the event in Lemma A.5, which has probability at least $1 - Ce^{-cd}$, and set $\tau_\circ = \tau_{\mathrm{e}}$. Equations (79) and (123) give Equation (147), while loss monotonicity and Lemma A.5 give Equation (148). The uniform lower bounds Equations (149) and (150) follow from Corollary B.3. It remains to verify the additional condition Equation (160) of Proposition D.2.

By Lemma A.5, $\varepsilon \lesssim s^2$. If $p<1$, then $\psi_p(\tau) = 1$, so $\varepsilon\psi_p(\tau_+) \lesssim s^2 \lesssim \delta$. If $p=1$, then $\psi_p(\tau_+) \lesssim (\log d)^{(L-1)/L}$, and assumption B.1 gives $\varepsilon\psi_p(\tau_+) \lesssim s^2(\log d)^{(L-1)/L} \lesssim \delta$. If $p>1$, then $\psi_p(\tau_+) \lesssim \log d$. Since $a(p-1) > 0$, for all sufficiently large d, $\log d \lesssim d^{a(p-1)}$, and hence assumption B.1 gives $\varepsilon\psi_p(\tau_+) \lesssim s^2 d^{a(p-1)} \lesssim \delta$. Thus, after choosing the threshold δ_0 in assumption B.1 sufficiently small, all the conditions of Proposition D.2 are satisfied.

Proposition D.2 therefore gives $\sigma(\tau) \asymp S_d(\tau)^{1/(2L)}$. Since $\|\boldsymbol{\theta}(\tau)\| = \sigma(\tau)^L$,

$$\|\boldsymbol{\theta}(\tau)\|^2 = \sigma(\tau)^{2L} \asymp S_d(\tau). \tag{199}$$

Combining this relation with Lemma C.1, and increasing τ_0 by a fixed amount if necessary, proves the corollary.

□

Corollary D.4. *Suppose $p<1$ and consider Gaussian initialization with $s=1$. Then there exist constants $c, C, \tau_0 > 0$, independent of d and τ, such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $\tau_0 \leq \tau \leq \tau_+$,*

$$\|\boldsymbol{\theta}(\tau)\|^2 \asymp \tau^{1-p}. \tag{200}$$

Proof. Work on the event in Lemma A.5, and set $\tau_\circ = c_\lambda^{-1}$. Then Equation (123) gives $S_d(\tau_\circ) \asymp 1$. Loss monotonicity and Lemma A.5 give $\mathcal{L}(v) \lesssim 1$ and hence $\|\boldsymbol{g}(v)\| \lesssim 1$ for $0 \le v \le \tau_\circ$. Thus, Equations (21) and (114) give

$$r(\tau_\circ) \le r(0) + \int_0^{\tau_\circ} \|\overline{\boldsymbol{K}}_L(v)\boldsymbol{g}(v)\|\, dv \lesssim 1. \tag{201}$$

Hence $\sigma(\tau_\circ) \lesssim 1$. Since $\tau_\circ$ is fixed and $S_d(\tau_\circ) \asymp 1$, this estimate verifies Equation (147), while the loss bound above verifies Equation (148). The uniform lower bounds Equations (149) and (150) follow from Lemma B.7.

Since $p < 1$, we have $\psi_p(\tau) = 1$, while Lemma A.5 gives $\varepsilon \lesssim 1$. Hence Equation (160) holds for some fixed $\bar{\varepsilon} > 0$, independent of d. Proposition D.2 therefore yields

$$\|\boldsymbol{\theta}(\tau)\|^2 = \sigma(\tau)^{2L} \asymp S_d(\tau). \tag{202}$$

By Lemma C.1, $S_d(\tau) \asymp (1+\tau)^{1-p}$, and the result follows after choosing a fixed τ_0 sufficiently large. □

E Common Spectral Law for the Loss

E.1 Difficult Tasks

For $p < 1$, we derive the loss scaling from the predictor-norm and loss bounds established in Proposition D.2.

Proposition E.1. *Suppose $p < 1$ and that the conditions of Proposition D.2 hold. Then*

$$\mathcal{L}(\tau) \asymp (1+\tau)^{-p} \tag{203}$$

for $\tau_\circ \le \tau \le \tau_+$.

Proof. From Proposition D.2 and Equation (124), we have

$$S_d(\tau) \asymp (1+\tau)^{1-p} \tag{204}$$

for $\tau_\circ \le \tau \le \tau_+$. Therefore, from Equation (163),

$$\mathcal{L}(\tau) \lesssim (1+\tau)^{-p} \tag{205}$$

holds.

Next, we consider the lower bound. Fix $\mu_0 > 0$ sufficiently small. In particular, assume that $\mu_0/(1+\tau_\circ) \le \mu_\star$. For these values of c_μ and μ_0, reduce c_0 further if necessary so that $2c_\mu c_0 < \mu_0$. Set $\mu(\tau) := \mu_0/(1+\tau)$. For any $\tau_\circ \le \tau \le \tau_+$, we have $\mu(\tau) \le \mu_0/(1+\tau_\circ) \le \mu_\star$. Moreover, for sufficiently large d, $1 + \tau_+ \le 2\tau_+$, so

$$\mu(\tau) \ge \frac{\mu_0}{1+\tau_+} \ge \frac{\mu_0}{2c_0} d^{-a} \ge c_\mu d^{-a}. \tag{206}$$

Therefore, $c_\mu d^{-a} \le \mu(\tau) \le \mu_\star$ holds. The lower bound in Lemma C.2 gives

$$\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| \gtrsim \mu_0^{p/2}(1+\tau)^{-p/2}. \tag{207}$$

From Proposition D.2 and Equation (124), for $\tau_\circ \le \tau \le \tau_+$ it holds that $r(\tau)^2 = \sigma(\tau)^{2L} \lesssim (1+\tau)^{1-p}$, so

$$\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}(\tau)\| \le \mu(\tau)^{1/2} r(\tau) \lesssim \mu_0^{1/2}(1+\tau)^{-p/2}. \tag{208}$$

Consequently,

$$\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}(\tau)\| \lesssim \mu_0^{(1-p)/2}\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|. \tag{209}$$

Since $p < 1$, taking μ_0 sufficiently small gives

$$\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}(\tau)\| \le \frac{1}{2}\|\boldsymbol{P}_{\le\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|. \tag{210}$$

Therefore,

$$\begin{aligned}\|\boldsymbol{P}_{\leq\mu(\tau)}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\theta}(\tau))\| &\geq \|\boldsymbol{P}_{\leq\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| - \|\boldsymbol{P}_{\leq\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}(\tau)\| \\ &\geq \frac{1}{2}\|\boldsymbol{P}_{\leq\mu(\tau)}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| \\ &\gtrsim \mu_0^{p/2}(1+\tau)^{-p/2}.\end{aligned} \tag{211}$$

Hence, for $\tau_\circ \leq \tau \leq \tau_+$,

$$\mathcal{L}(\tau) \geq \frac{1}{2}\|\boldsymbol{P}_{\leq\mu(\tau)}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\theta}(\tau))\|^2 \geq c(1+\tau)^{-p}. \tag{212}$$

□

E.2 Boundary and Easier Tasks

Throughout this subsection, suppose that $p \geq 1$ and that assumption B.1 holds. We choose $0 < \delta_0 \leq s_0^2$ sufficiently small to satisfy the conditions imposed below, and fix an arbitrary $0 < \delta \leq \delta_0$. For all sufficiently large d, we work on the event in Lemma A.5, which has probability at least $1 - Ce^{-cd}$. We use Proposition D.2 with $\tau_\circ = \tau_{\mathrm{e}}$, as established in the proof of Corollary D.3.

Lemma E.2. *Suppose $0 < s \leq s_0$. Then, for all sufficiently large d and every $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$,*

$$\begin{aligned}\boldsymbol{\beta}-\boldsymbol{\theta}(\tau) = &\left(\boldsymbol{I} - \sigma(\tau)^{L-1}\boldsymbol{\Sigma}\int_0^\tau e^{-(\tau-v)\boldsymbol{\Sigma}}\sigma(v)^{-(L-1)}\,dv\right)\boldsymbol{\beta} \\ &- \sigma(\tau)^{L-1}e^{-(\tau-\tau_{\mathrm{e}})\boldsymbol{\Sigma}}\boldsymbol{\xi}_{\mathrm{e}} - \sigma(\tau)^{L-1}\int_{\tau_{\mathrm{e}}}^\tau e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{f}_K(v)\,dv.\end{aligned} \tag{213}$$

Proof. Substitute Equation (110) into Equation (111) and rearrange. □

Lemma E.3. *Suppose $p \geq 1$ and that $\sigma(\tau) \asymp (s^L+\tau)^{1/L}$ for $0 \leq \tau \leq \tau_{\mathrm{e}}$ and $\sigma(\tau) \asymp S_d(\tau)^{1/(2L)}$ for $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$. Then, for every $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$ and every eigenvalue λ_i of $\boldsymbol{\Sigma}$,*

$$1 - C\lambda_i(1+\tau) \leq 1 - \sigma(\tau)^{L-1}\lambda_i\int_0^\tau e^{-(\tau-v)\lambda_i}\sigma(v)^{-(L-1)}\,dv \leq 1. \tag{214}$$

Proof. Note that $e^{-(\tau-v)\lambda_i} \leq 1$ for all $0 \leq v \leq \tau$. Since $\sigma(v) \gtrsim (s^L+v)^{1/L}$ for $0 \leq v \leq \tau_{\mathrm{e}}$,

$$\int_0^{\tau_{\mathrm{e}}}\sigma(v)^{-(L-1)}\,dv \leq C\int_0^{\tau_{\mathrm{e}}}(s^L+v)^{-(L-1)/L}\,dv \leq C. \tag{215}$$

For $p = 1$, combining $\sigma(\tau) \asymp S_d(\tau)^{1/(2L)}$ with Lemma C.1 gives $\sigma(\tau)^{L-1} \asymp \psi(\tau)^{(L-1)/(2L)}$ for $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$.

Also, since $\psi(v) \geq 1$, and for $v \geq \sqrt{1+\tau}-1$,

$$\psi(v) = 1 + \log(1+v) \geq 1 + \frac{1}{2}\log(1+\tau) \geq \frac{1}{2}\psi(\tau), \tag{216}$$

we obtain

$$\begin{aligned}\int_{\tau_{\mathrm{e}}}^\tau \psi(v)^{-(L-1)/(2L)}\,dv &\leq \int_0^{\sqrt{1+\tau}-1}\psi(v)^{-(L-1)/(2L)}\,dv + \int_{\sqrt{1+\tau}-1}^\tau \psi(v)^{-(L-1)/(2L)}\,dv \\ &\leq \sqrt{1+\tau} + C(1+\tau)\psi(\tau)^{-(L-1)/(2L)} \\ &\leq C(1+\tau)\psi(\tau)^{-(L-1)/(2L)}.\end{aligned} \tag{217}$$

Therefore,

$$\begin{aligned}\sigma(\tau)^{L-1}\int_{\tau_{\mathrm{e}}}^\tau \sigma(v)^{-(L-1)}\,dv &\leq C\psi(\tau)^{(L-1)/(2L)}\int_{\tau_{\mathrm{e}}}^\tau \psi(v)^{-(L-1)/(2L)}\,dv \\ &\leq C(1+\tau).\end{aligned} \tag{218}$$

The contribution from $[0, \tau_e]$ is at most $C\sigma(\tau)^{L-1} \le C(1+\tau)$, so the same bound holds over the full interval $[0, \tau]$.

For $p > 1$, using the same argument and the fact that $\sigma(\tau) \asymp 1$, we get

$$\sigma(\tau)^{L-1} \int_0^\tau \sigma(v)^{-(L-1)}\, dv \le C(1+\tau). \tag{219}$$

Therefore,

$$0 \le \sigma(\tau)^{L-1}\lambda_i \int_0^\tau e^{-(\tau-v)\lambda_i}\sigma(v)^{-(L-1)}\, dv \le C\lambda_i(1+\tau). \tag{220}$$

This proves Equation (214). □

Define

$$\overline{\mathcal{L}}_p := \sup_{\tau_e \le \tau \le \tau_+} (1+\tau)^p \mathcal{L}(\tau). \tag{221}$$

Lemma E.4. *Under the assumptions stated at the beginning of this subsection, for $\tau_e \le \tau \le \tau_+$ and $\mu > 0$, we have*

$$\begin{aligned}\left\| \boldsymbol{P}_{\le\mu}\boldsymbol{\Sigma}^{1/2}\sigma(\tau)^{L-1}\left(e^{-(\tau-\tau_e)\boldsymbol{\Sigma}}\boldsymbol{\xi}_e + \int_{\tau_e}^\tau e^{-(\tau-v)\boldsymbol{\Sigma}}\boldsymbol{f}_K(v)\, dv\right)\right\| \\ \le C\mu^{1/2}\left(sS_d(\tau_+)^{(L-1)/(2L)} + \varepsilon\overline{\mathcal{L}}_p^{1/2}S_d(\tau_+)^{(L-1)/L}\right).\end{aligned} \tag{222}$$

Proof. From Proposition D.2, for $\tau_e \le \tau \le \tau_+$,

$$\sigma(\tau)^2 \asymp S_d(\tau)^{1/L} \asymp \begin{cases}\psi(\tau)^{1/L}, & p = 1,\\ 1, & p > 1.\end{cases} \tag{223}$$

Step 1: Upper bound for $\int_{\tau_e}^{\tau_+} \|\boldsymbol{f}_K(v)\|\, dv$ If $L = 1$, then $\varepsilon = 0$ and hence $\boldsymbol{f}_K \equiv \boldsymbol{0}$, so Equation (229) holds immediately. For the remainder of this step, assume $L \ge 2$.

Define $x_j := 2^j\tau_e$ for $j = 0, 1, \ldots$ and let $J := \max\{j \ge 0 : x_j \le \tau_+\}$. Then, we can decompose $[\tau_e, \tau_+]$ as $\bigcup_{j=0}^{J-1}[x_j, 2x_j] \cup [x_J, \tau_+] = \bigcup_{j=0}^{J}[x_j, \min\{2x_j, \tau_+\}]$.

By the Cauchy–Schwarz inequality and the relation to the loss,

$$\begin{aligned}\int_{x_j}^{\min\{2x_j,\tau_+\}} \|\boldsymbol{g}(v)\|\, dv &\le x_j^{1/2}\left\{\int_{x_j}^{\min\{2x_j,\tau_+\}} \|\boldsymbol{g}(v)\|^2\, dv\right\}^{1/2} \\ &\le c_K^{-1/2}x_j^{1/2}\{\mathcal{L}(x_j) - \mathcal{L}(\min\{2x_j,\tau_+\})\}^{1/2} \\ &\le C\overline{\mathcal{L}}_p^{1/2}(1+x_j)^{(1-p)/2}.\end{aligned} \tag{224}$$

From Proposition D.2 and the monotonicity of S_d, for $x_j \le v \le \min\{2x_j, \tau_+\}$ we have

$$\sigma(v)^{-(L+1)} \le CS_d(v)^{-(L+1)/(2L)} \le CS_d(x_j)^{-(L+1)/(2L)}. \tag{225}$$

Therefore,

$$\begin{aligned}\int_{x_j}^{\min\{2x_j,\tau_+\}} \|\boldsymbol{f}_K(v)\|\, dv &\le C\varepsilon \sup_{x_j \le v \le \min\{2x_j,\tau_+\}} \sigma(v)^{-(L+1)} \int_{x_j}^{\min\{2x_j,\tau_+\}} \|\boldsymbol{g}(v)\|\, dv \\ &\le C\varepsilon\overline{\mathcal{L}}_p^{1/2}(1+x_j)^{(1-p)/2}S_d(x_j)^{-(L+1)/(2L)}.\end{aligned} \tag{226}$$

For $p = 1$, we have $(1 + x_j)^{(1-p)/2} = 1$. Using $S_d(x_j) \asymp \psi(x_j) \asymp 1 + j$ and $J + 1 = 1 + \left\lfloor \log_2 \frac{\tau_+}{\tau_e} \right\rfloor \le C\psi(\tau_+)$, we get

$$\begin{aligned}\sum_{j=0}^{J}(1+j)^{-(L+1)/(2L)} &\le 1 + \int_0^J (1+x)^{-(L+1)/(2L)}\,dx \\ &\le C(1+J)^{(L-1)/(2L)} \\ &\le C S_d(\tau_+)^{(L-1)/(2L)}.\end{aligned} \tag{227}$$

For $p > 1$, because $S_d(x_j) \asymp 1$, it follows that

$$\sum_{j=0}^{J}(1+x_j)^{(1-p)/2} S_d(x_j)^{-(L+1)/(2L)} \le C\sum_{j\ge 0} 2^{j(1-p)/2} \le C. \tag{228}$$

Therefore,

$$\int_{\tau_e}^{\tau_+} \|\boldsymbol{f}_K(v)\|\,dv \le C\varepsilon \overline{\mathcal{L}}_p^{1/2} S_d(\tau_+)^{(L-1)/(2L)}. \tag{229}$$

Step 2: Upper bound for the two terms involving $\boldsymbol{\xi}_e$ and $\boldsymbol{f}_K$ From Equation (107) and $\sigma(\tau)^{L-1} \le C S_d(\tau_+)^{(L-1)/(2L)}$, we get

$$\left\| \boldsymbol{P}_{\le\mu} \boldsymbol{\Sigma}^{1/2} \sigma(\tau)^{L-1} e^{-(\tau-\tau_e)\boldsymbol{\Sigma}} \boldsymbol{\xi}_e \right\| \le \mu^{1/2}\sigma(\tau)^{L-1}\|\boldsymbol{\xi}_e\| \le C\mu^{1/2} s S_d(\tau_+)^{(L-1)/(2L)}. \tag{230}$$

From Equation (229), we get

$$\begin{aligned}\left\| \boldsymbol{P}_{\le\mu} \boldsymbol{\Sigma}^{1/2} \sigma(\tau)^{L-1} \int_{\tau_e}^{\tau} e^{-(\tau-v)\boldsymbol{\Sigma}} \boldsymbol{f}_K(v)\,dv \right\| &\le \mu^{1/2}\sigma(\tau)^{L-1} \int_{\tau_e}^{\tau} \|\boldsymbol{f}_K(v)\|\,dv \\ &\le C\mu^{1/2}\varepsilon \overline{\mathcal{L}}_p^{1/2} S_d(\tau_+)^{(L-1)/L}.\end{aligned} \tag{231}$$

□

Proposition E.5. *Suppose $p \ge 1$ and consider Gaussian initialization satisfying assumption B.1. Then there exist constants $c, C > 0$ such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $\tau_e \le \tau \le \tau_+$,*

$$\mathcal{L}(\tau) \asymp (1+\tau)^{-p}. \tag{232}$$

The implicit constants are uniform over $0 < \delta \le \delta_0$.

Proof. Fix $\alpha > p$. Let $\tau_0 \ge \max\{\tau_e, 2\}$ be sufficiently large that $\alpha/(2c_K(1+\tau_0)) \le \mu_\star$. Define $\mu(v) := \alpha/(2c_K(1+v))$ for $v \ge \tau_0$.

Step 1: Upper bound for $\mathcal{L}(\tau)$ For $\tau_0 \le v \le \tau_+$, using Equation (214) and the upper bound in Lemma C.2, we obtain

$$\left\| \boldsymbol{P}_{\le\mu(v)} \boldsymbol{\Sigma}^{1/2} \left[\boldsymbol{I} - \sigma(v)^{L-1}\boldsymbol{\Sigma} \int_0^v e^{-(v-w)\boldsymbol{\Sigma}} \sigma(w)^{-(L-1)}\,dw \right] \boldsymbol{\beta} \right\| \le C\mu(v)^{p/2} \tag{233}$$

Therefore, combining this with Equation (222), we have

$$\begin{aligned}\mathcal{L}_{\le\mu(v)}(v) &= \frac{1}{2}\|\boldsymbol{P}_{\le\mu(v)}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\theta}(v))\|^2 \\ &\le C\left(\mu(v)^p + \mu(v)s^2 S_d(\tau_+)^{(L-1)/L} + \mu(v)\varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right) \\ &\le C\left[(1+v)^{-p} + (1+v)^{-1}\left(s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right)\right].\end{aligned} \tag{234}$$

Applying Equation (142) from Lemma C.3 with initial time τ_0, for $\tau_0 \le \tau \le \tau_+$,

$$\begin{aligned}\mathcal{L}(\tau) &\le \left(\frac{1+\tau_0}{1+\tau}\right)^{\alpha} \mathcal{L}(\tau_0) \\ &\quad + C(1+\tau)^{-\alpha} \int_{\tau_0}^{\tau} (1+v)^{\alpha-p-1}\, dv \\ &\quad + C(1+\tau)^{-\alpha} \int_{\tau_0}^{\tau} (1+v)^{\alpha-2} \left[s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right] dv \\ &\le C(1+\tau)^{-\alpha} + \frac{C}{\alpha-p}(1+\tau)^{-p} \\ &\quad + \frac{C}{\alpha-1}(1+\tau)^{-1}\left(s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right) \\ &\le C\left[(1+\tau)^{-p} + (1+\tau)^{-1}\left(s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right)\right].\end{aligned} \tag{235}$$

For $\tau \in [\tau_e, \tau_0]$, we have $(1+\tau)^p \mathcal{L}(\tau) \le C$. For sufficiently large d, we have $\tau_0 \le \tau_+$. Thus, multiplying both sides of Equation (235) by $(1+\tau)^p$, there exist constants $C_0, C_1 > 0$ such that

$$\overline{\mathcal{L}}_p \le C_0 + C_1\left(s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 \overline{\mathcal{L}}_p S_d(\tau_+)^{2(L-1)/L}\right)(1+\tau_+)^{p-1}. \tag{236}$$

For $p = 1$, we have $S_d(\tau_+) \asymp \log d$, and Equation (77) and $\varepsilon \lesssim s^2$ give

$$\begin{aligned}&s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 S_d(\tau_+)^{2(L-1)/L} \\ &\quad \le C\left[s^2 (\log d)^{(L-1)/L} + \{s^2 (\log d)^{(L-1)/L}\}^2\right] \\ &\quad \le C(\delta + \delta^2)\end{aligned} \tag{237}$$

For $p > 1$, we have $S_d(\tau_+) \asymp 1$. Since Equation (77) gives $s^2 d^{a(p-1)} \le \delta$ and $s^2 \le \delta$,

$$\left(s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 S_d(\tau_+)^{2(L-1)/L}\right)(1+\tau_+)^{p-1} \le C(1+s^2) s^2 d^{a(p-1)} \le C(\delta+\delta^2). \tag{238}$$

We further require δ_0 to be sufficiently small that, for every $0 < \delta \le \delta_0$ and all sufficiently large d,

$$C_1 s^2 S_d(\tau_+)^{(L-1)/L}(1+\tau_+)^{p-1} \le C_0, \qquad C_1 \varepsilon^2 S_d(\tau_+)^{2(L-1)/L}(1+\tau_+)^{p-1} \le \frac{1}{2}. \tag{239}$$

Substituting these bounds into Equation (236), we deduce

$$\overline{\mathcal{L}}_p \le 2C_0 + \frac{1}{2}\overline{\mathcal{L}}_p \tag{240}$$

Hence, $\overline{\mathcal{L}}_p \le 4C_0$. Therefore, by the definition of Equation (221), for $\tau_e \le \tau \le \tau_+$,

$$\mathcal{L}(\tau) \lesssim (1+\tau)^{-p} \tag{241}$$

Step 2: Lower bound of $\mathcal{L}(\tau)$ Fix a sufficiently small constant $\mu_0 > 0$ such that $\mu_0 \le \mu_\star$. If necessary, we decrease c_0 further to ensure that $4c_\mu c_0 < \mu_0$. By step 1, we have $\overline{\mathcal{L}}_p \le C$. For any $\tau_e \le \tau \le \tau_+$, choose $\mu(\tau) \in [\mu_0/(2(1+\tau)),\ \mu_0/(1+\tau)]$. For sufficiently large d, $1+\tau_+ \le 2\tau_+$ holds. Therefore,

$$c_\mu d^{-a} \le \frac{\mu_0}{4c_0} d^{-a} \le \frac{\mu_0}{2(1+\tau_+)} \le \frac{\mu_0}{2(1+\tau)} \le \mu(\tau) \le \frac{\mu_0}{1+\tau} \le \mu_0 \le \mu_\star \tag{242}$$

holds.

For every coordinate i such that $\lambda_i(1+\tau) \le \mu_0$, Lemma E.3 shows that the coefficient multiplying β_i in the first term on the right-hand side of Equation (213) is at least $1 - C\mu_0$. In particular, if we

take μ_0 sufficiently small, we can guarantee $1 - C\mu_0 \geq 1/2$. Therefore, together with the lower bound in Lemma C.2,

$$\begin{aligned}\left\| \boldsymbol{P}_{\leq \mu(\tau)} \boldsymbol{\Sigma}^{1/2} \left[\boldsymbol{I} - \sigma(\tau)^{L-1} \boldsymbol{\Sigma} \int_0^\tau e^{-(\tau-v)\boldsymbol{\Sigma}} \sigma(v)^{-(L-1)}\, dv \right] \boldsymbol{\beta} \right\| &\geq \frac{1}{2} \|\boldsymbol{P}_{\leq \mu(\tau)} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}\| \\ &\geq c\mu(\tau)^{p/2} \\ &\gtrsim (1+\tau)^{-p/2}.\end{aligned} \tag{243}$$

On the other hand, by Lemma E.4,

$$\begin{aligned}&\left\| \boldsymbol{P}_{\leq \mu(\tau)} \boldsymbol{\Sigma}^{1/2} \sigma(\tau)^{L-1} \left(e^{-(\tau-\tau_{\mathrm{e}})\boldsymbol{\Sigma}} \boldsymbol{\xi}_{\mathrm{e}} + \int_{\tau_{\mathrm{e}}}^\tau e^{-(\tau-v)\boldsymbol{\Sigma}} \boldsymbol{f}_K(v)\, dv \right) \right\| \\ &\leq C\mu(\tau)^{1/2} \left(s S_d(\tau_+)^{(L-1)/(2L)} + \varepsilon S_d(\tau_+)^{(L-1)/L} \right) \\ &\leq C\mu_0^{-(p-1)/2} \mu(\tau)^{p/2} (1+\tau_+)^{(p-1)/2} \left(s S_d(\tau_+)^{(L-1)/(2L)} + \varepsilon S_d(\tau_+)^{(L-1)/L} \right)\end{aligned} \tag{244}$$

and, as in the arguments for Equations (237) and (238),

$$\begin{aligned}&(1+\tau_+)^{(p-1)/2} \left\{ s S_d(\tau_+)^{(L-1)/(2L)} + \varepsilon S_d(\tau_+)^{(L-1)/L} \right\} \\ &\leq \sqrt{2} \left[\left\{ s^2 S_d(\tau_+)^{(L-1)/L} + \varepsilon^2 S_d(\tau_+)^{2(L-1)/L} \right\} (1+\tau_+)^{p-1} \right]^{1/2} \\ &\leq C(\delta + \delta^2)^{1/2}.\end{aligned} \tag{245}$$

Finally, since the spectral cutoff μ_0 has already been fixed, we require δ_0 to be sufficiently small that

$$C\mu_0^{-(p-1)/2} (\delta_0 + \delta_0^2)^{1/2} \leq \frac{c}{2}, \tag{246}$$

where c is the constant in the lower bound for the first term above. Therefore, for every $0 < \delta \leq \delta_0$ and all sufficiently large d, Equation (213) gives

$$\|\boldsymbol{P}_{\leq \mu(\tau)} \boldsymbol{\Sigma}^{1/2} (\boldsymbol{\beta} - \boldsymbol{\theta}(\tau))\| \geq c\mu(\tau)^{p/2}. \tag{247}$$

Therefore, for $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$,

$$\begin{aligned}\mathcal{L}(\tau) &\geq \mathcal{L}_{\leq \mu(\tau)}(\tau) \\ &= \frac{1}{2} \|\boldsymbol{P}_{\leq \mu(\tau)} \boldsymbol{\Sigma}^{1/2} (\boldsymbol{\beta} - \boldsymbol{\theta}(\tau))\|^2 \\ &\geq c\mu(\tau)^p \\ &\gtrsim (1+\tau)^{-p}.\end{aligned} \tag{248}$$

Step 3: Combining the upper and lower bounds Combining Equations (241) and (248) yields the desired conclusion.

□

E.3 Common Spectral Law under Small Initialization

Combining the cases $p < 1$ and $p \geq 1$ gives the common spectral loss law under small initialization. The following proposition gives a precise version of Proposition 4.4.

Proposition E.6 (Common spectral law under small initialization)**.** *Consider Gaussian initialization satisfying assumption B.1. Then there exist constants $c, C > 0$, independent of d, s, and τ, such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$,*

$$\mathcal{L}(\tau) \asymp (1+\tau)^{-p} \tag{249}$$

Proof. If $p < 1$, the proof of Corollary D.3 verifies that, under assumption B.1, the conditions of Proposition D.2 hold with $\tau_\circ = \tau_{\mathrm{e}}$. Proposition E.1 therefore gives $\mathcal{L}(\tau) \asymp (1+\tau)^{-p}$ for $\tau_{\mathrm{e}} \leq \tau \leq \tau_+$. If $p \geq 1$, the same conclusion follows directly from Proposition E.5. Combining the two cases proves the claim. □

F TRAINING-TIME LOSS SCALING

Using the predictor-norm estimates in Appendix D, we relate reparameterized time τ to training time t. We then express the spectral loss laws in Propositions E.1 and E.6 in terms of training time.

The following theorem gives a precise version of Theorem 4.5.

Theorem F.1 (Training-time loss scaling under small initialization)**.** *Consider Gaussian initialization satisfying assumption B.1. Let t_e be defined by $\tau(t_e) = \tau_e$, set $\bar{t} = t - t_e$, and let t_+ be defined by $\tau(t_+) = \tau_+$, with $\bar{t}_+ := t_+ - t_e$. Then there exist constants $c, C, \bar{t}_0 > 0$, independent of d, s, and $\bar{t}$, such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $\bar{t}_0 \le \bar{t} \le \bar{t}_+$, we have*

$$\mathcal{L}(t_e + \bar{t}) \asymp \begin{cases} \bar{t}^{-pL/[1+p(L-1)]}, & p < 1, \\ \bar{t}^{-1}(\log \bar{t})^{-(L-1)/L}, & p = 1, \\ \bar{t}^{-p}, & p > 1. \end{cases} \tag{250}$$

Proof. By Equation (9), $d\tau/dt = \sigma^{2(L-1)}$. Since $\bar{t} = t - t_e$ and $\|\boldsymbol{\theta}\| = \sigma^L$, we obtain

$$\frac{d\bar{t}}{d\tau} = \sigma^{-2(L-1)} = \|\boldsymbol{\theta}(\tau)\|^{-2(L-1)/L}. \tag{251}$$

Equivalently,

$$\bar{t} = \int_{\tau_e}^{\tau} \|\boldsymbol{\theta}(v)\|^{-2(L-1)/L}\, dv. \tag{252}$$

Choose a sufficiently large fixed constant $\tau_0 > \tau_e$ such that the bounds in Corollary D.3 hold for all $\tau_0 \le \tau \le \tau_+$. By Corollary B.3, we get $\sigma(\tau)^2 \ge c > 0$ for $\tau \ge \tau_e$. Therefore

$$\int_{\tau_e}^{\tau_0} \|\boldsymbol{\theta}(v)\|^{-2(L-1)/L}\, dv = \int_{\tau_e}^{\tau_0} \sigma(v)^{-2(L-1)}\, dv = O(1). \tag{253}$$

Therefore

$$\bar{t} = O(1) + \int_{\tau_0}^{\tau} \|\boldsymbol{\theta}(v)\|^{-2(L-1)/L}\, dv. \tag{254}$$

Since the $O(1)$ term is uniform in d and s, we may choose the fixed constant $\bar{t}_0 > 0$ sufficiently large so that, for $\bar{t} \ge \bar{t}_0$, this term can be absorbed into the comparisons below.

Combining Corollary D.3 with Equation (254), we obtain, for $\bar{t}_0 \le \bar{t} \le \bar{t}_+$,

$$\bar{t} \asymp \begin{cases} \displaystyle\int_{\tau_0}^{\tau} v^{-(1-p)(L-1)/L}\, dv, & p < 1, \\ \displaystyle\int_{\tau_0}^{\tau} (\log v)^{-(L-1)/L}\, dv, & p = 1, \\ \displaystyle\int_{\tau_0}^{\tau} dv, & p > 1. \end{cases} \tag{255}$$

Evaluating these integrals gives

$$\bar{t} \asymp \begin{cases} \tau^{[1+p(L-1)]/L}, & p < 1, \\ \tau(\log \tau)^{-(L-1)/L}, & p = 1, \\ \tau, & p > 1. \end{cases} \tag{256}$$

For $p = 1$, the estimate $\int_{\tau_0}^{\tau} (\log v)^{-(L-1)/L}\, dv \asymp \tau(\log \tau)^{-(L-1)/L}$ follows from the same splitting argument used in Equation (217). The relations in Equation (256) can be inverted as

$$\tau \asymp \begin{cases} \bar{t}^{L/[1+p(L-1)]}, & p < 1, \\ \bar{t}(\log \bar{t})^{(L-1)/L}, & p = 1, \\ \bar{t}, & p > 1. \end{cases} \tag{257}$$

In the case $p = 1$, we used the fact that $\bar{t} \asymp \tau(\log \tau)^{-(L-1)/L}$ implies $\log \bar{t} \asymp \log \tau$. Finally, Proposition E.6 gives $\mathcal{L}(\tau) \asymp \tau^{-p}$ on this interval. Substituting Equation (257) yields

$$\mathcal{L}(t_e + \bar{t}) \asymp \begin{cases} \bar{t}^{-pL/[1+p(L-1)]}, & p < 1, \\ \bar{t}^{-1}(\log \bar{t})^{-(L-1)/L}, & p = 1, \\ \bar{t}^{-p}, & p > 1. \end{cases} \tag{258}$$

□

The following corollary gives a precise version of Corollary 5.1.

Corollary F.2 (Training-time loss scaling under standard initialization for difficult tasks)**.** *Suppose $p < 1$ and consider Gaussian initialization with $s = 1$. Let t_+ be defined by $\tau(t_+) = \tau_+$. Then there exist constants $c, C, t_0 > 0$, independent of d and t, such that, for all sufficiently large d, with probability at least $1 - Ce^{-cd}$, uniformly for $t_0 \le t \le t_+$, we have $\mathcal{L}(t) \asymp t^{-pL/[1+p(L-1)]}$.*

Proof. Let $\tau_0 > 0$ be the fixed constant in Corollary D.4. Then $\|\boldsymbol{\theta}(\tau)\|^2 \asymp \tau^{1-p}$ uniformly for $\tau_0 \le \tau \le \tau_+$, while Proposition E.1 gives $\mathcal{L}(\tau) \asymp \tau^{-p}$. By Equation (9), $d\tau/dt = \sigma^{2(L-1)}$, and $\|\boldsymbol{\theta}\| = \sigma^L$.

By Lemma B.7, the training time elapsed over $0 \le \tau \le \tau_0$ is bounded uniformly in d. Choosing t_0 sufficiently large, we therefore obtain, uniformly for $t_0 \le t \le t_+$,

$$t = O(1) + \int_{\tau_0}^{\tau} \|\boldsymbol{\theta}(v)\|^{-2(L-1)/L}\, dv \asymp \int_{\tau_0}^{\tau} v^{-(1-p)(L-1)/L}\, dv \asymp \tau^{[1+p(L-1)]/L}. \tag{259}$$

Therefore,

$$\tau \asymp t^{L/[1+p(L-1)]} \tag{260}$$

and substituting this into $\mathcal{L}(\tau) \asymp \tau^{-p}$ gives

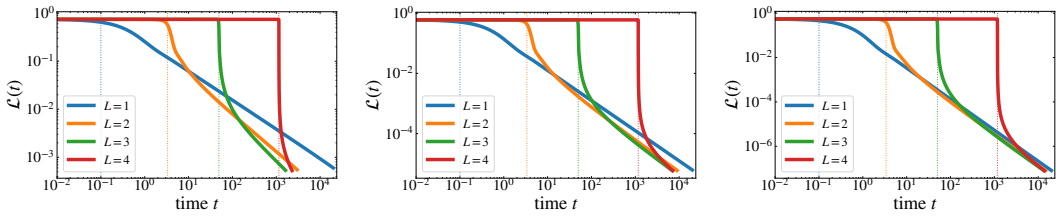
$$\mathcal{L}(t) \asymp t^{-pL/[1+p(L-1)]}. \tag{261}$$

□

G Experimental Details

G.1 Synthetic Power-Law Experiments

Figures 1 to 3. These three figures use the same trajectories. We set $a = 2$ and $c_\lambda = c_\beta = 1$, take $d = 2048$, and use $p \in \{0.6, 1.0, 1.4\}$, or equivalently $b \in \{0.1, 0.5, 0.9\}$. We consider $L \in \{1, 2, 3, 4\}$, with every hidden layer having width d, and initialize each layer with scale $s = d^{-1/2}$. We minimize the exact population loss by gradient descent with a learning rate of 0.01 and accumulate τ by the trapezoidal rule. Each curve is from a single random initialization. We set the entrance time by $\tau_e = 0.1$ and use the corresponding interpolated value of t_e in the shifted time $\bar{t} = t - t_e$. The corresponding loss curves without a time shift are shown in Figure 8.



(a) $p = 0.6$ (b) $p = 1.0$ (c) $p = 1.4$

Figure 8: The same trajectories as in Figure 1 are shown in training time t, without a time shift. Vertical dotted lines mark t_e, where $\tau(t_e) = 0.1$, for $L = 1, 2, 3, 4$.

Figure 4. We use the same task, dimension, widths, initialization scale, and optimization settings as above, now with $p \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$ and five random seeds for each $L \in \{1, 2, 3, 4\}$. For $p \neq 1$, over $50 \leq \tau \leq 5000$, we fit $\log \mathcal{L} = \log A - q \log(\bar{t} - B)$ by least squares over A, B, and q, with $B < \min \bar{t}$. At $p = 1$, we interpolate each trajectory at logarithmically spaced values of τ over the same window and fit $\log \mathcal{L}(\bar{t}) + \frac{L-1}{L} \log \log \bar{t} = \log A - q \log(\bar{t} - B)$ by least squares over A, B, and q, again with $B < \min \bar{t}$. The markers and error bars show the mean and sample standard deviation of q over the five seeds. Since even the largest standard deviation across seeds is approximately 0.004, the error bars are smaller than the markers.

Figure 5. We set $a = 2.25$, $b = -0.175$, $c_\lambda = c_\beta = 1$, and hence $p = 0.4$, and use $L \in \{1, 2, 3, 4\}$ with hidden width d and standard initialization $s = 1$. We use $d \in \{128, 256, 512, 1024, 2048, 4096, 8192\}$; panel (a) displays $d = 8192$, whereas panel (b) uses the full dimension sweep. For each pair (L, d), we use five random seeds for initialization. All runs use gradient descent with a learning rate of 0.002. To select a fitting window before the finite-dimensional spectral cutoff, we use the decoupled-mode proxy

$$\mathcal{L}_{\text{proxy}}(\tau) := \frac{1}{2} \sum_{i \geq 1} \lambda_i \beta_i^2 e^{-2\lambda_i \tau}. \tag{262}$$

This is the loss obtained by approximating each residual mode as $\beta_i - \theta_i(\tau) = \beta_i e^{-\lambda_i \tau}$. Since $\lambda_i \beta_i^2 = c_\lambda c_\beta^2 i^{-1-ap}$, the fraction of this infinite-dimensional proxy carried by the modes $i > d$, which are absent from the dimension-d experiment, is approximated by

$$r_{\text{tail}}(\tau; p, d) := \frac{\sum_{i>d} \lambda_i \beta_i^2 e^{-2\lambda_i \tau}}{\sum_{i \geq 1} \lambda_i \beta_i^2 e^{-2\lambda_i \tau}} \approx \frac{\int_d^\infty x^{-1-ap} e^{-2c_\lambda \tau x^{-a}} \, dx}{\int_0^\infty x^{-1-ap} e^{-2c_\lambda \tau x^{-a}} \, dx} = P\left(p, \frac{2c_\lambda \tau}{d^a}\right). \tag{263}$$

Here, for $\alpha > 0$ and $x \geq 0$, the regularized lower incomplete gamma function is

$$P(\alpha, x) = \frac{\int_0^x u^{\alpha-1} e^{-u} \, du}{\int_0^\infty u^{\alpha-1} e^{-u} \, du}, \tag{264}$$

and $P^{-1}(\alpha, \cdot)$ denotes its inverse with respect to the second argument. Setting $r_{\text{tail}} = 0.1$ and using $a = 2.25$, $p = 0.4$, and $c_\lambda = 1$ gives $\tau_{\max} = d^{2.25} P^{-1}(0.4, 0.1)/2$, and we set $\tau_{\min} = \max\{1, \tau_{\max}/1000\}$. Each run stops at the first iterate with $\tau \geq \tau_{\max}$. For each seed, we linearly interpolate t and $\mathcal{L}$ in τ at logarithmically spaced points in this window and fit $\log \mathcal{L} = \log A - q \log t$ by ordinary least squares, without a time shift. In panel (a), the theoretical lines have fixed exponents $pL/[1 + p(L-1)]$ and heights fitted in log space using the fitting points from all five seeds. Panel (b) reports the mean and sample standard deviation of the fitted exponents over the five seeds.

G.2 Extensions beyond the theoretical setting

Figure 6. We use all $N = 5{,}002{,}240$ examples in the five parts of CIFAR-5M. Each $32 \times 32 \times 3$ RGB image is scaled by $1/255$, flattened, and centered by the empirical mean, giving $d = 3072$. We likewise center the airplane-versus-rest target and compute the full empirical covariance $\boldsymbol{\Sigma}$ and input–target cross moment $\boldsymbol{e}$. We rescale the inputs so that $\|\boldsymbol{\Sigma}\|_{\text{op}} = 1$ and rescale the target direction so that $\boldsymbol{e}^\top \boldsymbol{\Sigma}^\dagger \boldsymbol{e} = 1$. We report the excess empirical squared loss relative to the best linear predictor, equivalently the squared loss against the corresponding realizable linear teacher. In the covariance eigenbasis, this teacher has coefficients $\beta_i = e_i/\lambda_i$. Motivated by Wei et al. (2022), who study eigenvalue decay and target alignment in neural representations, we estimate the effective task exponent p_{eff} by fitting

$$\sum_{i>k} \lambda_i \beta_i^2 \propto \lambda_k^{p_{\text{eff}}} \tag{265}$$

using ordinary least squares in log–log coordinates. The fit uses the unique integer ranks obtained by rounding logarithmically spaced values from 10 to $\text{round}(0.9d) = 2765$. The upper limit omits approximately the last 10% of ranks from the fit to reduce finite-dimensional endpoint effects. This gives $p_{\text{eff}} \approx 0.478$, as shown in Figure 9. We train networks with $L \in \{1, 2, 3, 4\}$ and hidden width 3072 from random initialization at scale $s = 1$, using full-batch gradient descent on these empirical sufficient statistics. All runs use a learning rate of 0.002 and continue to $\tau = 10^6$. Each curve is from a single random initialization.

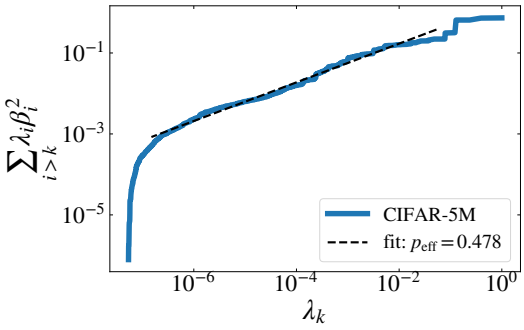


Figure 9: The cumulative sum $\sum_{i>k} \lambda_i \beta_i^2$ versus λ_k for the CIFAR-5M task. The dashed line is a power-law fit with $p_{\text{eff}} \approx 0.478$, using $10 \le k \le \text{round}(0.9d) = 2765$.

Figure 7. We draw fresh Gaussian inputs from the synthetic task with $a = 2$, $c_\lambda = c_\beta = 1$, and $d = 2048$, and generate noiseless targets $y = \boldsymbol{\beta}^\top \boldsymbol{x}$. We use $b = 0.1$ for $p = 0.6$ and $b = 0.9$ for $p = 1.4$. The bias-free MLPs have $L \in \{2, 3, 4\}$ trainable linear maps, hidden width 2048, ReLU activations after the first $L - 1$ maps, and a scalar linear output. The hidden maps use standard He initialization and the output map uses standard fan-in initialization, both at scale one. Each curve is from a single random initialization. We train by online SGD for $100{,}000$ steps with fresh batches of size 4096, a learning rate of 0.01, and neither momentum nor weight decay. We evaluate the loss on a fixed set of $65{,}536$ independent samples.