LEARNING RESHAPES POWER-LAW ANISOTROPY IN INTERNAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Power-law anisotropy in internal representations has been observed across a wide range of biological and artificial neural systems, from state-of-the-art language models to the mouse cerebral cortex. This anisotropy is a key geometric property of high-dimensional information processing and underlies a variety of theoretical analyses. However, the mechanism by which it emerges from input structure and task-driven learning has remained unclear. Here, we characterize this formation process by exactly solving the learning dynamics of a wide two-layer linear neural network in a teacher–student setting with power-law input and teacher structures. We show that, in the feature-learning regime, the local power-law exponent of the internal-representation spectrum evolves nonmonotonically over the course of training and exhibits up to four distinct asymptotic regimes across modes and training times. By contrast, in the lazy regime, the exponent remains essentially unchanged. We further demonstrate numerically that similar exponent dynamics arise in more realistic nonlinear networks. Together, these results suggest a general mechanism by which the dynamic interaction between input statistics and task structure gives rise to power-law internal representations.

1 INTRODUCTION

Neural networks transform inputs from the external world into high-dimensional internal representations. Understanding the structure of these representations is essential for explaining how networks extract information and form features relevant to a task. One fundamental quantity characterizing this geometry is the eigenspectrum of the covariance matrix, or kernel, of the internal representations.

Recent studies have reported that the eigenvalues of internal-representation covariance matrices decay as a power law with rank in a wide range of neural systems Gauthaman et al. (2025); Morales-Gregorio et al. (2024); Pachitariu et al. (2026); Morales et al. (2023); Nakamuta et al. (2025), including mouse visual cortex Stringer et al. (2019), computer-vision models Ghosh et al. (2022), and large language models Li et al. (2026). Such power-law anisotropy implies that a large fraction of the variance is concentrated along a small number of directions, while information remains distributed across many higher-order directions. In addition to being observed broadly in prior work, power-law anisotropy has been linked to the efficiency and robustness of information coding Stringer et al. (2019); Tatsukawa & Teramae (2025); Safavi et al. (2024); Nassar et al. (2020), generalization performance Ghosh et al. (2022), and to neural scaling laws in training Bordelon et al. (2024a); Maloney et al. (2022); Kramp et al. (2026). Power-law representation spectra may therefore provide an important connection between the microscopic geometry of internal representations and macroscopic performance.

Most existing theories, however, treat the eigenspectrum of the internal representation as a given quantity that remains fixed during learning. In actual neural networks, the representation itself changes through training. Merely assuming a fixed power-law exponent therefore cannot explain where the power-law spectrum originates or how learning reorganizes it. In particular, because the eigenspectrum of the input covariance of natural data often itself exhibits power-law decay Field (1987); Ruderman & Bialek (1993); Stringer et al. (2019); Corral et al. (2015); Font-Clos et al. (2013), one must distinguish whether an observed internal representation simply inherits the input structure or is reshaped by task learning (Fig. 1). Although several previous studies have assumed

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

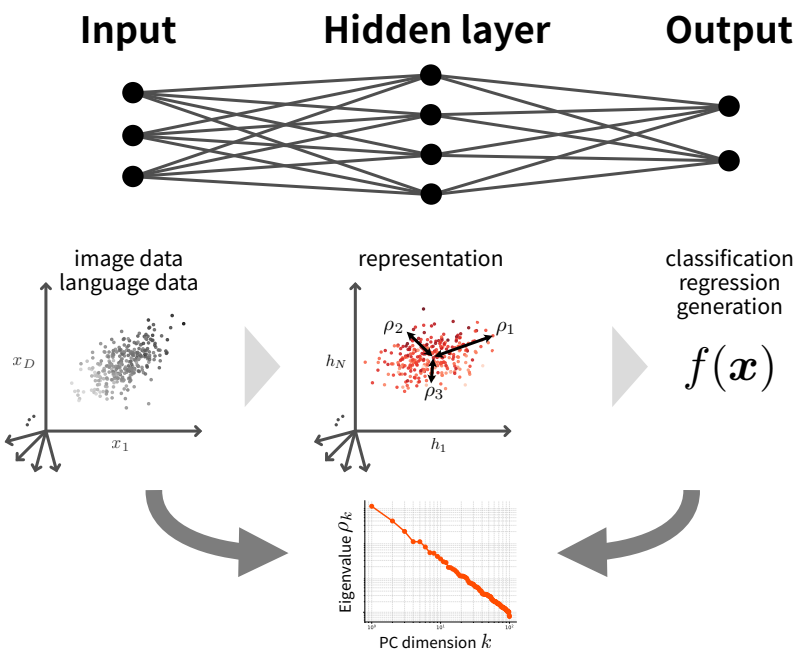


Figure 1: Conceptual overview of this study. Previous works have extensively investigated how power-law anisotropy of internal representations affects network outputs and training dynamics, whereas the mechanism by which such representations form has remained insufficiently understood. Here, we clarify how input structure and learning jointly shape these representations.

power-law input spectra and derived the dynamics of internal representations Bordelon et al. (2025); Wortsman & Loureiro (2025), to the best of our knowledge, none of them has characterized how the power-law exponent which quantifies the degree of spectral anisotropy evolves during learning. It also remains unclear whether the spectrum retains a single power-law exponent across all modes during learning or instead exhibits multiple local exponents over different mode ranges, and how such changes depend on the learning regime.

1.1 Our contributions

To characterize how learning reshapes the eigenspectrum of internal representation, we study a wide two-layer linear network with aligned power-law input and teacher structures characterized by exponents α and β, respectively. Our main contributions are as follows:

1. **Exact solution for the time evolution of representation spectra.** We derive an exact analytical expression for the time evolution of the internal-representation eigenspectrum (Result 3.1, Eq. 12). This expression explicitly decomposes each eigenvalue into an input-derived component present at initialization and a teacher-dependent component formed through learning.
2. **Representation spectra exhibit four asymptotic exponent regions during learning.** We analytically characterize the power-law exponent, which quantifies spectral anisotropy (Result 3.2 , Eq. 15, 16). In the feature-learning regime, the exponent can evolve nonmonotonically, increasing from α toward $3\alpha+\beta$ before decreasing toward either $\alpha+\beta$ or $\alpha+\beta/2$. By contrast, in the lazy regime, the learning-induced component remains small relative to the initialization component, and the exponent stays close to that of the input.
3. **Finite-width nonlinear networks validate the theory.** We validate the theory in finite-width nonlinear networks and observe similar learning-induced changes, confirming that the theoretically predicted spectral transitions are observed beyond the linear model (Result 3.3).

Oveall, our results show that power-law internal representations do not merely reflect the statistical structure of the input data, but are dynamically reshaped by the anisotropy of the teacher signal and by mode-dependent learning timescales. They thus provide an analytical starting point for understanding the power-law internal representations observed across neural networks as a consequence of the interaction between input structure and learning.

1.2 Related Works

Power-law spectra of internal representations Power-law covariance eigenspectra have been observed in biological neural populations: in mouse V1, the eigenvalues decay approximately as k^{-1}, with related scaling reported across species and brain regions Stringer et al. (2019); Kong et al. (2022); Meshulam et al. (2019); Morales et al. (2023); Wang et al. (2025). A recent reanalysis of the mouse V1 data found that a broken power law provides a better fit than a single power law Pospisil & Pillow (2025), supporting a local, mode-dependent description of the exponent. Analogous power-law feature-covariance spectra and their evolution during training have been empirically observed in vision models Agrawal et al. (2022); Ghosh et al. (2022) and large language models Li et al. (2026).

Input structure and spectral theories of fixed representations Power-law decay of covariance spectra has long been reported for natural images as well Field (1987); Ruderman & Bialek (1993). However, the persistence of a power-law spectrum in V1 responses even when the input images are spatially whitened indicates that internal representations cannot be explained solely by direct inheritance of the input covariance Stringer et al. (2019). In random-feature models, the input power-law exponent has been shown to carry over to the feature-covariance spectrum up to logarithmic corrections Paquette et al. (2026). Theories based on fixed kernels or fixed features have used eigenspectra and target alignment to derive learning order, generalization curves, and neural scaling laws Kramp et al. (2026); Bordelon et al. (2024a; 2025); Maloney et al. (2022). Beyond random-feature models, however, it remains unknown how training deforms an inherited input exponent when the internal kernel evolves over time.

Learning dynamics of deep linear networks Although deep linear networks implement linear input–output maps, the gradient dynamics of their factorized weights are nonlinear, and these networks have therefore been studied as analytically tractable feature-learning models. Exact learning curves for aligned singular modes have been derived, revealing plateaus and stage-like learning Saxe et al. (2013). The exact temporal evolution of the hidden representational similarity matrix and the neural tangent kernel was subsequently derived as well Braun et al. (2022); Domine et al. (2025). Building on these results, we derive the hidden-representation covariance in the population limit for power-law input and teacher structures that are simultaneously diagonalizable, with particular emphasis on the power-law decay of its eigenspectrum.

2 Setting

We consider a mean-field parameterization of a two-layer linear neural network in a teacher-student setting with input dimension D, hidden width N, and output dimension P:

$$\boldsymbol{f}(\boldsymbol{x}, t) = \frac{1}{\gamma_0 N} \sum_{i=1}^{N} \boldsymbol{a}_i(t)\, \boldsymbol{w}_i(t)^\top \boldsymbol{x}, \tag{1}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{\Theta}^\top \boldsymbol{x}. \tag{2}$$

Here, $\boldsymbol{y}(\boldsymbol{x})$ denotes teacher output, and $\boldsymbol{\Theta}$ is teacher matrix. The dimensions of the variables are $\boldsymbol{x}, \boldsymbol{w}_i \in \mathbb{R}^D$, $\boldsymbol{a}_i, \boldsymbol{f}, \boldsymbol{y} \in \mathbb{R}^P$, and $\boldsymbol{\Theta} \in \mathbb{R}^{D\times P}$. We assume

$$P = D \ll N, \qquad N \to \infty. \tag{3}$$

Thus, the input and output dimensions take the same finite value, while the hidden-layer width is taken to infinity. Although conventional regression and classification problems typically have $P \ll D$, the choice $P \geq D$ is natural for architectures such as large language models, diffusion models, and autoencoders. As will become clear below, $P \geq D$ is also required for the spectrum to evolve over a broad range of modes. We therefore focus on the simplest such setting, $P = D$.

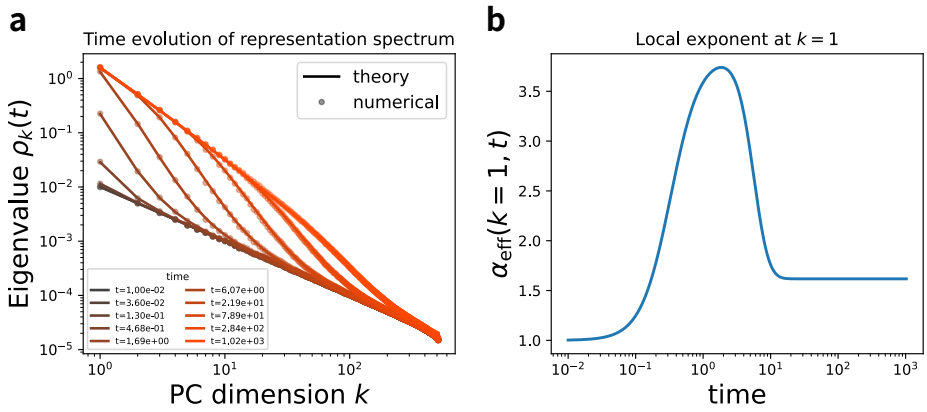


Figure 2: Evolution of the internal representations of a two-layer linear model whose power-law exponent changes during training. *(a)* Eigenvalue spectra of the internal representation at different training times. Solid lines denote the analytical solution, markers denote the numerical solution, and color indicates training time. *(b)* Local power-law exponent at each time, evaluated at $k = 1$; equivalently, this is the slope of the spectrum on logarithmic axes.

Following previous work Bordelon & Pehlevan (2022); Bordelon et al. (2024c;b), the parameter γ_0 controls the learning regime. Although this deep linear model can represent only linear input–output maps, its learning dynamics are nonlinear and exhibit a variety of rich phenomena, making it an important object of study in machine-learning theory.

We assume that the input data $\boldsymbol{x}$ and the teacher matrix $\boldsymbol{\Theta}$ have the following aligned structure: the input covariance matrix and the teacher Gram matrix are diagonal in the same orthonormal basis, and their eigenvalues decay as power laws with exponents α and β, respectively:

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \tag{4}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), \tag{5}$$

$$\lambda_k \sim k^{-\alpha}, \tag{6}$$

$$\boldsymbol{\Theta}\boldsymbol{\Theta}^\top = \mathrm{diag}(\mu_1, \ldots, \mu_D), \tag{7}$$

$$\mu_k \sim k^{-\beta}. \tag{8}$$

The weights $\boldsymbol{w}_i(t)$ and $\boldsymbol{a}_i(t)$ are initialized independently from Gaussian distributions with variances σ_w^2 and σ_a^2, respectively, and are trained by gradient flow on the mean-squared-error loss with learning rate ν. Although the temporal evolution of the parameters and loss is known from previous work Saxe et al. (2013), neither the power-law anisotropy of the internal representation nor the evolution of its exponent has been studied. We therefore analytically derive the temporal evolution of the eigenvalues of the internal-representation covariance in the infinite-data limit and characterize the corresponding power-law exponents.

3 Results

3.1 Exact solution for the time evolution of representation spectra

We first derive the time evolution of the internal-representation eigenvalue spectrum in the two-layer linear network. The internal-representation kernel is defined as

$$\Phi_t(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{G}^w(t) \boldsymbol{x}', \tag{9}$$

where

$$\boldsymbol{G}^w(t) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{w}_i(t) \boldsymbol{w}_i(t)^\top \tag{10}$$

is the Gram matrix of the first-layer weights. Although our object of interest is the spectrum of Φ_t, in the population limit its nonzero eigenvalues coincide with those of

$$\boldsymbol{M}(t) = \boldsymbol{\Lambda}^{1/2} \boldsymbol{G}^w(t) \boldsymbol{\Lambda}^{1/2} \tag{11}$$

(Appendix A). We therefore analyze the eigenvalue spectrum $\rho_k(t)$ of $\boldsymbol{M}(t)$.

In our setting, because the input covariance and the teacher Gram matrix are diagonalized by the same basis, the eigenvalues ρ_k evolve independently mode by mode and are given exactly by

$$\begin{aligned}
\rho_k(t) &= \underbrace{\sigma_w^2 \lambda_k}_{\substack{\text{input-derived} \\ \text{initialization component}}} + \underbrace{\sigma_w^2 \lambda_k R(u_k(t))}_{\substack{\text{teacher-derived component} \\ \text{formed through learning}}}, \\
R(u_k(t)) &= \frac{S}{2\sigma_w^2} \left[-1 + \sqrt{1 + 4u_k(t)^2} \right], \\
u_k(t) &= \frac{v_k(t)}{1 - v_k(t)^2}, \\
v_k(t) &= \frac{r_k^+ \big(E_k(t) - 1\big)}{(r_k^+)^2 + E_k(t)}, \\
E_k(t) &= \exp\left(\frac{\lambda_k \sqrt{S^2 + 4\gamma_0^2 \theta_k^2}}{\gamma_0^2} t \right), \\
r_k^+ &= \frac{-S + \sqrt{S^2 + 4\gamma_0^2 \theta_k^2}}{2\gamma_0 \theta_k},
\end{aligned} \tag{12}$$

where $S = \sigma_w^2 + \sigma_a^2$ and $\theta_k = \sqrt{\mu_k}$ (Appendix D). This derivation uses the wide-network limit $P = D \ll N$, with $N \to \infty$. Throughout the evolution described above, $\rho_k(t)$ remains monotonically decreasing in k; hence, the relative ordering of the modes is preserved at all times (Appendix F). The input-mode index k can therefore be identified directly with the rank index of the spectrum.

Equation 12 shows that the spectrum consists of two components. The first is inherited from the input data and is already present at initialization, whereas the second is induced by the teacher signal through learning. Thus, the anisotropy of the internal representation evolves not only according to the anisotropy of the input distribution but also according to the anisotropy of the teacher signal and the learning dynamics.

The eigenspectrum dynamics predicted by the analytical solution above closely match those obtained numerically for finite-width networks (Fig. 2(a), Appendix G). At initialization, the spectrum $\rho_k(0)$ has the same power-law exponent α as the input variance λ_k. As training proceeds, the modes acquire the teacher-derived component in order from low to high mode index. The evolution of the spectrum demonstrates that the local effective power-law exponent $\alpha_{\text{eff}}(k, t)$ can change during training (Fig. 2(b). We define this exponent formally and characterize its temporal evolution in the next subsection.

3.2 Representation spectra exhibit four asymptotic exponent domains during learning

We next investigate how the power-law exponent of the eigenvalue spectrum, which characterizes the anisotropy of the internal representation, changes during learning. After taking the infinite-width limit $N \to \infty$, we consider sufficiently large input and output dimensions D and P, and assume that, over the mode range $1 \le k \le D$, the input variance and the mode strength of the teacher matrix obey

$$\lambda_k = \lambda_0 k^{-\alpha}, \qquad \mu_k = \theta_k^2 = \mu_0 k^{-\beta}. \tag{13}$$

We define the local effective power-law exponent of the representation spectrum, which depends on time t and mode k, by

$$\alpha_{\text{eff}}(k, t) := -\frac{\partial \log \rho_k(t)}{\partial \log k}. \tag{14}$$

Here, the derivative is defined after continuously interpolating the discrete sequence of modes.

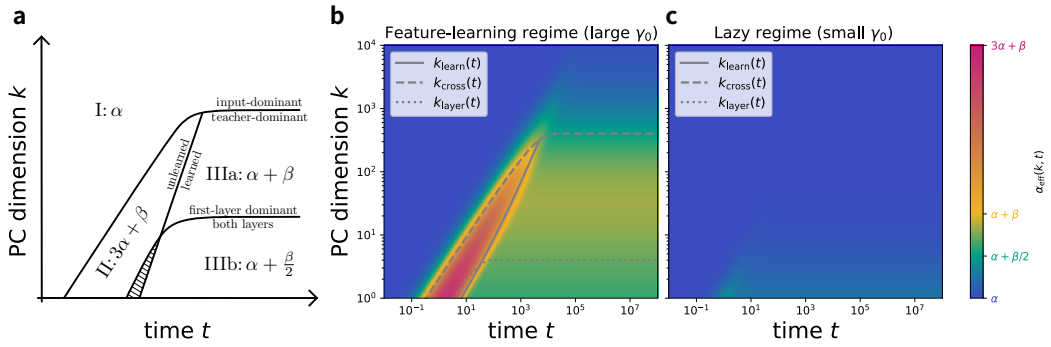


Figure 3: Evolution of the local power-law exponent α_{eff}, defined as the negative slope of the eigenspectrum on log–log axes. In all panels, the horizontal axis represents training time and the vertical axis represents PC dimension (mode rank). In the heatmaps, color indicates the local exponent $\alpha_{\text{eff}}(k,t)$ around each mode k at each training time t. *(a)* Schematic of the plateaus and their boundaries in the local exponent around each mode during training. The hatched regions exhibit decay that is not described by a single power law. *(b)* In the feature-learning regime, the internal-representation spectrum exhibits as many as four asymptotic exponents. *(c)* In the lazy regime, the exponent of the internal representation changes very little.

For the setting considered here, the local exponent $\alpha_{\text{eff}}(k,t)$ takes different values in distinct domains of the (k,t) plane (Fig. 3). We therefore first determine the boundaries separating these domains and then derive the asymptotic value of the local power-law exponent within each domain.

Substituting the analytical expression for the eigenspectrum derived in Result 3.1 (Eq. 12) into the definition of the power-law exponent yields the following exact expression (Appendix E):

$$\begin{cases} \alpha_{\text{eff}}(k,t) = \alpha + \dfrac{R(u_k(t))}{1+R(u_k(t))}\left(1+\dfrac{1}{g(u_k(t))}\right)p(k,t) \\ p(k,t) = \dfrac{(u_{k,\infty}-u_k(t))g(u_k(t))}{u_k(t)}\displaystyle\int_0^{u_k(t)}\left[\frac{\alpha}{u_{k,\infty}-z}+\frac{u_{k,\infty}\beta/2}{(u_{k,\infty}-z)^2}\right]\frac{dz}{g(z)} \end{cases} \tag{15}$$

where $g(x) := \sqrt{1+4x^2}$, $u_{k,\infty} := \gamma_0\theta_k/S$.

The form of the exact solution above indicates that the main changes in the power-law exponent occur when $R(u_k(t)) = 1$, $u_k(t) \to u_{k,\infty}$, and $u_k(t) = 1$. Accordingly, we define the boundaries between the asymptotic domains as follows (Appendix E):

- $k_{\text{cross}}(t)$ is defined by $R(u_k(t)) = 1$, or equivalently $u_k(t) = u_{\text{cross}} := \sqrt{\sigma_w^2(S+\sigma_w^2)}/S$, separating the input-dominated region from the teacher-dominated region;
- $k_{\text{learn}}(t)$ is defined by $u_k(t) = 0.9u_{k,\infty}$, separating the unlearned region from the learned region; and
- $k_{\text{layer}}(t)$ is defined by $u_k(t) = 1$, separating the region in which only the first layer learns from that in which both layers learn.

Near these boundaries, where $u_k(t)$ is close to the corresponding threshold, the spectrum is generally not described by a single power law. Sufficiently far from the boundaries, however, the local exponent approaches a constant asymptotic value. The regions are characterized as follows (Fig. 3 and Appendix E).

- **Region I: Initialization-dominated.** Above $k_{\text{cross}}(t)$, the learning-induced change in the internal representation is smaller than the initialization component. The exponent approaches α in this region. This domain contains not only unlearned modes but also modes for which the teacher signal is so weak that the initialization component remains dominant

even after learning. In particular, all sufficiently large k belong to this domain. Consequently, changes in the local power-law exponent are observed only over a finite range of modes.

- **Region II: Transient steepening during early learning.** Below $k_{\mathrm{cross}}(t)$ and above $k_{\mathrm{learn}}(t)$, with $u_k(t) \ll 1$, modes have only recently begun to learn, yet the learning-induced component already exceeds the initialization component. The exponent rises toward $3\alpha + \beta$. By contrast, in a domain where $u_k(t) \gg 1$, the eigenvalue spectrum $\rho_k(t)$ then has no single power-law exponent.
- **Region IIIa: Learned modes with first-layer-dominated learning.** Below both $k_{\mathrm{cross}}(t)$ and $k_{\mathrm{learn}}(t)$, but above $k_{\mathrm{layer}}(t)$, the exponent approaches $\alpha + \beta$. A necessary condition for this asymptotic region to exist is that the initialization scale of the second layer be much larger than that of the first, $\sigma_a \gg \sigma_w$ (Appendix E.4). The analytical solution in Appendix D shows that learning in this region substantially changes the first layer relative to its initialization. Region IIIa therefore corresponds to learning primarily in the first layer.
- **Region IIIb: Learned modes with learning in both layers.** Within the learned, teacher-dominated region, modes below $k_{\mathrm{layer}}(t)$ approach an exponent of $\alpha + \beta/2$. In contrast to Region IIIa, both layers learn in this region.

Collecting the results above, the internal-representation spectrum in the (k, t) plane exhibits as many as four asymptotic exponents in bulk domains sufficiently far from the boundaries (Fig. 3(a, b)):

$$\alpha_{\mathrm{eff}}(k,t) \simeq \begin{cases} \alpha, & u_k(t) \ll u_{\mathrm{cross}}, \\ 3\alpha+\beta, & u_k(t) \gg u_{\mathrm{cross}}, \quad u_k(t) \ll u_{k,\infty}, \quad u_k(t) \ll 1, \\ \alpha+\beta, & u_k(t) \gg u_{\mathrm{cross}}, \quad u_k(t) \simeq u_{k,\infty}, \quad u_k(t) \ll 1, \\ \alpha+\dfrac{\beta}{2}, & u_k(t) \gg u_{\mathrm{cross}}, \quad u_k(t) \simeq u_{k,\infty}, \quad u_k(t) \gg 1. \end{cases} \tag{16}$$

When a fixed mode k is followed over time, its location relative to k_{layer} determines which of the following sequences it exhibits:

$$\alpha \longrightarrow 3\alpha+\beta \longrightarrow \begin{cases} \alpha+\beta, \\ \alpha+\dfrac{\beta}{2}. \end{cases} \tag{17}$$

Whether all four regions exist depends on the hyperparameters. In particular, in the feature-learning regime characterized by $\sigma_a \gg \sigma_w$ and $\gamma_0 \gg (\sigma_a^2 + \sigma_w^2)/\theta_0$ and for sufficiently larger D, all four regions can coexist (see Appendix E). Consequently, the local effective exponent first steepens transiently and then decreases. This nonmonotonic change reflects the mode-dependent learning rates set by the input eigenvalues and the process by which the anisotropy of the teacher signal is incorporated into the internal representation.

By contrast, in the lazy regime satisfying $\gamma_0 \ll \sqrt{\sigma_w^2(2\sigma_w^2 + \sigma_a^2)}/\theta_0$, the learned component does not exceed the initialization component even for the strongest teacher mode. Hence $R(u_k(t)) \ll 1$ is maintained for every mode, and

$$\rho_k(t) \simeq \sigma_w^2 \lambda_k, \qquad \alpha_{\mathrm{eff}}(k,t) \simeq \alpha \tag{18}$$

(Fig. 3(c)). The nonmonotonic exponent evolution and the multiple bulk exponents derived above are therefore specific to the feature-learning regime, in which the internal representation changes substantially from its initial value.

3.3 Similar exponent dynamics appear in nonlinear networks

Finally, to test whether the exponent changes found in the solvable model also arise in more general models, we perform a numerical study using nonlinear networks. We train a network with a nonlinear activation function in a teacher–student setting with the same type of power-law input variance and teacher signal, and measure the temporal evolution of the internal-representation spectrum (Appendix G).

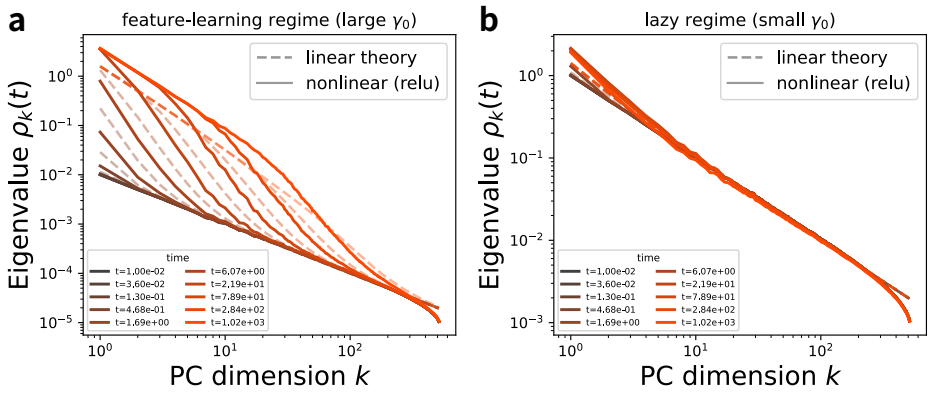


Figure 4: Numerical validation using a nonlinear student network. As training progresses, the slope of the internal-representation spectrum changes in the feature-learning regime but remains close to its initial value in the lazy regime.

The slope of the internal-representation spectrum also changes over the course of learning in the nonlinear network (Fig. 4). In particular, in the feature-learning regime, the spectrum changes successively from low to high mode index, and its exponent evolves from α to $\alpha + \beta$ over a finite range of modes. In the lazy regime, by contrast, it remains at α. These results agree quantitatively with the predictions of the linear theory.

This result suggests that the mechanism analyzed exactly in the two-layer linear model is not peculiar to linear networks, but may capture a basic mechanism of internal-representation formation in a broader class of neural networks.

4 Conclusion

We analyzed how a wide, two-layer linear neural network reshapes the spectrum of its internal representations during learning in a teacher–student setting where the input data and the teacher map have aligned power-law structure. We derived the exact temporal evolution of the eigenvalues of the hidden representation kernel in the population limit and showed that the internal-representation spectrum decomposes into a baseline component inherited from the input covariance at initialization and an increment formed through training.

Using this decomposition, we showed that, in the feature-learning regime and over finite mode ranges sufficiently far from the boundaries, the local effective power-law exponent can take four asymptotic values:

$$\alpha, \qquad 3\alpha + \beta, \qquad \alpha + \beta, \qquad \alpha + \frac{\beta}{2}. \tag{19}$$

These exponents correspond, respectively, to dominance by the initialization component, mode-dependent representation changes at early learning times, post-learning deformation driven primarily by the first layer, and deformation distributed across both layers. In particular, at early learning times, the mode-dependent learning rates set by the input eigenvalues transiently steepen the spectrum. The spectrum then transitions to an exponent determined by the strength of the teacher map and the allocation of learning across layers. In the lazy regime, by contrast, the training-induced component never exceeds the initialization component, and the power-law exponent of the internal representation remains nearly unchanged from the input exponent α.

Numerical experiments with nonlinear networks further showed that the main exponent transitions predicted by the theory, as well as the difference between the feature-learning and lazy regimes, persist beyond the linear model. This result suggests that the interaction among input anisotropy, the modal structure of the teacher signal, and mode-dependent learning rates may also govern internal-representation formation in more general networks.

Our analysis relies on simultaneous diagonalizability of the input covariance and the teacher matrix, infinite width, and population-limit gradient flow. Future work should extend the analysis to misaligned input and teacher structure, finite-width and finite-sample effects, deeper networks, and general nonlinear activation functions. It will also be important to determine the extent to which the predicted exponent transitions and boundaries can be observed in real data and large-scale models.

Taken together, these results show that power-law internal representations should not be regarded as possessing a single fixed exponent. Rather, their exponents are reshaped over time and across modes through the interaction between input statistics and task learning. This work provides a foundation for analytically understanding how power-law representations are inherited and transformed in neural networks.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic datasets, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to draft parts of a research paper, summarize or analyse existing literature, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. We manually checked the manuscript text, analytical calculations, numerical computations, and cited references. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Mathematical details are provided in the appendix section. The code used to generate the figures is provided in the supplementary material.

REFERENCES

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

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

Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 37:35824–35878, 2024b.

Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *International Conference on Learning Representations*, volume 2024, pp. 22088–22127, 2024c.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. In *International Conference on Learning Representations*, volume 2025, pp. 51909–51939, 2025.

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.

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, 10(7):e0129031, 2015.

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*, volume 2025, pp. 102485–102536, 2025.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Francesc Font-Clos, Gemma Boleda, and Alvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New Journal of Physics*, 15(9):093033, 2013.

Raj Magesh Gauthaman, Brice Ménard, and Michael F Bonner. Universal scale-free representations in human visual cortex. *PLoS computational biology*, 21(11):e1013714, 2025.

Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.

Nathan CL Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022.

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.

Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake Richards. Tracing the representation geometry of language models from pretraining to post-training. *Advances in Neural Information Processing Systems*, 38:54691–54724, 2026.

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

Leenoy Meshulam, Jeffrey L Gauthier, Carlos D Brody, David W Tank, and William Bialek. Coarse graining, fixed points, and scaling in a large population of neurons. *Physical review letters*, 123 (17):178103, 2019.

Guillermo B Morales, Serena Di Santo, and Miguel A Muñoz. Quasiuniversal scaling in mouse-brain neuronal activity stems from edge-of-instability critical dynamics. *Proceedings of the National Academy of Sciences*, 120(9):e2208998120, 2023.

Aitor Morales-Gregorio, Anno C Kurth, Junji Ito, Alexander Kleinjohann, Frédéric V Barthélemy, Thomas Brochier, Sonja Grün, and Sacha J van Albada. Neural manifolds in v1 change with top-down signals from v4 targeting the foveal region. *Cell Reports*, 43(7), 2024.

Asahi Nakamuta, Dai Akita, He Zhang, Yuta Kawahara, Hirokazu Takahashi, and Jun-nosuke Teramae. Self-organization of high-dimensional geometry of neural activity in culture. *bioRxiv*, pp. 2025–01, 2025.

Josue Nassar, Piotr Sokol, SueYeon Chung, Kenneth D Harris, and Il Memming Park. On 1/n neural representation and robustness. *Advances in neural information processing systems*, 33: 6211–6222, 2020.

Marius Pachitariu, Lin Zhong, Alexa Gracias, Amanda Minisi, Crystall Lopez, and Carsen Stringer. A critical initialization for biological neural networks. *Nature*, pp. 1–7, 2026.

Elliot Paquette, Ke Liang Xiao, and Yizhe Zhu. Power-law spectrum of the random feature model. *arXiv preprint arXiv:2603.14578*, 2026.

Dean A Pospisil and Jonathan W Pillow. Revisiting the high-dimensional geometry of population responses in the visual cortex. *Proceedings of the National Academy of Sciences*, 122(45): e2506535122, 2025.

Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

Shervin Safavi, Matthew Chalk, Nikos K Logothetis, and Anna Levina. Signatures of criticality in efficient coding networks. *Proceedings of the National Academy of Sciences*, 121(41): e2302730121, 2024.

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.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

Tsuyoshi Tatsukawa and Jun-nosuke Teramae. The cortical critical power law balances energy and information in an optimal fashion. *Proceedings of the National Academy of Sciences*, 122(21): e2418218122, 2025.

Zezhen Wang, Weihao Mai, Yuming Chai, Kexin Qi, Hongtai Ren, Chen Shen, Shiwu Zhang, Guodong Tan, Yu Hu, and Quan Wen. The geometry and dimensionality of brain-wide activity. *Elife*, 14:RP100666, 2025.

Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization. *arXiv preprint arXiv:2510.04780*, 2025.

APPENDIX

Throughout the appendices, we define the following three Gram matrices:

$$\boldsymbol{G}^w(t) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_i(t)\boldsymbol{w}_i(t)^\top, \tag{20}$$

$$\boldsymbol{G}^a(t) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{a}_i(t)\boldsymbol{a}_i(t)^\top, \tag{21}$$

$$\boldsymbol{G}^{wa}(t) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_i(t)\boldsymbol{a}_i(t)^\top. \tag{22}$$

A THE ANISOTROPY OF THE INTERNAL REPRESENTATION CORRESPONDS TO THE EIGENVALUES OF $\boldsymbol{M}(t)$

The kernel of the internal representation can be written as

$$\Phi_t(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{G}^w(t)\boldsymbol{x}'. \tag{23}$$

We wish to determine the eigenvalues of this kernel. In particular, in the population limit, where the number of samples is infinity, empirical averages can be replaced by expectations over the data distribution. The kernel eigenvalue problem then becomes equivalent to that of the following integral operator:

$$(T_\Phi u)(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}'}\left[\boldsymbol{x}^\top \boldsymbol{G}^w(t)\boldsymbol{x}' u(\boldsymbol{x}')\right] = \rho u(\boldsymbol{x}). \tag{24}$$

Because the left-hand side is linear in $\boldsymbol{x}$, every eigenfunction $u(\boldsymbol{x})$ associated with a nonzero eigenvalue is also linear in $\boldsymbol{x}$. We use this fact to derive a matrix $\boldsymbol{M}(t)$ whose eigenvalues coincide with the nonzero eigenvalues of this operator.

Let

$$\psi_k(\boldsymbol{x}) = \frac{x_k}{\sqrt{\lambda_k}} \tag{25}$$

be an orthonormal basis for the subspace of linear functions. These functions are indeed orthonormal under the expectation inner product on the function space containing $u(\boldsymbol{x})$. The action of T_Φ on this basis is

$$(T_\Phi \psi_l)(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}'}\left[\boldsymbol{x}^\top \boldsymbol{G}^w \boldsymbol{x}' \frac{x'_l}{\sqrt{\lambda_l}}\right] \tag{26}$$

$$= \boldsymbol{x}^\top \boldsymbol{G}^w \sqrt{\lambda_l}\boldsymbol{e}_l \tag{27}$$

$$= \sum_k x_k G^w_{kl}\sqrt{\lambda_l} \tag{28}$$

$$= \sum_k \sqrt{\lambda_k} G^w_{kl}\sqrt{\lambda_l}\psi_k(\boldsymbol{x}). \tag{29}$$

Thus, defining

$$\boldsymbol{M}(t) := \boldsymbol{\Lambda}^{1/2}\boldsymbol{G}^w(t)\boldsymbol{\Lambda}^{1/2}, \tag{30}$$

we obtain the matrix representation

$$T_\Phi \psi_l = \sum_k M_{kl}\psi_k. \tag{31}$$

Using the fact that the eigenfunction $u(\boldsymbol{x})$ is linear, we can write $u(\boldsymbol{x}) = \sum_l c_l \psi_l(\boldsymbol{x})$, and hence

$$T_\Phi u = T_\Phi \left(\sum_l c_l \psi_l \right) \tag{32}$$

$$= \sum_l c_l T_\Phi \psi_l \tag{33}$$

$$= \sum_l c_l \sum_k M_{kl} \psi_k \tag{34}$$

$$= \sum_k \left(\sum_l M_{kl} c_l \right) \psi_k. \tag{35}$$

Therefore, the eigenvalue problem for T_Φ becomes

$$\sum_k \left(\sum_l M_{kl} c_l \right) \psi_k = \rho \sum_k c_k \psi_k. \tag{36}$$

Since the ψ_k are linearly independent, uniqueness of the basis expansion gives

$$\sum_l M_{kl} c_l = \rho c_k, \tag{37}$$

$$\therefore \qquad \boldsymbol{M}\boldsymbol{c} = \rho \boldsymbol{c}. \tag{38}$$

This is precisely the eigenvalue problem for $\boldsymbol{M}$. Hence, the nonzero eigenvalues of T_Φ coincide with those of $\boldsymbol{M}(t)$. Consequently, the anisotropy of the internal representation can be characterized by the eigenvalues of

$$\boldsymbol{M}(t) = \boldsymbol{\Lambda}^{1/2} \boldsymbol{G}^w(t) \boldsymbol{\Lambda}^{1/2}. \tag{39}$$

B Warm-up: Derivation of the eigenvalues when the second layer is fixed

As a warm-up, we solve the time evolution of the first-layer weights $\boldsymbol{w}_i(t)$ when the second-layer weights $\boldsymbol{a}_i$ are fixed and only the first layer is trained. This yields an analytical solution for $\boldsymbol{M}(t)$. As shown below, because $\boldsymbol{\Theta}\boldsymbol{\Theta}^\top$ is assumed to be diagonal, $\boldsymbol{M}(t)$ remains diagonal through training, and the dynamics of its eigenvalues ρ_k decouple across k.

The student network can be written as

$$\boldsymbol{f}(\boldsymbol{x}, t) = \frac{1}{\gamma_0} \boldsymbol{G}^{wa}(t)^\top \boldsymbol{x}. \tag{40}$$

In the population limit, define

$$\boldsymbol{R}(t) := \boldsymbol{\Theta} - \frac{1}{\gamma_0} \boldsymbol{G}^{wa}(t). \tag{41}$$

The loss is then

$$L(t) = \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \left[\| \boldsymbol{f}(\boldsymbol{x}, t) - \boldsymbol{y}(\boldsymbol{x}) \|^2 \right] \tag{42}$$

$$= \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \left[\sum_p \left(\sum_k R_{kp}(t) x_k \right)^2 \right] \tag{43}$$

$$= \frac{1}{2} \sum_p \sum_k R_{kp}(t) \lambda_k R_{kp}(t) \tag{44}$$

$$= \frac{1}{2} \operatorname{Tr} \left[\boldsymbol{R}(t)^\top \boldsymbol{\Lambda} \boldsymbol{R}(t) \right]. \tag{45}$$

We now determine the time evolution of $\boldsymbol{M}(t)$ under gradient-flow training in the infinite-data limit.

The first-layer weights obey

$$\dot{w}_{ik}(t) = -\nu \frac{\partial L}{\partial w_{ik}} \tag{46}$$

$$= -\nu \sum_{p=1}^{P} \frac{\partial L}{\partial G^{wa}_{kp}} \frac{\partial G^{wa}_{kp}}{\partial w_{ik}} \tag{47}$$

$$= \frac{\nu}{\gamma_0} \sum_{p=1}^{P} \lambda_k R_{kp}(t) \frac{\partial G^{wa}_{kp}}{\partial w_{ik}} \tag{48}$$

$$= \frac{\nu}{\gamma_0 N} \sum_{p=1}^{P} \lambda_k R_{kp}(t) a_{ip} \tag{49}$$

$$= \frac{1}{\gamma_0} \sum_{p=1}^{P} \lambda_k R_{kp}(t) a_{ip}, \tag{50}$$

$$\therefore \qquad \dot{\boldsymbol{w}}_i(t) = \frac{1}{\gamma_0} \boldsymbol{\Lambda} \boldsymbol{R}(t) \boldsymbol{a}_i, \tag{51}$$

where we set the learning rate to $\nu = N$ so that the parameter dynamics remain $O(1)$ as $N \to \infty$. Combining this equation with the definition of $\boldsymbol{G}^{wa}$ gives

$$\dot{\boldsymbol{G}}^{wa}(t) = \frac{1}{\gamma_0 N} \boldsymbol{\Lambda} \boldsymbol{R}(t) \sum_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \tag{52}$$

$$= \frac{\sigma_a^2}{\gamma_0} \boldsymbol{\Lambda} \left(\boldsymbol{\Theta} - \frac{1}{\gamma_0} \boldsymbol{G}^{wa}(t) \right). \tag{53}$$

Here, using $N \gg D$, we set

$$\frac{1}{N} \sum_i \boldsymbol{a}_i \boldsymbol{a}_i^\top = \sigma_a^2 \boldsymbol{I}. \tag{54}$$

Solving the differential equation for $\boldsymbol{G}^{wa}$ yields

$$G^{wa}_{kp}(t) = \gamma_0 \Theta_{kp} + \left[G^{wa}_{kp}(0) - \gamma_0 \Theta_{kp} \right] \exp\left(-\frac{\sigma_a^2}{\gamma_0^2} \lambda_k t \right). \tag{55}$$

We next obtain the time evolution of $\boldsymbol{M}(t)$. From

$$\boldsymbol{G}^w(t) := \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{w}_i(t) \boldsymbol{w}_i(t)^\top, \tag{56}$$

we have

$$\dot{\boldsymbol{G}}^w(t) = \frac{1}{N} \sum_i \left[\boldsymbol{w}_i(t) \dot{\boldsymbol{w}}_i(t)^\top + \dot{\boldsymbol{w}}_i(t) \boldsymbol{w}_i(t)^\top \right] \tag{57}$$

$$= \frac{1}{\gamma_0 N} \sum_i \left[\boldsymbol{w}_i(t) \boldsymbol{a}_i^\top \boldsymbol{R}(t)^\top \boldsymbol{\Lambda} + \boldsymbol{\Lambda} \boldsymbol{R}(t) \boldsymbol{a}_i \boldsymbol{w}_i(t)^\top \right] \tag{58}$$

$$= \frac{1}{\sigma_a^2} \left[\boldsymbol{G}^{wa}(t) \dot{\boldsymbol{G}}^{wa}(t)^\top + \dot{\boldsymbol{G}}^{wa}(t) \boldsymbol{G}^{wa}(t)^\top \right], \tag{59}$$

$$\therefore \qquad \boldsymbol{G}^w(t) - \boldsymbol{G}^w(0) = \frac{1}{\sigma_a^2} \left[\boldsymbol{G}^{wa}(t) \boldsymbol{G}^{wa}(t)^\top - \boldsymbol{G}^{wa}(0) \boldsymbol{G}^{wa}(0)^\top \right]. \tag{60}$$

In the limit $N \to \infty$,

$$\boldsymbol{G}^{wa}(0) = \frac{1}{N} \sum_i \boldsymbol{w}_i(0) \boldsymbol{a}_i^\top = \boldsymbol{0}, \tag{61}$$

$$\boldsymbol{G}^w(0) = \sigma_w^2 \boldsymbol{I}. \tag{62}$$

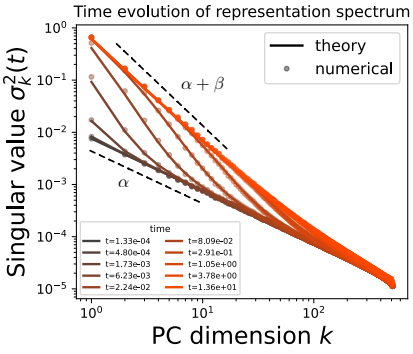


Figure 5: Eigenvalue dynamics for the model with a fixed second layer.

Combining these results, we obtain

$$G^w_{kl}(t) = \sigma_w^2 \delta_{kl} + \frac{1}{c} \sum_p \Theta_{kp} \Theta_{lp} \left[1 - \exp(-c\lambda_k t)\right] \left[1 - \exp(-c\lambda_l t)\right], \tag{63}$$

where

$$c = \frac{\sigma_a^2}{\gamma_0^2}. \tag{64}$$

Since

$$\sum_p \Theta_{kp} \Theta_{lp} = \mu_k \delta_{kl} \tag{65}$$

by assumption, it follows that

$$G^w_{kl}(t) = \left\{ \sigma_w^2 + \frac{\mu_k}{c} \left[1 - \exp(-c\lambda_k t)\right]^2 \right\} \delta_{kl}. \tag{66}$$

Therefore,

$$M_{kl}(t) = \left\{ \lambda_k \sigma_w^2 + \frac{\lambda_k \mu_k}{c} \left[1 - \exp(-c\lambda_k t)\right]^2 \right\} \delta_{kl}, \tag{67}$$

and the eigenvalues $\rho_k(t)$ of $\boldsymbol{M}(t)$ are

$$\rho_k(t) = \lambda_k \sigma_w^2 + \frac{\lambda_k \mu_k}{c} \left[1 - \exp(-c\lambda_k t)\right]^2. \tag{68}$$

The dynamics of this eigenvalue spectrum are illustrated in Fig. 5.

C Warm-up: Derivation of the power-law exponent boundaries when the second layer is fixed

To characterize the power-law exponent around mode k at time t, we introduce

$$\alpha_{\text{eff}}(k,t) := -\frac{\partial \log \rho_k(t)}{\partial \log k}. \tag{69}$$

We derive this quantity from the analytical expression for $\rho_k(t)$. Substituting

$$\lambda_k = \lambda_0 k^{-\alpha}, \qquad \mu_k = \mu_0 k^{-\beta}, \tag{70}$$

we obtain

$$\rho_k(t) = \lambda_0 \sigma_w^2 k^{-\alpha} + \frac{\lambda_0 \mu_0 k^{-(\alpha+\beta)}}{c} \left[1 - \exp(-c\lambda_0 t k^{-\alpha})\right]^2 \tag{71}$$

$$= \lambda_0 \sigma_w^2 k^{-\alpha} \left[1 + \frac{\mu_0 k^{-\beta}}{c\sigma_w^2} \left[1 - \exp(-c\lambda_0 t k^{-\alpha})\right]^2\right] \tag{72}$$

$$= \lambda_0 \sigma_w^2 k^{-\alpha} \left[1 + Q(k,t)\right], \tag{73}$$

where

$$Q(k,t) = \frac{\mu_0 k^{-\beta}}{c\sigma_w^2} \left[1 - \exp(-z_k(t))\right]^2, \tag{74}$$

$$z_k(t) = c\lambda_0 t k^{-\alpha}. \tag{75}$$

Taking logarithms gives

$$\log \rho_k(t) = \log(\lambda_0 \sigma_w^2) - \alpha \log k + \log[1 + Q(k,t)], \tag{76}$$

$$\therefore \qquad \alpha_{\text{eff}}(k,t) = \alpha - \frac{k}{1+Q(k,t)} \frac{\partial Q(k,t)}{\partial k}. \tag{77}$$

The derivative of $Q(k,t)$ is

$$\frac{\partial Q(k,t)}{\partial k} = -\frac{\beta \mu_0 k^{-\beta-1}}{c\sigma_w^2} \left[1 - \exp(-z_k(t))\right]^2 \tag{78}$$

$$+ \frac{2\mu_0 k^{-\beta}}{c\sigma_w^2} \left[1 - \exp(-z_k(t))\right] \exp(-z_k(t)) \frac{\partial z_k(t)}{\partial k} \tag{79}$$

$$= -\frac{\mu_0 k^{-\beta}}{c\sigma_w^2} \left[1 - \exp(-z_k(t))\right]^2 \left[\frac{\beta}{k} + \frac{2}{\exp(z_k(t)) - 1} \frac{\alpha z_k(t)}{k}\right] \tag{80}$$

$$= -\frac{Q(k,t)}{k} \left[\beta + \frac{2\alpha z_k(t)}{\exp(z_k(t)) - 1}\right]. \tag{81}$$

Thus,

$$\alpha_{\text{eff}}(k,t) = \alpha + \frac{Q(k,t)}{1+Q(k,t)} \left[\beta + \frac{2\alpha z_k(t)}{\exp(z_k(t)) - 1}\right]. \tag{82}$$

This expression gives the analytical time evolution of the local power-law exponent. We next use it to derive approximate boundaries that partition the (t,k) plane.

The time dependence of $\rho_k(t)$ enters only through

$$z_k(t) = c\lambda_0 t k^{-\alpha}. \tag{83}$$

Specifically, since $Q(k,t)$ is a sigmoidal function with respect to t, a mode k is unlearned at time t when $z_k(t) \ll 1$, and learned when $z_k(t) \gg 1$. Defining

$$k_{\text{learn}}(t) = (c\lambda_0 t)^{1/\alpha}, \tag{84}$$

the mode is unlearned for $k \gg k_{\text{learn}}(t)$ and learned for $k \ll k_{\text{learn}}(t)$. On a log–log plot, the boundary between the learned and unlearned regions is the line

$$\log k_{\text{learn}}(t) = \frac{1}{\alpha} \left[\log t + \log(c\lambda_0)\right]. \tag{85}$$

We now obtain the asymptotic power-law exponents in each region .

C.1 Learned modes

For learned modes satisfying $k \ll k_{\text{learn}}(t)$, we have $z_k(t) \gg 1$, and hence

$$Q(k,t) \simeq \frac{\mu_0 k^{-\beta}}{c\sigma_w^2}. \tag{86}$$

Therefore,

$$\rho_k(t) \simeq \lambda_0 \sigma_w^2 k^{-\alpha} + \frac{\lambda_0 \mu_0}{c} k^{-(\alpha+\beta)}. \tag{87}$$

The first term is the input-induced anisotropy present at initialization, whereas the second is the anisotropy induced by learning. Because the second term decays more rapidly with k, the spectrum must ultimately decay with exponent α at sufficiently large k. This need not hold, however, at finite k. In particular, for

$$k \ll k_1 := \left(\frac{\mu_0}{c\sigma_w^2}\right)^{1/\beta}, \tag{88}$$

the spectrum decays with exponent $\alpha + \beta$. On a log–log plot, this boundary is the constant line

$$\log k_1 = \frac{1}{\beta} \log\left(\frac{\mu_0}{c\sigma_w^2}\right). \tag{89}$$

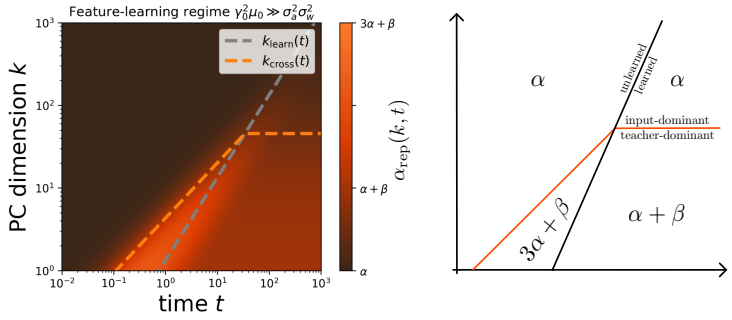


Figure 6: Asymptotic exponent regions for the model with a fixed second layer.

C.2 Unlearned modes

For $k \gg k_{\text{learn}}(t)$, we have

$$Q(k,t) \simeq \frac{\mu_0 k^{-\beta} z_k(t)^2}{c\sigma_w^2}. \tag{90}$$

It follows that

$$\rho_k(t) \simeq \lambda_0 \sigma_w^2 k^{-\alpha} + \frac{\lambda_0 \mu_0 k^{-(\alpha+\beta)}}{c} z_k(t)^2 \tag{91}$$

$$= \lambda_0 \sigma_w^2 k^{-\alpha} + c\lambda_0^3 \mu_0 t^2 k^{-(3\alpha+\beta)}. \tag{92}$$

As in the learned case, the second term decays more rapidly with k, so the asymptotic exponent at sufficiently large k is necessarily α. At finite k, however, modes satisfying

$$k \ll k_2(t) := \left(\frac{c\lambda_0^2 \mu_0}{\sigma_w^2} t^2 \right)^{1/(2\alpha+\beta)} \tag{93}$$

decay with exponent $3\alpha+\beta$. On a log–log plot, this boundary is the line

$$\log k_2(t) = \frac{1}{2\alpha+\beta} \left[2\log t + \log\left(\frac{c\lambda_0^2 \mu_0}{\sigma_w^2} \right) \right]. \tag{94}$$

From the two cases above, the intersection of k_1 and $k_2(t)$ is

$$(\log t^*, \log k^*) = \left(\frac{\alpha}{\beta} \log \frac{\mu_0}{c\sigma_w^2} - \log(c\lambda_0),\ \frac{1}{\beta} \log \frac{\mu_0}{c\sigma_w^2} \right), \tag{95}$$

which lies on $k_{\text{learn}}(t)$. Defining

$$k_{\text{cross}}(t) = \begin{cases} k_1, & t > t^*, \\ k_2(t), & t \le t^*, \end{cases} \tag{96}$$

we can summarize the asymptotic behavior as

$$\alpha_{\text{eff}}(k,t) \simeq \begin{cases} \alpha, & k > k_{\text{cross}}(t), \\ 3\alpha+\beta, & k_{\text{learn}}(t) < k < k_{\text{cross}}(t), \\ \alpha+\beta, & \text{otherwise}. \end{cases} \tag{97}$$

These three regions and their asymptotic exponents are illustrated in Fig. 6.

D Analytical eigenvalue solution when both layers are trained

The network output is

$$\boldsymbol{f}(\boldsymbol{x},t) = \frac{1}{\gamma_0 N} \sum_{i=1}^{N} \boldsymbol{a}_i(t) \boldsymbol{w}_i(t)^\top \boldsymbol{x} \tag{98}$$

$$= \frac{1}{\gamma_0} \boldsymbol{G}^{wa}(t)^\top \boldsymbol{x}. \tag{99}$$

Thus, in the population limit, using

$$\boldsymbol{R}(t) = \boldsymbol{\Theta} - \frac{1}{\gamma_0}\boldsymbol{G}^{wa}(t), \tag{100}$$

the loss can be written as

$$L(t) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\left\|\boldsymbol{f}(\boldsymbol{x},t) - \boldsymbol{y}(\boldsymbol{x})\right\|^2\right] \tag{101}$$

$$= \frac{1}{2}\operatorname{Tr}\left[\boldsymbol{R}(t)^\top \boldsymbol{\Lambda}\boldsymbol{R}(t)\right]. \tag{102}$$

The gradient-flow dynamics are therefore

$$\dot{w}_{ik}(t) = -\nu\frac{\partial L}{\partial w_{ik}} \tag{103}$$

$$= -\nu\sum_{p=1}^{P}\frac{\partial L}{\partial G^{wa}_{kp}}\frac{\partial G^{wa}_{kp}}{\partial w_{ik}} \tag{104}$$

$$= \frac{\nu}{\gamma_0 N}\sum_{p=1}^{P}\lambda_k R_{kp}(t)a_{ip}(t) \tag{105}$$

$$= \frac{1}{\gamma_0}\sum_{p=1}^{P}\lambda_k R_{kp}(t)a_{ip}(t), \tag{106}$$

$$\therefore \qquad \dot{\boldsymbol{w}}_i(t) = \frac{1}{\gamma_0}\boldsymbol{\Lambda}\boldsymbol{R}(t)\boldsymbol{a}_i(t), \tag{107}$$

and

$$\dot{a}_{ip}(t) = -\nu\frac{\partial L}{\partial a_{ip}} \tag{108}$$

$$= -\nu\sum_{k=1}^{D}\frac{\partial L}{\partial G^{wa}_{kp}}\frac{\partial G^{wa}_{kp}}{\partial a_{ip}} \tag{109}$$

$$= \frac{\nu}{\gamma_0 N}\sum_{k=1}^{D}\lambda_k R_{kp}(t)w_{ik}(t) \tag{110}$$

$$= \frac{1}{\gamma_0}\sum_{k=1}^{D}\lambda_k R_{kp}(t)w_{ik}(t), \tag{111}$$

$$\therefore \qquad \dot{\boldsymbol{a}}_i(t) = \frac{1}{\gamma_0}\boldsymbol{R}(t)^\top\boldsymbol{\Lambda}\boldsymbol{w}_i(t). \tag{112}$$

Here again, we set $\nu = N$.

These weight dynamics yield the following Gram-matrix dynamics:

$$\dot{\boldsymbol{G}}^{wa}(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\dot{\boldsymbol{w}}_i(t)\boldsymbol{a}_i(t)^\top + \boldsymbol{w}_i(t)\dot{\boldsymbol{a}}_i(t)^\top\right] \tag{113}$$

$$= \frac{1}{\gamma_0 N}\sum_{i=1}^{N}\left[\boldsymbol{\Lambda}\boldsymbol{R}(t)\boldsymbol{a}_i(t)\boldsymbol{a}_i(t)^\top + \boldsymbol{w}_i(t)\boldsymbol{w}_i(t)^\top\boldsymbol{\Lambda}\boldsymbol{R}(t)\right] \tag{114}$$

$$= \frac{1}{\gamma_0}\left[\boldsymbol{\Lambda}\boldsymbol{R}(t)\boldsymbol{G}^a(t) + \boldsymbol{G}^w(t)\boldsymbol{\Lambda}\boldsymbol{R}(t)\right], \tag{115}$$

$$\dot{\boldsymbol{G}}^w(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\dot{\boldsymbol{w}}_i(t)\boldsymbol{w}_i(t)^\top + \boldsymbol{w}_i(t)\dot{\boldsymbol{w}}_i(t)^\top\right] \tag{116}$$

$$= \frac{1}{\gamma_0 N}\sum_{i=1}^{N}\left[\boldsymbol{\Lambda}\boldsymbol{R}(t)\boldsymbol{a}_i(t)\boldsymbol{w}_i(t)^\top + \boldsymbol{w}_i(t)\boldsymbol{a}_i(t)^\top\boldsymbol{R}(t)^\top\boldsymbol{\Lambda}\right] \tag{117}$$

$$= \frac{1}{\gamma_0}\left[\boldsymbol{\Lambda}\boldsymbol{R}(t)\boldsymbol{G}^{wa}(t)^\top + \boldsymbol{G}^{wa}(t)\boldsymbol{R}(t)^\top\boldsymbol{\Lambda}\right], \tag{118}$$

and

$$\dot{\boldsymbol{G}}^a(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\dot{\boldsymbol{a}}_i(t)\boldsymbol{a}_i(t)^\top + \boldsymbol{a}_i(t)\dot{\boldsymbol{a}}_i(t)^\top\right] \tag{119}$$

$$= \frac{1}{\gamma_0 N}\sum_{i=1}^{N}\left[\boldsymbol{R}(t)^\top\boldsymbol{\Lambda}\boldsymbol{w}_i(t)\boldsymbol{a}_i(t)^\top + \boldsymbol{a}_i(t)\boldsymbol{w}_i(t)^\top\boldsymbol{\Lambda}\boldsymbol{R}(t)\right] \tag{120}$$

$$= \frac{1}{\gamma_0}\left[\boldsymbol{R}(t)^\top\boldsymbol{\Lambda}\boldsymbol{G}^{wa}(t) + \boldsymbol{G}^{wa}(t)^\top\boldsymbol{\Lambda}\boldsymbol{R}(t)\right]. \tag{121}$$

Because the weights are initialized independently and $N \to \infty$,

$$\boldsymbol{G}^{wa}(0) = \boldsymbol{0}, \qquad \boldsymbol{G}^{w}(0) = \sigma_w^2\boldsymbol{I}, \qquad \boldsymbol{G}^{a}(0) = \sigma_a^2\boldsymbol{I}. \tag{122}$$

Moreover, by assumption, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}\boldsymbol{\Theta}^\top$ are diagonal. Since an orthogonal rotation in the output space leaves the loss invariant, we may set, without loss of generality,

$$\boldsymbol{\Theta} = \mathrm{diag}(\theta_0, \theta_1, \ldots) = \mathrm{diag}(\sqrt{\mu_0}, \sqrt{\mu_1}, \ldots). \tag{123}$$

The derived Gram-matrix dynamics shows that if $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$, $\boldsymbol{G}^{wa}$, $\boldsymbol{G}^{w}$, and $\boldsymbol{G}^{a}$ are all diagonal, then their time derivatives are also diagonal. Because they are diagonal at $t = 0$, it follows that $\boldsymbol{G}^{wa}(t)$, $\boldsymbol{G}^{w}(t)$, and $\boldsymbol{G}^{a}(t)$ remain diagonal for all t.

It therefore suffices to consider the independent dynamics of the diagonal entries. Define

$$\xi_k(t) = G^{wa}_{kk}(t), \tag{124}$$

$$\eta_k(t) = G^{w}_{kk}(t), \tag{125}$$

$$\zeta_k(t) = G^{a}_{kk}(t). \tag{126}$$

The initial conditions are

$$\xi_k(0) = 0, \qquad \eta_k(0) = \sigma_w^2, \qquad \zeta_k(0) = \sigma_a^2, \tag{127}$$

and the equations of motion are

$$\dot{\xi}_k(t) = \frac{\lambda_k}{\gamma_0}\left(\theta_k - \frac{\xi_k(t)}{\gamma_0}\right)\left[\eta_k(t) + \zeta_k(t)\right], \tag{128}$$

$$\dot{\eta}_k(t) = \dot{\zeta}_k(t) = \frac{2\lambda_k\xi_k(t)}{\gamma_0}\left(\theta_k - \frac{\xi_k(t)}{\gamma_0}\right). \tag{129}$$

We seek the eigenvalues of

$$\boldsymbol{M}(t) = \boldsymbol{\Lambda}^{1/2}\boldsymbol{G}^{w}(t)\boldsymbol{\Lambda}^{1/2}. \tag{130}$$

In the present setting, these are

$$\rho_k(t) = \lambda_k\eta_k(t). \tag{131}$$

The system has two conserved quantities, allowing us to eliminate two variables. First, because $\dot{\eta}_k - \dot{\zeta}_k = 0$,

$$\eta_k(t) - \zeta_k(t) = \sigma_w^2 - \sigma_a^2 \qquad \text{(constant)}. \tag{132}$$

Next, define

$$s_k(t) := \eta_k(t) + \zeta_k(t). \tag{133}$$

Then

$$\dot{s}_k = \frac{4\lambda_k \xi_k}{\gamma_0}\left(\theta_k - \frac{\xi_k}{\gamma_0}\right), \tag{134}$$

$$\frac{ds_k}{d\xi_k} = \frac{\dot{s}_k}{\dot{\xi}_k} = \frac{4\xi_k}{s_k}, \tag{135}$$

$$s_k\, ds_k = 4\xi_k\, d\xi_k, \tag{136}$$

$$s_k(t)^2 - s_k(0)^2 = 4\xi_k(t)^2 - 4\xi_k(0)^2, \tag{137}$$

$$s_k(t)^2 - 4\xi_k(t)^2 = (\sigma_w^2 + \sigma_a^2)^2 \qquad \text{(constant)}. \tag{138}$$

Using these two conserved quantities, $\eta_k(t)$ can be expressed as a function of $\xi_k(t)$:

$$\begin{cases} \eta_k(t) + \zeta_k(t) = \sqrt{4\xi_k(t)^2 + (\sigma_w^2 + \sigma_a^2)^2}, \\ \eta_k(t) - \zeta_k(t) = \sigma_w^2 - \sigma_a^2, \end{cases} \tag{139}$$

and hence

$$\eta_k(t) = \sigma_w^2 + \frac{1}{2}\left[-(\sigma_w^2 + \sigma_a^2) + \sqrt{4\xi_k(t)^2 + (\sigma_w^2 + \sigma_a^2)^2}\right]. \tag{140}$$

Thus, solving

$$\dot{\xi}_k = \frac{\lambda_k}{\gamma_0}\left(\theta_k - \frac{\xi_k}{\gamma_0}\right)\sqrt{4\xi_k(t)^2 + (\sigma_w^2 + \sigma_a^2)^2} \tag{141}$$

allows us to obtain the analytical eigenvalue solution through the sequence $\xi_k(t) \to \eta_k(t) \to \rho_k(t)$. We solve for $\xi_k(t)$ by a change of variables and separation of variables. Let

$$S := \sigma_w^2 + \sigma_a^2, \quad m_k := \gamma_0\theta_k, \quad b_k := \frac{\lambda_k}{\gamma_0^2}, \quad u_k := \frac{\xi}{S}, \quad \ell_k := \operatorname{arsinh}(2u_k). \tag{142}$$

Then

$$\begin{cases} \text{left-hand side} = \dot{\xi}_k = \dfrac{S}{2}\cosh(\ell_k)\dot{\ell}_k, \\ \text{right-hand side} = b_k\left(m_k - \dfrac{S}{2}\sinh(\ell_k)\right)S\cosh(\ell_k). \end{cases} \tag{143}$$

The initial-value problem is therefore

$$\dot{\ell}_k = b_k\left(2m_k - S\sinh(\ell_k)\right), \qquad \ell_k(0) = 0. \tag{144}$$

Next, let

$$v_k = \tanh\left(\frac{\ell_k}{2}\right). \tag{145}$$

Then

$$\begin{cases} \dot{v}_k = \dfrac{\dot{\ell}_k}{2\cosh^2(\ell_k/2)} = \dfrac{1 - v_k^2}{2}\dot{\ell}_k, \\ \sinh(\ell_k) = \dfrac{2v_k}{1 - v_k^2}. \end{cases} \tag{146}$$

It follows that

$$\dot{v}_k = b_k\left[m_k(1 - v_k^2) - Sv_k\right], \qquad v_k(0) = 0. \tag{147}$$

This equation can be solved by partial fractions. Let

$$r_k^{\pm} = \frac{-S \pm \sqrt{S^2 + 4m_k^2}}{2m_k} \tag{148}$$

be the two roots of

$$m_k v_k^2 + Sv_k - m_k = 0. \tag{149}$$

Then

$$b_k t = \frac{1}{m_k(r_k^+ - r_k^-)} \int_0^{v_k(t)} \left(\frac{1}{v - r_k^-} - \frac{1}{v - r_k^+} \right) dv \tag{150}$$

$$= \frac{1}{m_k(r_k^+ - r_k^-)} \log \left| \frac{r_k^+ \left(v_k(t) - r_k^-\right)}{r_k^- \left(v_k(t) - r_k^+\right)} \right| \tag{151}$$

$$= \frac{1}{\sqrt{S^2 + 4m_k^2}} \log \left| \frac{r_k^+ \left(v_k(t) - r_k^-\right)}{r_k^- \left(v_k(t) - r_k^+\right)} \right|. \tag{152}$$

This equation can be inverted to obtain $v_k(t)$. Define

$$E_k(t) := \exp\left[b_k \sqrt{S^2 + 4m_k^2}\, t \right]. \tag{153}$$

Then

$$\frac{r_k^+ \left(v_k(t) - r_k^-\right)}{r_k^- \left(v_k(t) - r_k^+\right)} = E_k(t), \tag{154}$$

$$\therefore \qquad v_k(t) = \frac{r_k^+ r_k^- \left[1 - E_k(t)\right]}{r_k^+ - r_k^- E_k(t)} \tag{155}$$

$$= \frac{r_k^+ \left[E_k(t) - 1\right]}{(r_k^+)^2 + E_k(t)}, \tag{156}$$

where we used $r_k^+ r_k^- = -1$ in the last line.

The analytical solution can therefore be expressed through the sequence $E_k(t) \to v_k(t) \to u_k(t) \to \eta_k(t)$:

$$\eta_k(t) = \sigma_w^2 + \frac{S}{2}\left[-1 + \sqrt{1 + 4u_k(t)^2} \right], \tag{157}$$

$$u_k(t) = \frac{v_k(t)}{1 - v_k(t)^2}, \tag{158}$$

$$v_k(t) = \frac{r_k^+ \left[E_k(t) - 1\right]}{(r_k^+)^2 + E_k(t)}, \tag{159}$$

$$E_k(t) = \exp\left[\frac{\lambda_k \sqrt{S^2 + 4\gamma_0^2 \theta_k^2}}{\gamma_0^2} t \right], \tag{160}$$

$$r_k^+ = \frac{-S + \sqrt{S^2 + 4\gamma_0^2 \theta_k^2}}{2\gamma_0 \theta_k}, \tag{161}$$

$$S = \sigma_w^2 + \sigma_a^2. \tag{162}$$

This provides the analytical solution for the eigenvalues $\rho_k(t) = \lambda_k \eta_k(t)$. As explained in the next section, however, the full analytical solution is not required to evaluate the power-law exponents.

E Derivation of the power-law exponents when both layers are trained

E.1 Exact local exponent

We derive the analytical expression for the power-law exponent presented in Result 3.2. The derivation uses the following equations:

$$\begin{cases} \rho_k(t) = \lambda_k \sigma_w^2 \left[1 + R(u_k)\right], \\ R(u_k) = \dfrac{S}{2\sigma_w^2} \left[-1 + g(u_k)\right], \\ \dot{u}_k = \dfrac{\lambda_k S}{\gamma_0^2} (u_{k,\infty} - u_k) g(u_k), \end{cases}$$

with the initial condition $u_k(0) = 0$, where

$$u_{k,\infty} := \frac{\gamma_0 \theta_k}{S}, \qquad S := \sigma_w^2 + \sigma_a^2, \qquad g(x) := \sqrt{1 + 4x^2}.$$

We first define

$$p(k,t) := -\frac{\partial \log u_k(t)}{\partial \log k}$$

and express $\alpha_{\text{eff}}(k,t)$ in terms of $p(k,t)$. From the analytical expression for the eigenspectrum, we obtain

$$\begin{aligned}
\alpha_{\text{eff}}(k,t) &= -\frac{\partial \log \rho_k(t)}{\partial \log k} \\
&= \alpha + \frac{R(u_k(t))}{1 + R(u_k(t))} \left(-\frac{\partial \log R(u_k(t))}{\partial \log k} \right) \\
&= \alpha + \frac{R(u_k(t))}{1 + R(u_k(t))} \left(1 + \frac{1}{g(u_k(t))} \right) \left(-\frac{\partial \log u_k(t)}{\partial \log k} \right) \\
&= \alpha + \frac{R(u_k(t))}{1 + R(u_k(t))} \left(1 + \frac{1}{g(u_k(t))} \right) p(k,t).
\end{aligned}$$

Next, we determine $p(k,t)$ from the evolution equation for $u_k(t)$. Separating variables and integrating yields

$$\frac{\lambda_k S t}{\gamma_0^2} = F(u_k(t), u_{k,\infty}) := \int_0^{u_k(t)} \frac{dz}{(u_{k,\infty} - z) g(z)}.$$

Differentiating both sides with respect to $\log k$ at fixed t, the left-hand side becomes

$$\begin{aligned}
\frac{\partial}{\partial \log k} \frac{\lambda_k S t}{\gamma_0^2} &= k \frac{\partial}{\partial k} \frac{\lambda_k S t}{\gamma_0^2} \\
&= -\alpha \frac{\lambda_k S t}{\gamma_0^2} \\
&= -\alpha F(u_k(t), u_{k,\infty}),
\end{aligned}$$

while the right-hand side becomes

$$\frac{\partial F}{\partial \log k} = u_k(t) \frac{\partial F}{\partial u_k(t)} \frac{\partial \log u_k(t)}{\partial \log k} + \frac{\partial F}{\partial u_{k,\infty}} \frac{\partial u_{k,\infty}}{\partial \log k}.$$

Evaluating the derivatives and rearranging, we obtain

$$p(k,t) = \frac{(u_{k,\infty} - u_k(t)) g(u_k(t))}{u_k(t)} \int_0^{u_k(t)} \left[\frac{\alpha}{u_{k,\infty} - z} + \frac{u_{k,\infty} \beta / 2}{(u_{k,\infty} - z)^2} \right] \frac{dz}{g(z)}.$$

E.2 Boundaries

Since $u_k(t)$ increases monotonically from zero toward $u_{k,\infty} = \gamma_0 \theta_k / S$, we can define the hitting time $t_k^*(u)$ for each threshold $0 < u < u_{k,\infty}$:

$$t_k^*(u) = \frac{\gamma_0^2}{\lambda_k S} \int_0^u \frac{dz}{(u_{k,\infty} - z)\sqrt{1 + 4z^2}}. \tag{163}$$

The boundaries k_{cross}, k_{learn}, and k_{layer} are then defined implicitly by

$$t = t^*_{k_{\mathrm{cross}}(t)}(u_{\mathrm{cross}}), \qquad t = t^*_{k_{\mathrm{learn}}(t)}(0.9u_{k,\infty}), \qquad t = t^*_{k_{\mathrm{layer}}(t)}(1),$$

whenever the corresponding thresholds are reachable. The uniqueness of these boundaries follows from the preservation of the ordering of $u_k(t)$ across modes throughout the dynamics, as shown in Appendix F.

E.3 Exponents within each domain

We first show that

$$0 \le p(k,t) - \left[\alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}\right] \le 2\left(\alpha + \frac{\beta}{2}\right) u_k(t)^2, \tag{*}$$

holds for $t \ge 0$, where

$$h(x) := \frac{1-x}{x}\log\frac{1}{1-x}.$$

Using the fact that

$$1 \le \frac{g(u_k(t))}{g(z)} \le g(u_k(t))$$

throughout the integration interval in the exact expression, we obtain

$$\begin{aligned} p(k,t) &\ge \frac{u_{k,\infty} - u_k(t)}{u_k(t)} \int_0^{u_k(t)} \left[\frac{\alpha}{u_{k,\infty} - z} + \frac{u_{k,\infty}\beta/2}{(u_{k,\infty} - z)^2}\right] dz \\ &= \alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}, \\ p(k,t) &\le g(u_k(t)) \left[\alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}\right]. \end{aligned}$$

Furthermore,

$$\begin{aligned} &p(k,t) - \left[\alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}\right] \\ &\quad\le \big(g(u_k(t)) - 1\big)\left[\alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}\right] \\ &\quad\le \big(g(u_k(t)) - 1\big)\left(\alpha + \frac{\beta}{2}\right) \qquad \big(h(x) \le 1\big) \\ &\quad\le 2u_k(t)^2\left(\alpha + \frac{\beta}{2}\right) \qquad \big(g(x) \le 1 + 2x^2\big). \end{aligned}$$

This establishes inequality (*). We now evaluate the power-law exponent in each regime.

Region I. We bound each factor in the exact expression

$$\alpha_{\mathrm{eff}}(k,t) = \alpha + \frac{R(u_k(t))}{1 + R(u_k(t))}\left(1 + \frac{1}{g(u_k(t))}\right) p(k,t).$$

First,

$$\begin{aligned} \frac{R(u_k(t))}{1 + R(u_k(t))}\left(1 + \frac{1}{g(u_k(t))}\right) &\le 2R(u_k(t)) && \big(g(u) \ge 1,\ R(u) \ge 0\big) \\ &\le \frac{2S}{\sigma_w^2} u_k(t)^2 && \big(g(x) \le 1 + 2x^2\big) \\ &\le 4\left(\frac{u_k(t)}{u_{\mathrm{cross}}}\right)^2, \end{aligned}$$

where the last inequality follows from

$$u_{\text{cross}}^2 = \frac{\sigma_w^2(S+\sigma_w^2)}{S^2} \qquad \text{and} \qquad S \geq \sigma_w^2.$$

Moreover, inequality (*), together with $h(x) \leq 1$, gives

$$0 \leq p(k,t) \leq \left(\alpha + \frac{\beta}{2}\right)\left(1 + 2u_k(t)^2\right) = O(1)$$

in Region I. Consequently,

$$\alpha_{\text{eff}}(k,t) = \alpha + O\left(\left(\frac{u_k(t)}{u_{\text{cross}}}\right)^2\right).$$

Region II. As above, we estimate each factor separately. In the regime $u_{\text{cross}} \ll u_k(t) \ll \min(1, u_{k,\infty})$, we have

$$\begin{aligned}
&\frac{R(u_k(t))}{1+R(u_k(t))}\left(1 + \frac{1}{g(u_k(t))}\right) \\
&\quad = \left(1 + O\left(R(u_k(t))^{-1}\right)\right)\left(2 + O\left(u_k(t)^2\right)\right) \\
&\quad = 2 + O\left(\left(\frac{u_{\text{cross}}}{u_k(t)}\right)^2 + u_k(t)^2\right).
\end{aligned}$$

By inequality (*),

$$\begin{aligned}
p(k,t) &= \left[\alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2}\right] + O\left(u_k(t)^2\right) \\
&= \alpha + \frac{\beta}{2} + O\left(\frac{u_k(t)}{u_{k,\infty}} + u_k(t)^2\right),
\end{aligned}$$

where we used $h(x) = 1 + O(x)$ as $x \to 0$.

Consequently,

$$\alpha_{\text{eff}}(k,t) = 3\alpha + \beta + O\left(\left(\frac{u_{\text{cross}}}{u_k(t)}\right)^2 + \frac{u_k(t)}{u_{k,\infty}} + u_k(t)^2\right).$$

By contrast, in a domain where $1 \ll u_k(t) \ll u_{k,\infty}$, $p(k,t)$ remains mode-dependent. The eigenvalue spectrum $\rho_k(t)$ then has no single power-law exponent.

Region IIIa. As in Region II, the first factor satisfies

$$\frac{R(u_k(t))}{1+R(u_k(t))}\left(1 + \frac{1}{g(u_k(t))}\right) = 2 + O\left(\left(\frac{u_{\text{cross}}}{u_k(t)}\right)^2 + u_k(t)^2\right).$$

For the second factor, inequality (*) gives

$$\begin{aligned}
p(k,t) &= \alpha h\left(\frac{u_k(t)}{u_{k,\infty}}\right) + \frac{\beta}{2} + O\left(u_k(t)^2\right) \\
&\to \frac{\beta}{2} + O\left(u_k(t)^2\right) \qquad \left(u_k(t) \to u_{k,\infty}\right),
\end{aligned}$$

since $h(x) \to 0$ as $x \to 1^-$.

Consequently,

$$\alpha_{\text{eff}}(k,t) \to \alpha + \beta + O\left(\left(\frac{u_{\text{cross}}}{u_k(t)}\right)^2 + u_k(t)^2\right).$$

Region IIIb. For the first factor, we have

$$\begin{aligned}&\frac{R(u_k(t))}{1+R(u_k(t))}\left(1+\frac{1}{g(u_k(t))}\right)\\&\quad=\left(1+O\left(R(u_k(t))^{-1}\right)\right)\left(1+O\left(u_k(t)^{-1}\right)\right)\\&\quad=1+O\left(\left(\frac{u_{\mathrm{cross}}}{u_k(t)}\right)^2+u_k(t)^{-1}\right).\end{aligned}$$

For the second factor, the bound in inequality (*) is not sufficiently sharp, since $u_k(t) \gg 1$ in this regime. We therefore use integration by parts to obtain a sharper bound.

Integrating both sides of the identity

$$\frac{d}{dz}\frac{z}{(u_{k,\infty}-z)g(z)}=\frac{u_{k,\infty}}{(u_{k,\infty}-z)^2g(z)}-\frac{4z^2}{(u_{k,\infty}-z)g(z)^3}$$

gives

$$\begin{aligned}&\frac{u_k(t)}{(u_{k,\infty}-u_k(t))g(u_k(t))}\\&\quad=\int_0^{u_k(t)}\frac{u_{k,\infty}\,dz}{(u_{k,\infty}-z)^2g(z)}-\int_0^{u_k(t)}\frac{4z^2}{1+4z^2}\frac{dz}{(u_{k,\infty}-z)g(z)}.\end{aligned}$$

Substituting this identity into the exact expression for $p(k,t)$ yields

$$\begin{aligned}p(k,t)=\frac{\beta}{2}+\frac{(u_{k,\infty}-u_k(t))g(u_k(t))}{u_k(t)}\\\times\int_0^{u_k(t)}\frac{\alpha+\frac{\beta}{2}\frac{4z^2}{1+4z^2}}{(u_{k,\infty}-z)g(z)}\,dz.\end{aligned}$$

Using

$$0\le\frac{4z^2}{1+4z^2}\le 1,$$

we obtain

$$\begin{aligned}0\le p(k,t)-\frac{\beta}{2}\le\left(\alpha+\frac{\beta}{2}\right)\frac{(u_{k,\infty}-u_k(t))g(u_k(t))}{u_k(t)}\\\times\int_0^{u_k(t)}\frac{dz}{(u_{k,\infty}-z)g(z)}.\end{aligned}$$

To estimate the remaining integral, we use the decomposition

$$\begin{aligned}&\frac{u_{k,\infty}g(u_k(t))}{u_k(t)}\int_0^{u_k(t)}\frac{dz}{(u_{k,\infty}-z)g(z)}\\&\quad=\frac{g(u_k(t))}{u_k(t)}\int_0^{u_k(t)}\frac{dz}{g(z)}+\int_0^{u_k(t)}\frac{g(u_k(t))}{u_k(t)}\frac{z}{g(z)}\frac{dz}{u_{k,\infty}-z}.\end{aligned}$$

The first integral satisfies

$$\int_0^{u_k(t)}\frac{dz}{g(z)}=\frac{\operatorname{arsinh}(2u_k(t))}{2}=O\left(\log(4u_k(t))\right).$$

Moreover, since $z/g(z)$ is increasing for $z \ge 0$,

$$\frac{g(u_k(t))}{u_k(t)}\frac{z}{g(z)}\le 1\qquad(0\le z\le u_k(t)).$$

Thus,

$$\int_0^{u_k(t)} \frac{g(u_k(t))}{u_k(t)} \frac{z}{g(z)} \frac{dz}{u_{k,\infty} - z} \leq \int_0^{u_k(t)} \frac{dz}{u_{k,\infty} - z} = \log \frac{u_{k,\infty}}{u_{k,\infty} - u_k(t)}.$$

Combining these estimates with $g(u_k(t))/u_k(t) = O(1)$ for $u_k(t) \gg 1$, we obtain

$$\frac{(u_{k,\infty} - u_k(t))g(u_k(t))}{u_k(t)} \int_0^{u_k(t)} \frac{dz}{(u_{k,\infty} - z)g(z)} = O\left(\frac{u_{k,\infty} - u_k(t)}{u_{k,\infty}} \log \frac{4u_k(t)u_{k,\infty}}{u_{k,\infty} - u_k(t)} \right).$$

Consequently,

$$p(k,t) = \frac{\beta}{2} + O\left(\frac{u_{k,\infty} - u_k(t)}{u_{k,\infty}} \log \frac{4u_k(t)u_{k,\infty}}{u_{k,\infty} - u_k(t)} \right),$$

with $p(k,t) \geq \beta/2$.

For any fixed k, the error term vanishes as $u_k(t) \to u_{k,\infty}^-$. at sufficiently large t. Consequently, sufficiently far to the right within Region IIIb,

$$\alpha_{\text{eff}}(k,t) \to \alpha + \frac{\beta}{2} + O\left(\left(\frac{u_{\text{cross}}}{u_{k,\infty}} \right)^2 + u_{k,\infty}^{-1} \right).$$

E.4 Conditions for the existence of the asymptotic regions

We now derive conditions on the hyperparameters α, β, γ_0, σ_w^2, σ_a^2, λ_0, and μ_0 under which each region exists.

Region I. At $t = 0$, we always have $R_k(0) = 0 < 1$, so this region always exists.

Region II. Since the upper end of this region is the k-coordinate of the intersection between $k_{\text{cross}}(t)$ and $k_{\text{learn}}(t)$, this region exists when the coordinate is much larger than 1. Because $t_k^*(u)$ is monotonically increasing, the k-coordinate of the intersection is determined by

$$\frac{0.9\gamma_0\theta_0 k^{-\beta/2}}{S} = \frac{\sqrt{\sigma_w^2(2\sigma_w^2 + \sigma_a^2)}}{S}. \tag{164}$$

Therefore, the condition for this region to exist over a wide range within $1 \leq k \leq D$ is

$$\left(\frac{0.9\gamma_0\theta_0}{\sqrt{\sigma_w^2(2\sigma_w^2 + \sigma_a^2)}} \right)^{2/\beta} \gg 1. \tag{165}$$

Moreover, Region II requires $u_{\text{cross}} \ll u_k(t) \ll \min(1, u_{k,\infty})$. In particular, we must have $u_{\text{cross}} \ll 1$, which requires

$$\sigma_a^2 \gg \sigma_w^2.$$

Thus, the initialization variance of the second layer must be much larger than that of the first layer.

Region IIIa. Similarly, denoting the upper and lower boundaries of this region by k_u and k_l, respectively, we have $u_{k_u,\infty} = u_{\text{cross}}$ and $u_{k_l,\infty} = 1$. The condition for this region to exist over a broad range is

$$\max(1, k_l) \ll \min(D, k_u). \tag{166}$$

A necessary condition, obtained by solving $k_l \ll k_u$, is

$$\frac{\gamma_0\theta_0}{S} \ll \frac{\gamma_0\theta_0}{\sqrt{\sigma_w^2(2\sigma_w^2 + \sigma_a^2)}}, \tag{167}$$

$$\therefore \sigma_a^2 \gg \frac{-1 + \sqrt{5}}{2}\sigma_w^2. \tag{168}$$

In other words, the initialization scale of the second layer must be sufficiently larger than that of the first layer for Region IIIa to exist. We consider sufficiently large input dimension D, with all other parameters held fixed and $\beta > 0$. Since $k_l = (\gamma_0\theta_0/S)^{2/\beta}$ is independent of D, the condition $k_l \ll D$ is automatically satisfied in this limit.

Region IIIb. The conditions for this region are

$$k < k_u \qquad \text{and} \qquad k < k_l. \tag{169}$$

The required condition is therefore

$$\left[\min\left(\frac{\gamma_0\theta_0}{S}, \frac{\gamma_0\theta_0}{\sqrt{\sigma_w^2(2\sigma_w^2+\sigma_a^2)}}\right)\right]^{2/\beta} \gg 1. \tag{170}$$

Combining the above results, the four regions coexist when

$$\sigma_w \ll \sigma_a \qquad \text{and} \qquad \frac{\sigma_w^2+\sigma_a^2}{\theta_0} \ll \gamma_0 \ll \frac{D^{\beta/2}(\sigma_w^2+\sigma_a^2)}{\theta_0}.$$

The conditions that the initialization scale of the second layer is larger than that of the first layer and that γ_0 is large generally correspond to the feature-learning regime, in which the second layer changes significantly. Conversely, when

$$\gamma_0 \ll \frac{\sqrt{\sigma_w^2(2\sigma_w^2+\sigma_a^2)}}{\theta_0},$$

the entire range $1 \le k \le D$ belongs to Region I, and the power-law exponent of the spectrum remains nearly unchanged. In this paper, we therefore refer to this situation as the lazy regime, in contrast to the feature-learning regime.

F Proof that the ordering of the eigenvalues $\rho_k(t)$ is preserved

Consider the system

$$\begin{cases} \rho_k(t) = \lambda_k\left\{\sigma_w^2 + \dfrac{S}{2}\left[-1+\sqrt{1+\dfrac{4\xi_k(t)^2}{S^2}}\right]\right\}, \\ \dot{\xi}_k = \dfrac{\lambda_k}{\gamma_0}\left(\theta_k - \dfrac{\xi_k}{\gamma_0}\right)\sqrt{4\xi_k(t)^2+S^2}, \\ \xi_k(0) = 0, \\ \lambda_k = \lambda_0 k^{-\alpha}, \\ \theta_k = \theta_0 k^{-\beta/2}. \end{cases} \tag{171}$$

We show that if $k > l$, then

$$\rho_k(t) \le \rho_l(t) \tag{172}$$

for every t; that is, the ordering of the eigenvalues is preserved throughout training.

Define

$$\phi(\xi,\theta) := \frac{1}{\gamma_0}\left(\theta - \frac{\xi}{\gamma_0}\right)\sqrt{4\xi^2+S^2}, \tag{173}$$

and denote by $\xi(t;\lambda,\theta)$ the solution of the initial-value problem

$$\dot{\xi} = \lambda\phi(\xi,\theta), \qquad \xi(0) = 0. \tag{174}$$

For $\theta \ge 0$, the solution satisfies

$$0 \le \xi(t;\lambda,\theta) \le \gamma_0\theta, \tag{175}$$

and along this trajectory,

$$\phi(\xi(t;\lambda,\theta),\theta) \ge 0. \tag{176}$$

Hence, $\xi(t;\lambda,\theta)$ is monotonically increasing in t.

We next show that $\xi(t;\lambda,\theta)$ is monotonically increasing in both λ and θ. Let

$$p := \partial_\lambda \xi. \tag{177}$$

Differentiating the initial-value problem gives

$$\dot{p} = \phi(\xi,\theta) + \lambda \partial_\xi \phi(\xi,\theta) p, \qquad p(0) = 0. \tag{178}$$

Its solution is

$$p(t) = \int_0^t \phi(\xi(s),\theta) \exp\left[\lambda \int_s^t \partial_\xi \phi(\xi(u),\theta)\, du\right] ds \geq 0. \tag{179}$$

Therefore,

$$\partial_\lambda \xi(t;\lambda,\theta) \geq 0. \tag{180}$$

Similarly, let

$$q := \partial_\theta \xi. \tag{181}$$

Then

$$\dot{q} = \lambda \partial_\xi \phi(\xi,\theta) q + \lambda \partial_\theta \phi(\xi,\theta), \qquad q(0) = 0, \tag{182}$$

with

$$\partial_\theta \phi(\xi,\theta) = \frac{\sqrt{4\xi^2 + S^2}}{\gamma_0} \geq 0. \tag{183}$$

Thus,

$$q(t) = \lambda \int_0^t \partial_\theta \phi(\xi(s),\theta) \exp\left[\lambda \int_s^t \partial_\xi \phi(\xi(u),\theta)\, du\right] ds \geq 0, \tag{184}$$

and consequently

$$\partial_\theta \xi(t;\lambda,\theta) \geq 0. \tag{185}$$

Finally, because

$$\partial_\lambda \xi(t;\lambda,\theta) \geq 0 \qquad \text{and} \qquad \partial_\theta \xi(t;\lambda,\theta) \geq 0, \tag{186}$$

and because both λ_k and θ_k are ordered in descending order, $k > l$ implies

$$\begin{aligned} \xi_k(t) &= \xi(t;\lambda_k,\theta_k) && (187)\\ &\leq \xi(t;\lambda_l,\theta_k) && (188)\\ &\leq \xi(t;\lambda_l,\theta_l) && (189)\\ &= \xi_l(t). && (190) \end{aligned}$$

Combining this result with the fact that ρ_k is monotonically increasing in ξ_k and λ gives

$$\rho_k(t) \leq \rho_l(t). \tag{191}$$

Therefore, the input-dimension index k coincides with the eigenvalue rank.

G Numerical settings

For all plots, we used $\lambda_k = k^{-1}$ and $\mu_k = \zeta(2)^{-1} k^{-1}$; equivalently, $\alpha = \beta = \lambda_0 = 1$ and $\mu_0 = 1/\zeta(\alpha+\beta) = 1/\zeta(2)$. Thus, $\sum_{k=1}^{\infty} \lambda_k \mu_k = 1$. Inputs and targets were generated as $x \sim \mathcal{N}(0,\Lambda)$ and $y = \Theta^\top x$, with $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_D)$ and $\Theta = \mathrm{diag}(\sqrt{\mu_1},\ldots,\sqrt{\mu_D})$. The initialization scale of the second layer was fixed at $\sigma_a = 1.0$.

Figure 2. We trained the finite-width linear network in Eq. (1) with $D = P = 512$, $N = 8192$, and $M = 4096$ fixed i.i.d. training samples by full-batch gradient descent on the empirical mean-squared loss. The remaining parameters were

$$(\sigma_w, \gamma_0) = \left(0.1, 2.0 \cdot \frac{\sigma_w^2 + \sigma_a^2}{\theta_0}\right),$$

with learning rate $\nu = N$. At each checkpoint, letting $H_{mi}(t) = w_i(t)^\top x_m$, we plotted the nonzero eigenvalues of $H(t)H(t)^\top/(MN)$ in decreasing order.

Figure 4. We used the same data model and teacher, and replaced the student by

$$f_\phi(x,t) = \frac{1}{\gamma_0 N}\sum_{i=1}^{N} a_i(t)\mathrm{ReLU}\big(w_i(t)^\top x\big)\,.$$

The numerical settings were

$$(D,P,N,M) = (512, 512, 8192, 4096).$$

We used $\sigma_w = 0.1, \gamma_0 = 2.0(\sigma_w^2 + \sigma_a^2)/\theta_0$ in the feature-learning regime and $\sigma_w = 1.0, \gamma_0 = 0.5(\sigma_w^2 + \sigma_a^2)/\theta_0$ in the lazy regime. Spectra were recorded same as Figure 2 for preactivations.