SPECTRAL DYNAMICS IN DEEP NETWORKS: FEATURE LEARNING, OUTLIER ESCAPE, AND LEARNING RATE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the evolution of hidden-weight spectra in wide neural networks trained by (stochastic) gradient descent. We develop a two-level dynamical mean-field theory (DMFT) that jointly tracks bulk and outlier spectral dynamics for spiked ensembles whose spike directions remain statistically dependent on the random bulk. We apply this framework to two settings: (1) infinite-width nonlinear networks in mean-field/μP scaling and (2) deep linear networks in the proportional high-dimensional limit, where width, input dimension, and sample size diverge with fixed ratios. Our theory predicts how outliers evolve with training time, width, output scale, and initialization variance. In deep linear networks, μP yields width-consistent outlier dynamics and hyperparameter transfer, including width-stable growth of the leading NTK mode toward the edge of stability (EoS). In contrast, NTK parameterization exhibits strongly width-dependent outlier dynamics, despite converging to a stable large-width limit. We show that this bulk+outlier picture is descriptive of simple tasks with small output channels, but that tasks involving large numbers of outputs (ImageNet classification or GPT language modeling) are better described by a restructuring of the spectral bulk. We develop a toy model with extensive output channels that recapitulates this phenomenon and show that edge of the spectrum still converges for sufficiently wide networks.

1 INTRODUCTION

Understanding how large neural networks learn internal representations remains a central problem in deep learning theory. Motivated by both theoretical tractability and empirical observations about the benefits of larger neural network models (Kaplan et al., 2020; Hoffmann et al., 2022; Achiam et al., 2023), significant theoretical attention has been paid to various infinite width limits of neural networks in recent years (Jacot et al., 2018; Lee et al., 2019; Bahri et al., 2021; Bordelon et al., 2020; Loureiro et al., 2021). While some infinite width limits of neural networks generate a *kernel behavior* where hidden representations are frozen over the course of training, the mean-field / μP scaling enables strong feature learning even at infinite width (Mei et al., 2019; Geiger et al., 2020; Yang & Hu, 2021; Bordelon & Pehlevan, 2022). In this feature-learning regime (often termed the *rich regime*), training substantially modifies hidden weights and representations rather than merely fitting a linear model on top of static random features (Ghorbani et al., 2020; Vyas et al., 2022; Montanari & Urbani, 2025).

A natural way to study the evolution of hidden features is through the spectrum of hidden weight matrices (Mahoney, 2022; Hodgkinson et al., 2025). At initialization, these spectra are well described by random matrix theory: the empirical singular value distribution forms a Marchenko-Pastur bulk (Marčenko & Pastur, 1967; Potters & Bouchaud, 2020). During training, structured updates can create isolated singular values that detach from the random bulk. These outliers encode low-dimensional directions shaped by learning, while the bulk continues to reflect high-dimensional randomness and finite-width effects. In random matrix theory, the emergence of isolated eigenvalues from the bulk as signal strength increases is the **BBP phase transition** (Baik et al., 2005).

This perspective is especially relevant for questions of width scaling and hyperparameter transfer. Feature learning parameterizations such as μP and CompleteP are designed so that training dynamics remain comparable across widths and depths, generating consistent optimal hyperparameters across model sizes (Yang & Hu, 2021; Yang et al., 2022; Dey et al., 2025). However, a complete theoretical

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.

account of hyperparameter transfer is still lacking in regimes where both (1) model losses vary significantly across widths and (2) optimal learning rates remain consistent in μP but not in other parameterizations, such as NTK-P. The edge-of-stability (EoS) literature (Xing et al., 2018; Jastrzebski et al., 2020; Cohen et al., 2021; 2022; Andreyev & Beneventano, 2025) suggests that the maximum stable learning rate is controlled by sharpness or the NTK eigenvalue (Jiang et al., 2026). Therefore, width-stable learning-rate transfer requires width-stable dynamics of the corresponding top spectral outlier Ghosh et al. (2025). This makes the evolution of the full spectrum, and in particular its leading isolated modes, a natural object for testing and explaining when hyperparameter transfer succeeds or fails. A recent proposal by Ghosh et al. (2025) suggests that feature learning proceeds through effectively low-rank weight updates whose dominant directions stabilize before the loss converges. In this work, we provide an exact theoretical characterization of these spectral dynamics.

BBP for μP? In this work, we develop a theoretical framework for tracking the dynamics of singular values of hidden weight matrices in wide, randomly initialized deep networks. We show that learning induces spike-like directions in the weights, which are in general *statistically coupled* to the random bulk from initialization. This places the problem outside the scope of standard spiked random matrix models, which typically assume independence between spikes and noise (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012; Perry et al., 2018; Capitaine, 2018; Barbier et al., 2023). To address this, we introduce a *two-level dynamical mean field theory (DMFT)* formalism that jointly describes the training dynamics generating the structured directions and the spectral probe used to resolve bulk and outlier singular values. The result is a general framework for spiked matrix ensembles with statistically dependent spikes, together with a finite-dimensional criterion that predicts when outliers escape from the bulk.

Our main contributions are:

- We develop a two-level DMFT framework for random matrices with finite-rank spikes statistically coupled to the random bulk. This yields a finite-dimensional condition for BBP-style outlier escape in trained weight spectra.
- For infinite-width nonlinear networks in mean-field/μP scaling, the theory yields layerwise predictions for spectral outliers and their BBP transitions as a function of learning richness. This finite-outlier picture accurately captures the main spectral features observed in wide convolutional ResNets trained on CIFAR-10 image classification.
- We characterize width-dependent spectral dynamics in deep linear networks and show that μP yields substantially more consistent BBP transition times and outlier dynamics across width than NTK parameterization, providing a mechanistic signature of successful transfer across model size.
- For tasks with extensive output dimension, including compute-optimal language-model pre-training and large-class vision classification, training can restructure the entire spectral bulk rather than produce only finitely many isolated outliers. We capture this departure from the finite-rank bulk-plus-outlier picture with a simple linear-network toy model with extensive output channels. This connects to the generalized BBP for extensive-rank spikes (Forner et al., 2025), here treated through DMFT for training-induced spikes and arbitrarily deep NNs.

1.1 Related Works

Infinite Limits and Hyperparameter Transfer The recent emphasis on neural scaling laws has made it increasingly important to understand which aspects of training are stable across model size and which are not (Kaplan et al., 2020; Hoffmann et al., 2022). On the theory side, infinite-width limits have provided a complementary lens: depending on parameterization, these limits can yield kernel descriptions, mean-field descriptions, or nontrivial feature-learning limits that remain analytically tractable (Jacot et al., 2018; Mei et al., 2019; Yang & Hu, 2021; Yang et al., 2022). Our work fits into this broader program, but focuses specifically on the evolution of hidden-layer weights and kernel spectra and on whether the signal-carrying spectral structure of training is preserved across width.

Regimes of Training Lazy training is well approximated by linearization around initialization (Jacot et al., 2018; Chizat et al., 2019), whereas rich training substantially changes representations (Woodworth et al., 2020; Mei et al., 2019; Yang & Hu, 2021). Parameterizations such as μP preserve feature learning at infinite width and enable hyperparameter transfer (Yang et al., 2022; Vyas et al., 2023). We contribute to this literature by showing that the evolution of the spectral bulk and isolated

outliers provides a mechanistic diagnostic of both the lazy-to-rich transition and the width consistency underlying hyperparameter transfer.

Dynamical Mean Field Theory and Random Matrix Theory Originating in spin-glass physics (Sompolinsky & Zippelius, 1981; De Dominicis, 1978), dynamical mean field theory (DMFT) describes high-dimensional learning dynamics, including SGD in generalized linear models (Mignacco et al., 2020; Gerbelot et al., 2022), feature learning in two-layer networks (Montanari & Urbani, 2025), and infinite-width/depth training limits (Bordelon & Pehlevan, 2022; Bordelon et al., 2024b;c). Random matrix theory (RMT) characterizes weight-covariance spectra at initialization through the Marchenko–Pastur law (Marčenko & Pastur, 1967), while spiked models and the BBP transition describe when low-rank perturbations generate isolated eigenvalues outside the bulk (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012; Baik & Silverstein, 2006). We combine these approaches in a *two-level DMFT* that tracks spectral evolution throughout training, accounting for the statistical dependence between learned spikes and the random initialization.

Weight Spectra and Scaling Laws Recent studies connect heavy-tailed weight spectra to implicit self-regularization (Martin & Mahoney, 2020; 2021; Thamm et al., 2022; Wang et al., 2023). Theoretical studies characterize spectra in shallow networks after one gradient step or under Bayesian inference (Dandi et al., 2023; Moniri et al., 2023; Dandi et al., 2024; Cui et al., 2024), and link spectral phases to generalization and scaling laws in quadratic networks (Defilippis et al., 2025; Erba et al., 2026) and teacher-student attention models (Boncoraglio et al., 2025). Recent DMFT work also tracks the dynamics of extensive-width quadratic networks (Martin et al., 2026), recovering static predictions at long times (Maillard et al., 2024). We develop a theory for how spectral structure emerges from random initialization and evolves throughout multi-step (S)GD in wide nonlinear networks of arbitrary depth.

Deep Linear Networks Our work connects to the tradition of using deep linear networks as tractable models of nonlinear optimization dynamics (Saxe et al., 2013; Kunin et al., 2024; Dominé et al., 2024). Our work extends prior DMFT analyses that enable study of randomly initialized deep linear networks in a proportional scaling (Bordelon & Pehlevan, 2025). We extend these results with our two-level DMFT to track spectral outliers throughout training.

2 Results

We first present a general characterization of the singular-value spectrum of random matrices with statistically coupled finite-rank structure.

2.1 Singular Values of Random Matrices with Statistically Coupled Spikes

Let $\boldsymbol{W}(S) \in \mathbb{R}^{N_1 \times N_0}$ be a random matrix initialized with mean-zero i.i.d. entries of variance σ^2 with S spikes

$$\boldsymbol{W}(S) = \boldsymbol{W}(0) + \frac{1}{\sqrt{N_0}} \sum_{t \in [S]} \boldsymbol{g}(t) \boldsymbol{\phi}(t)^\top. \tag{1}$$

Our general result assumes that the spike vectors $\boldsymbol{g}(t)$ and $\boldsymbol{\phi}(t)$ have asymptotically independent entries, as occurs naturally in large-width limits of neural networks. Crucially, however, the spikes are not independent of the initialization. Instead, they are coupled to $\boldsymbol{W}(0)$ through their dependence on the dynamical random vectors $\boldsymbol{\chi}(t) = \frac{1}{\sqrt{N_0}} \boldsymbol{W}(0) \boldsymbol{\phi}(t)$ and $\boldsymbol{\xi}(t) = \frac{\sqrt{N_0}}{N_1} \boldsymbol{W}(0)^\top \boldsymbol{g}(t)$, being

$$\text{Feature Structural Condition:} \quad \phi_i(t) = \phi_t\left(\{\xi_i(s)\}_{s \in [S]}\right) \ , \ g_i(t) = g_t\left(\{\chi_i(s)\}_{s \in [S]}\right) \tag{2}$$

where ϕ_t and g_t are asymptotically deterministic (possibly nonlinear) functions that take in S arguments $\{\chi_i(s)\}$ or $\{\xi_i(s)\}_{s \in [S]}$ [1]. We first review the dynamical mean field theory (DMFT) equations describing the limiting joint distribution of $\boldsymbol{\phi}(t), \boldsymbol{g}(t), \boldsymbol{\chi}(t), \boldsymbol{\xi}(t)$. Our predictions follow from a statistical-mechanics saddle-point analysis and are supported by numerical experiments; a rigorous mathematical proof is beyond the scope of this work. The correlation and response functions characterize the macroscopic dynamics in the high-dimensional limit.

[1] In this setting, the dependence on the variables χ, ξ will be causal so $\phi(t) = \phi_t(\{\xi(s)\}_{s<t})$.

Lemma 1: Generic DMFT for the Asymptotic Limit Dynamics

Assume proportional scaling $N_0, N_1 \to \infty$ with $N_1/N_0 = \alpha$ in a system that satisfies Equations 1 & 2 for constant spikes $S = \mathcal{O}(1)$. Each of the vectors $\boldsymbol{\phi}(t), \boldsymbol{g}(t), \boldsymbol{\chi}(t), \boldsymbol{\xi}(t)$ have i.i.d. entries which can be viewed as *single-site* stochastic processes which obey

$$\chi(t) = u^{\chi}(t) + \frac{\sigma^2}{\alpha} \sum_s R^{\phi}(t,s) g(s) \ , \ u^{\chi}(t) \sim \mathcal{N}\left(0, \sigma^2 C^{\phi}(t,s)\right) \tag{3}$$

$$\xi(t) = u^{\xi}(t) + \sigma^2 \sum_s R^{g}(t,s) \phi(s) \ , \ u^{\xi}(t) \sim \mathcal{N}\left(0, \sigma^2 \alpha^{-1} C^{g}(t,s)\right) \tag{4}$$

and $\phi(t) = \phi_t(\{\xi\})$ and $g(t) = g_t(\{\chi\})$ are functions of the above processes. The correlation functions C^{ϕ}, C^{g} and response functions R^{ϕ}, R^{g} are determined by averages $\langle \cdot \rangle$ over the $\{u^{\chi}, u^{\xi}\}$ Gaussian processes

$$C^{\phi}(t,s) = \langle \phi(t)\phi(s) \rangle \ , \ C^{g}(t,s) = \langle g(t) g(s) \rangle \tag{5}$$

$$R^{\phi}(t,s) = \left\langle \frac{\partial \phi(t)}{\partial u^{\xi}(s)} \right\rangle \ , \ R^{g}(t,s) = \left\langle \frac{\partial g(t)}{\partial u^{\chi}(s)} \right\rangle . \tag{6}$$

After solving the above DMFT equations for the correlation functions C^{ϕ}, C^{g} and response functions R^{ϕ} and R^{g}, one is now in a position to compute the singular values of the weight matrix $\boldsymbol{W}(S)$. Below, we characterize the singular values in a large system limit with $S = \mathcal{O}_N(1)$ held fixed.

Result 1: Singular Values of Random Matrices with Statistically Coupled Spikes

The limiting eigenvalue density $\rho_{\boldsymbol{M}}(\lambda)$ of $\boldsymbol{M} = \frac{1}{N_1} \boldsymbol{W}(S)^{\top} \boldsymbol{W}(S)$ has Bulk + Spikes form

$$\rho_{\boldsymbol{M}}(\lambda) \to \rho_{\text{MP}}(\lambda; \alpha, \sigma) + \rho_{\text{spike}}(\lambda) \ , \quad N_0 \rho_{\text{spike}}(\lambda) \to \sum_{z \in O_S} \delta(\lambda - z) \tag{7}$$

where $\rho_{\text{MP}}(\lambda; \alpha, \sigma)$ is the Marchenko-Pastur eigenvalue density with aspect ratio α and scale σ. The outlier set $O_S = \{z \in \mathbb{R}_+ : \det \boldsymbol{A}(z) = 0\}$ are solutions to a zero-determinant condition for a matrix valued function $\boldsymbol{A}(z) \in \mathbb{R}^{4S \times 4S}$

$$\boldsymbol{A}(z) = \begin{bmatrix} \mathcal{G}(z)^{-1} \boldsymbol{I} & \boldsymbol{0} & -(\sigma^2 \alpha^{-1} \boldsymbol{R}^{\phi} + \boldsymbol{C}^{\phi}) & -\sigma^2 \boldsymbol{C}^{\phi} \\ \boldsymbol{0} & \mathcal{G}(z)^{-1} \boldsymbol{I} & -(\boldsymbol{R}^{\phi})^{\top} & -\sigma^2 (\boldsymbol{R}^{\phi})^{\top} \\ -(\sigma^2 \boldsymbol{R}^{g} + \boldsymbol{C}^{g}) & -\sigma^2 \alpha^{-1} \boldsymbol{C}^{g} & (1 - \frac{\sigma^2}{\alpha} \mathcal{G}(z)) \boldsymbol{I} & \boldsymbol{0} \\ -(\boldsymbol{R}^{g})^{\top} & -\sigma^2 \alpha^{-1} (\boldsymbol{R}^{g})^{\top} & \boldsymbol{0} & (1 - \frac{\sigma^2}{\alpha} \mathcal{G}(z)) \boldsymbol{I} \end{bmatrix} \tag{8}$$

where $\mathcal{G}(z) = \frac{\alpha}{2z\sigma^2} \left[z + \sigma^2(\alpha^{-1} - 1) - \sqrt{[z + \sigma^2(\alpha^{-1} - 1)]^2 - 4z\alpha^{-1}\sigma^2} \right]$ is the Stieltjes transform of the Wishart matrix and the matrices $\boldsymbol{R}^{\phi}, \boldsymbol{R}^{g}, \boldsymbol{C}^{\phi}, \boldsymbol{C}^{g} \in \mathbb{R}^{T \times T}$ are the correlation and response functions of the DMFT for the original system.

Buildup of Spike(s) Increasing the number of spikes S reveals the outlier dynamics. At initialization ($S = 0$), the outlier set O_S is empty. As updates to $\boldsymbol{W}$ accumulate, multiple outliers (up to S) can emerge from the bulk through a BBP-like transition (Baik et al., 2005), although the dynamics typically maintains a low-dimensional spike subspace. Since S usually grows linearly with time steps T, we define the **BBP time** as the first T for which O_S is non-empty, when the accumulated updates are strong enough to produce a detectable outlier. We illustrate this mechanism with toy models in Appendices A and B, and a closed-form one-step example in Appendix C.1.

This result covers two regimes studied below: infinite-width feature-learning networks at fixed batch size and training time (Section 3), and randomly initialized deep linear networks in a proportional limit of input dimension, sample size, and width (Section 4).

3 Infinite Width Limits of Feature Learning Networks

Consider an L-hidden-layer fully connected network with width N, input dimension D, parameters $\boldsymbol{\theta} = \text{Vec}\{\boldsymbol{W}^\ell\}_{\ell=0}^L$, hidden states $\boldsymbol{h}_\mu^\ell$, and output f_μ

$$f_\mu = \frac{1}{\gamma N}\boldsymbol{w}^L \cdot \phi(\boldsymbol{h}_\mu^L), \qquad \boldsymbol{h}_\mu^{\ell+1} = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell \phi(\boldsymbol{h}_\mu^\ell), \qquad \boldsymbol{h}_\mu^1 = \frac{1}{\sqrt{D}}\boldsymbol{W}^0 \boldsymbol{x}_\mu. \tag{9}$$

Here ϕ is applied elementwise and γ is an output multiplier controlling the richness of the dynamics: $\gamma \to 0$ yields a kernel/lazy limit, while the $1/(\gamma N)$ readout scaling corresponds to the mean-field/μP scaling and preserves feature learning as $N \to \infty$ (Chizat et al., 2019; Lee et al., 2019; Geiger et al., 2020; Yang & Hu, 2021; Bordelon & Pehlevan, 2022; 2023). The parameters are initialized i.i.d. with variance σ^2 and trained by SGD on minibatches $\mathcal{B}_t = \{\boldsymbol{x}_\mu\}_{\mu=1}^B$: $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta\gamma^2 N \frac{\partial}{\partial \boldsymbol{\theta}(t)} \frac{1}{B}\sum_{\boldsymbol{x}\in\mathcal{B}_t} \ell(f(\boldsymbol{x}))$.

Super-Wide Networks We first assume that the width N is diverging while the input dimension D, batchsize B and number of timesteps t are held fixed. These constraints enable exact concentration of the network's prediction and loss dynamics and also perfect interpolation when applied on finite datasets (Jacot et al., 2018; Bordelon & Pehlevan, 2022). This regime has been shown to capture the behavior of wide convolutional neural networks on simple tasks like CIFAR-10, but it is overly optimistic about model performance in the scaling law regime where total data is large compared to model size (Vyas et al., 2023; Hoffmann et al., 2022; Bordelon et al., 2024a). We will relax this constraint in Section 4, where finite $\frac{Bt}{N}$ and $\frac{D}{N}$ effects can be characterized in linear networks.

Lemma 2: DMFT For Infinite Width Neural Networks in μP Trained with (S)GD

Infinite width networks of any depth L in μP trained with SGD for $O_N(1)$ steps on $O_N(1)$ batchsize satisfies the structural assumptions needed for our Bulk + Outlier theory with $S = BT$ spikes. The signal strength of the spikes are controlled by output multiplier γ.

Spike Structure in Nonlinear Networks In this setting, the spike variables are $\boldsymbol{\phi}_\mu^\ell(t) \equiv \phi(\boldsymbol{h}_\mu^\ell(t))$ and $\boldsymbol{g}_\mu^\ell(t) \equiv N\gamma \frac{\partial f_\mu(t)}{\partial \boldsymbol{h}_\mu^\ell(t)}$ are the forward and backward pass features respectively for all data in each minibatch $\boldsymbol{x}_\mu \in \mathcal{B}_t$. The dependence on the initial condition for weights occurs through

$$\boldsymbol{\chi}_\mu^{\ell+1}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\boldsymbol{\phi}_\mu^\ell(t)\ ,\ \boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t) \tag{10}$$

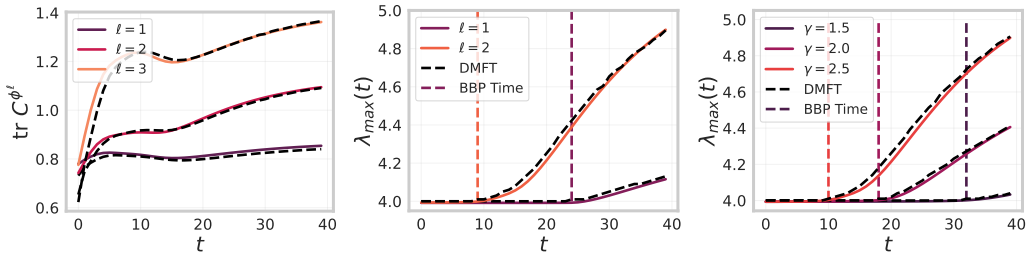
$$\boldsymbol{\phi}_\mu^\ell(t) \sim_{N\to\infty} \phi_{\mu,t}^\ell\left(\{\boldsymbol{\chi}_\nu^\ell(s), \boldsymbol{\xi}_\nu^\ell(s)\}_{\nu;s<t}; \gamma\right)\ ,\ \boldsymbol{g}_\mu^\ell(t) \sim_{N\to\infty} g_{\mu,t}^\ell\left(\{\boldsymbol{\chi}_\nu^\ell(s), \boldsymbol{\xi}_\nu^\ell(s)\}_{\nu;s<t}; \gamma\right) \tag{11}$$

where the functions $\phi_{\mu,t}^\ell$ and $g_{\mu,t}^\ell$ only act *elementwise* over the neurons. Further, each neuron statistically decouples as an i.i.d. random variable whose distribution can be identified from the DMFT. We give the explicit formulas for $\boldsymbol{\phi}_\mu^\ell(t), \boldsymbol{g}_\mu^\ell(t)$ spikes in the Appendix D, which allows Result 1 to be applied to nonlinear networks (see as well Bordelon & Pehlevan (2022; 2023)). In Figure 1 we illustrate our theory of the spike dynamics for a depth 4 ($L = 3$ hidden layer) $\tanh$ network trained on a single index model $y = q(\boldsymbol{\beta}\cdot\boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^{10}$ and $q(z) = z + \frac{1}{2}z^2$. The DMFT predicts the feature correlation functions, as well as the maximum eigenvalue dynamics through Result 1.

Bulk+Outlier Picture Holds in μP CNNs on CIFAR-10 DMFT can in principle be applied to architectures such as CNNs, recurrent networks, and transformers (Bordelon & Pehlevan, 2022; Clark et al., 2026; Bordelon et al., 2024b), but solving the exact outlier equations on realistic tasks is often computationally hard. We therefore test the bulk+outlier picture empirically in μP CNNs trained on CIFAR-10. As shown in Figure 2, across large channel widths N, a finite number of spikes detach from an otherwise stable bulk after thousands of training steps, consistent with a BBP-like transition in the *super-wide* regime.

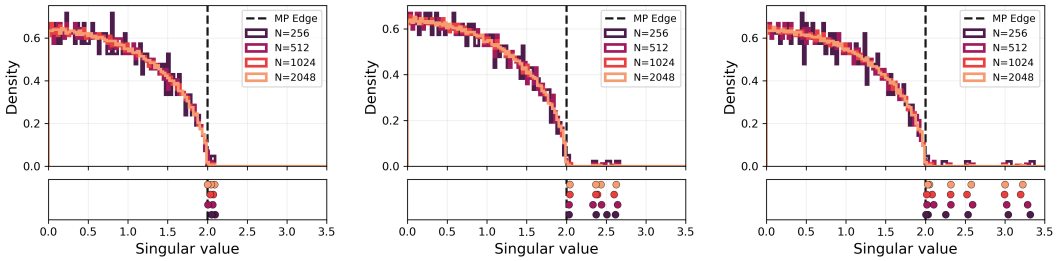
4 Linear Networks That Incorporate Width / Data Effects

In the previous section, we saw that the theory for nonlinear networks could accurately predict outlier dynamics for *super-wide networks*. However, this regime is incapable of characterizing variations in training dynamics across network widths.



(a) Feature Correlation Dynamics (b) BBP Transition $\gamma = 2.5$ (c) Last Layer $\ell = 2$

Figure 1: Deep ($L = 3$ hidden layers) nonlinear networks ($\phi = \tanh$) trained with gradient descent on a single-index polynomial target function with unit scale initialization $\sigma^2 = 1$. (a) The hidden feature kernel dynamics are predicted by the DMFT equations (dashed black lines). (b)-(c) The outlier dynamics of the hidden weights are accurately predicted by the $\boldsymbol{A}(z)$ matrix.



(a) $T = 100$ (b) $T = 250$ (c) $T = 1000$

Figure 2: Weight singular values of varying width N ResNets (second convolution in the first residual block) on CIFAR-10 at different snapshots of training. Consistent with the infinite width theory, the MP bulk is unshifted, while the outliers undergo deterministic (and asymptotically width-independent) dynamics. Surprisingly even width $N = 256$ models exhibit similar dynamics to $N = 2048$.

Proportionally Wide Linear Networks To access regimes where width can vary in a meaningful way that impacts performance and dynamics, we next turn to a setting where we can characterize networks that are wide but where width, input dimension, and data are comparable. This setting is challenging for general architectures and data distributions, but can be analyzed for *random and high dimensional data* in deep linear networks (see App. E or Bordelon & Pehlevan (2025)).

Lemma 3: DMFT For Proportional Width Linear Networks in μP Trained with (S)GD

Consider a randomly initialized width-N linear network with $C = O(1)$ output channels and linear targets $y_c(\boldsymbol{x}) = D^{-1/2}\boldsymbol{\beta}_c \cdot \boldsymbol{x} + \epsilon_c$. We study either full-batch GD on P random inputs in D dimensions or online SGD with batch size B, in the proportional limits

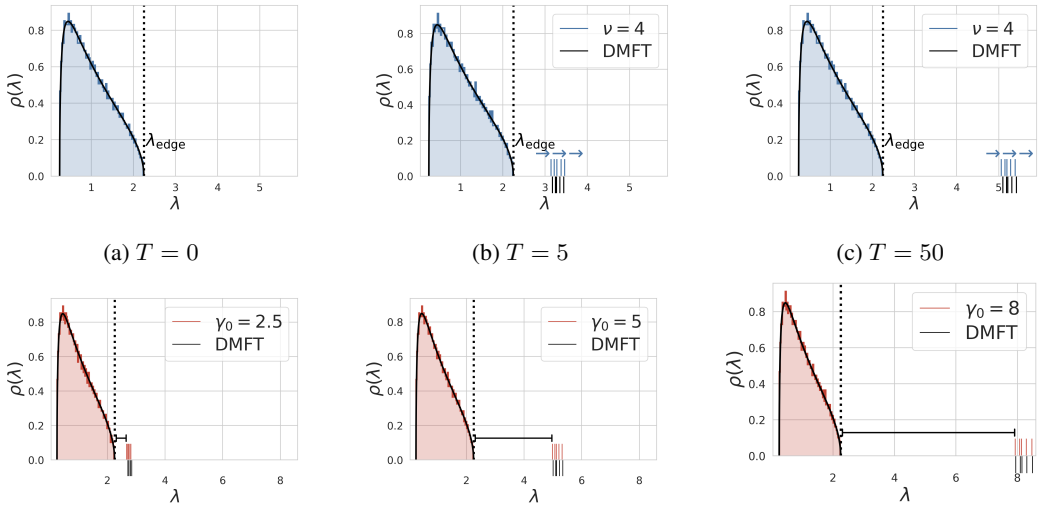
$$\begin{cases} P, N, D \to \infty, & \alpha = P/D, \quad \nu = N/D, \quad S = CT, & \text{full-batch GD}, \\ B, N, D \to \infty, & \alpha = B/D, \quad \nu = N/D, \quad S = CT, & \text{online SGD}. \end{cases} \tag{12}$$

(S)GD satisfies the conditions of Result 1, and the loss and weight dynamics are characterized by self-averaging correlation and response $\{C^{\phi}, C^{g}, R^{\phi}, R^{g}\}$ for each hidden layer $\ell \in [L]$.

The weight updates are formed as sums over both time steps $t \in [T]$ and class indices $c \in [C]$

$$\boldsymbol{W}^{\ell}(T) = \boldsymbol{W}^{\ell}(0) + \frac{\eta\gamma_0}{\sqrt{N}} \sum_{t=1}^{T} \sum_{c=1}^{C} \boldsymbol{g}_c^{\ell+1}(t) \boldsymbol{\phi}_c^{\ell}(t)^{\top}. \tag{13}$$

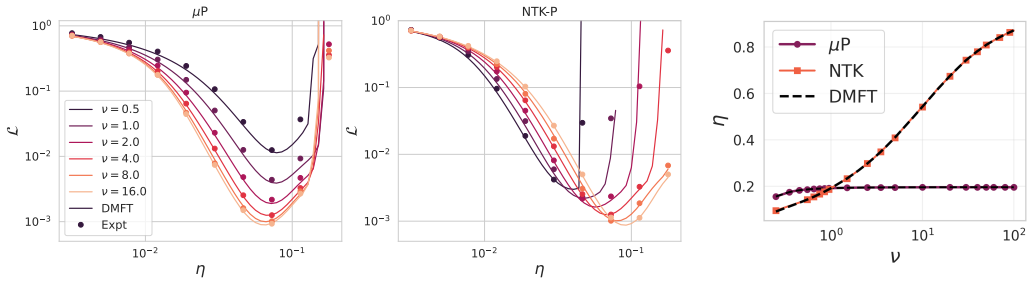
The DMFT equations provide the necessary description of the random (Gaussian in this case) fields $\boldsymbol{g}_c^{\ell}(t) \sim g_{c,t}^{\ell}(\{\boldsymbol{\chi}_{c'}^{\ell}(s), \boldsymbol{\xi}_{c'}^{\ell}(s)\}_{s<t})$ and $\boldsymbol{\phi}_c^{\ell}(t) \sim \phi_{c,t}^{\ell}(\{\boldsymbol{\chi}_{c'}^{\ell}(s), \boldsymbol{\xi}_{c'}^{\ell}(s)\}_{s<t})$, where $\sim$ indicates equivalence in the joint scaling limit and the functions $\phi_{c,t}^{\ell}(\cdot)$ and $g_{c,t}^{\ell}(\cdot)$ act elementwise over the



(a) $T = 0$ (b) $T = 5$ (c) $T = 50$

(d) $T = 100, \nu = 4$ (e) $T = 100, \nu = 4$ (f) $T = 100, \nu = 4$

Figure 3: Spectral density $\rho(\lambda)$ of the weight covariance of a $L = 2$ linear network. (a) At initialization ($T = 0$), the spectrum follows a Marchenko–Pastur (MP) distribution. (b)-(c) As training progresses under GD, the bulk remains unchanged, while a small number of eigenvalues detach from the bulk and evolve as outliers. (d)-(f) The richness γ_0 enhances the scale of the outliers.



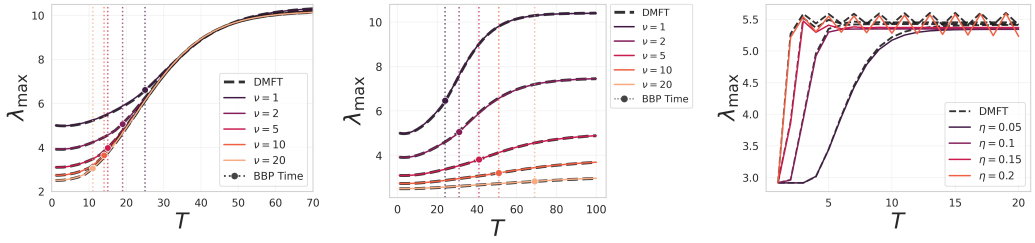
(a) Learning rate transfer under SGD with DMFT Prediction (b) $\eta_{\max}$

Figure 4: (a) DMFT can predict the success and failure of learning rate transfer across widths ν in the proportionally wide linear networks. (b) The maximum stable learning rate as a function of ν is weakly varying in μP and is predicted by the NTK top eigenvalue (predicted from our outlier theory).

vectors. Figure 3 shows C outliers emerging with training, accurately predicted by our theory. Neglecting spike–initialization correlations underestimates or misses outliers (see App. E.12.3). Similarly, Figure 3 illustrates at a fixed number of steps T, the impact of the output multiplier γ_0 on the strength of the outlier spike. Figures 4 and 5 connect these outlier dynamics to learning-rate transfer by showing that the maximal stable learning rate is controlled by the top NTK mode: in μP, this is nearly width-stable, whereas in NTK-P it remains strongly ν-dependent.

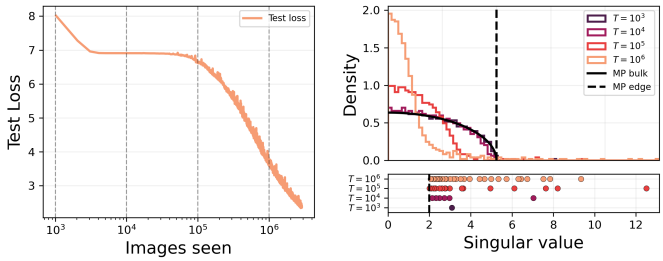
5 Extensive Output Channels: Beyond Constant Bulk + Dynamic Spikes

Although Result 1 captures the finite-outlier regime observed in simple models and CIFAR-10 CNNs, larger-output tasks exhibit a different spectral phenomenology. We first observe this departure in an ImageNet-scale ResNet18, where training modifies the bulk of the hidden-weight spectrum rather than producing only a few isolated outliers (Fig. 6). The same effect appears more prominently in GPT-style language models trained on next-token prediction on the C4 dataset (Fig. 7), where the output vocabulary is much larger than the hidden width. These observations motivate a large-output-channel regime in which training restructures an extensive part of the spectrum.



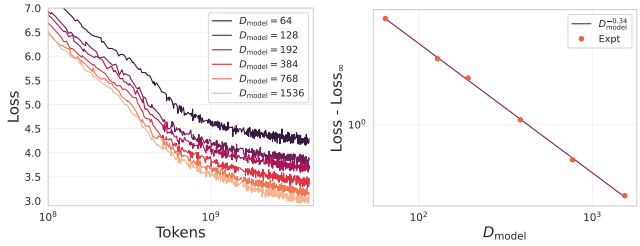
(a) μP (b) NTK-P (c) Weight covariance eigenvalue

Figure 5: Progressive sharpening of the top NTK mode in a $L = 1$ linear network. (a,b) Evolution of the top NTK mode $\lambda_{\max}$ for different aspect ratios ν and parameterizations. In μP, the post-BBP dynamics is nearly ν-independent, whereas in NTK-P it remains strongly ν-dependent and approaches a lazy large-ν limit. (c) $\lambda_{\max}(T)$ vs learning rates η: larger η speeds up sharpening until EoS.

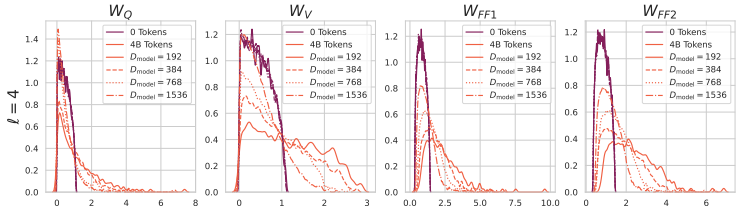


(a) Loss Dynamics (b) Weight Spectra vs T

Figure 6: Width-256 μP ResNet18 on ImageNet. Test loss drops around 10^5 images seen, while the MP-normalized spectrum of the first hidden-layer ResNet block deforms: the bulk shifts left and a right tail emerges beyond the MP edge.

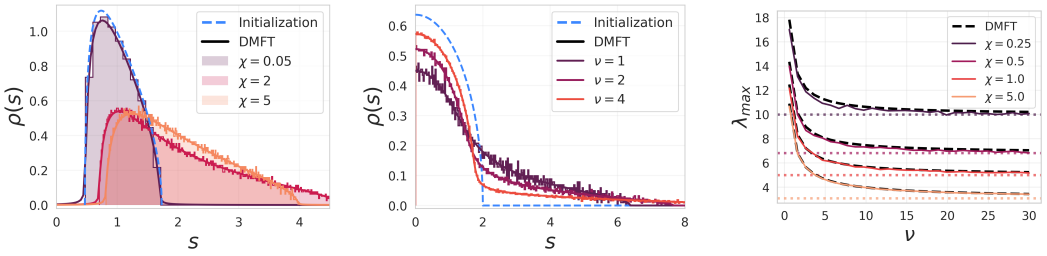


(a) Loss Dynamics Across Widths (b) Non-trivial Scaling Exponent



(c) Weight Spectra Before and After Training

Figure 7: Losses and spectral densities of weights before and after language model pretraining in μP on 4B tokens on the C4 dataset in models of varying widths $D_{\text{model}} = 64 \times$ heads. The models range from $H = 3$ heads to $H = 24$ heads with $L = 6$ layers and spectra are plotted before and after training. The spectral densities show a bulk reshaping, falling outside the scope of Result 1.



(a) $\gamma_0 = 1$ (b) $\gamma_0 = 2$ (c) $\gamma_0 = 1$

Figure 8: Large-output spectral densities. (a) Two-layer bulk shifts away from MP as χ increases. (b) Three-layer hidden-weight spectrum sharpens with width ν. (c) $\lambda_{\max}$ stabilizes across width and approaches the shifted-MP plateau $1 + \gamma_0^2(1 + \chi^{-1/2})^2$.

5.1 Candidate Theory: Large Output Channels

Motivated by the fact that the output vocabulary size for language models is often larger than the width (in Fig. 7 the vocabulary is $\sim$ 50k and the width for $H = 12$ is $N = 768$), we consider an extension of our linear network theory where the output channels C is of the same order as width N.

Result 2: Linear Network with Extensive Classes at One Step

Consider randomly initialized width-N linear network with jointly diverging output dimension C and input dimension D. For $\boldsymbol{M}^\ell = \frac{1}{N}\boldsymbol{W}^\ell(T)^\top \boldsymbol{W}^\ell(T)$, we study the joint limit

$$N, P, D, C \to \infty, \qquad \alpha = P/D, \quad \nu = N/D, \quad \chi = C/D, \tag{14}$$

in which the singular-value density converges as $\rho_{\boldsymbol{M}^\ell}(s) \to \rho^\ell(s; \alpha, \nu, \chi, \gamma_0)$. For $\gamma_0, \chi > 0$, this limiting bulk density is generally not Marchenko–Pastur. In the large $\nu, \alpha \to \infty$ limit, the leading finite-ν correction to the top eigenvalue for a two-layer model approaches

$$\lambda_{\max} = \lambda_{\max}^{\infty} + \frac{1}{\nu}\,\lambda_{\max}^{\infty}\left(\sqrt{\chi} + \frac{\chi}{\gamma_0^2(1+\sqrt{\chi})}\right) + \mathcal{O}(\nu^{-2}) \tag{15}$$

so, in the extensive-output regime, width-consistent $\lambda_{\max}$ can still occur if correction is small.

In the population limit $\alpha \to \infty$, one GD step gives, up to corrections changing $\boldsymbol{M}^\ell$ by $\mathcal{O}(D^{-1/2})$ in operator norm and thus preserving its limiting spectral density in the asymptotic limit (see Appendix F)

$$\boldsymbol{W}^\ell(1) = \boldsymbol{W}^\ell(0) + \frac{\eta\gamma_0}{\sqrt{CD}}\boldsymbol{G}^{\ell+1}(0)\boldsymbol{B}^\star\boldsymbol{H}^\ell(0)^\top. \tag{16}$$

Here $\boldsymbol{H}^\ell \in \mathbb{R}^{N\times D}$, $\boldsymbol{G}^{\ell+1} \in \mathbb{R}^{N\times C}$, and $\boldsymbol{B}^\star \in \mathbb{R}^{C\times D}$ denote the forward feature map, back-propagated readout, and target, respectively. These matrices are independent and asymptotically free at initialization Potters & Bouchaud (2020). Extending the analysis beyond one step requires tracking interlayer correlations and is left for future work. We illustrate the spectra across χ, γ_0, ν in Fig. 8.

6 Discussion

We developed a method to compute the dynamics of spectral outliers for spiked matrix ensembles where the spikes have a causal dependence on the random component of the matrix. This method enabled us to characterize the outlier dynamics of weights in (1) super wide nonlinear networks and (2) proportional width linear networks. We showed that this theory describes some regimes of network training, such as CNNs on CIFAR-10 but fails to capture others such as language model pretraining, for which we propose a simple toy model to account for an extensive number of classes.

Limitations The theory applies to (1) nonlinear networks in the super-wide regime, where a constant bulk coexists with finitely many outliers, and (2) proportionally wide linear networks with finitely many output channels. Outlier computation scales as $\mathcal{O}(S^3)$ in the number of spike directions. Our extensive-output analysis is currently restricted to a single GD step; extending it to full training is left for future work.

AI Use Statement. In this work, we used generative AI tools to help simplify the expressions in App E.15.1 in a more compact manner, which provided "critical ingredients for proving mathematical claims". In this work, we have **not** used generative AI tools for:

- help develop theoretical models or conceptual frameworks
- formulate mathematical claims
- assist in the writing of proofs
- propose or refine hypotheses
- design or provide feedback on research methodology or experiments
- implement methods

The following are not applicable to this work:

- generate synthetic data sets
- assist with translation
- clean and reformat dataset
- support qualitative and thematic data analysis
- interpret results

Additionally, we used generative AI tools for create and edit figures, write code, and to edit and polish the manuscript. We take responsibility for the final content of this work, including text, claims and figures produced with the aid of generative AI.

Reproducibility Statement. We detail the experimental settings in Appendix G and provide code to reproduce the figures in the supplementary material.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd, 2025. URL https://arxiv.org/abs/2412.20553.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

Jean Barbier, Francesco Camilli, Marco Mondelli, and Manuel Sáenz. Fundamental limits in structured principal component analysis and how to reach them. *Proceedings of the National Academy of Sciences*, 120(30):e2302028120, 2023.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2025.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *arXiv preprint arXiv:2205.09653*, 2022.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *arXiv preprint arXiv:2304.03408*, 2023.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. *arXiv preprint arXiv:2502.02531*, 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. *arXiv preprint arXiv:2402.01092*, 2024a.

Blake Bordelon, Hamza Tahir Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *arXiv preprint arXiv:2405.15712*, 2024b.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=KZJehvRKGD.

Mireille Capitaine. Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. In *Séminaire de Probabilités XLIX*, pp. 119–164. Springer, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

David G. Clark and L. F. Abbott. Theory of coupled neuronal-synaptic dynamics. *Phys. Rev. X*, 14:021001, Apr 2024. doi: 10.1103/PhysRevX.14.021001. URL https://link.aps.org/doi/10.1103/PhysRevX.14.021001.

David G Clark, Blake Bordelon, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Structure, disorder, and dynamics in task-trained recurrent neural circuits. *bioRxiv*, pp. 2026–03, 2026.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Jeremy M. Cohen, B. Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David E. Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability. *ArXiv*, abs/2207.14484, 2022. URL https://api.semanticscholar.org/CorpusID:251196732.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue M Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=iBDcaBLhz2.

Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue M Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. *arXiv preprint arXiv:2410.18938*, 2024.

C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2025.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*, 2025.

Clémentine CJ Dominé, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro AM Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv preprint arXiv:2409.14623*, 2024.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks. *Advances in Neural Information Processing Systems*, 38:88862–88901, 2026.

Niklas Forner, Alexander Maloney, and Bernd Rosenow. Bbp phase transition for an extensive number of outliers. *arXiv preprint arXiv:2511.18501*, 2025.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (11):113301, 2020.

Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. *arXiv preprint arXiv:2512.22768*, 2025.

Liam Hodgkinson, Zhichao Wang, and Michael W. Mahoney. Models of heavy-tailed mechanistic universality. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=r4XsTcuiqc.

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, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g87C4KwB.

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=QKo4c3LAz3.

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.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *arXiv preprint arXiv:2406.06158*, 2024.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Michael W Mahoney. Practice, theory, and theorems for random matrix theory in modern machine learning. 2022.

Antoine Maillard, Emanuele Troiani, Simon Martin, Lenka Zdeborová, and Florent Krzakala. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, 2024.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Charles H Martin and Michael W Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 505–513. SIAM, 2020.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464. PMLR, 2019.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024. URL https://arxiv.org/abs/2501.00656.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252, 2015.

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.

Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5):054124, 2022.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Limitations of the ntk for understanding generalization in deep learning. *arXiv preprint arXiv:2206.10012*, 2022.

Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *arXiv preprint arXiv:2305.18411*, 2023.

Zhichao Wang, Andrew Engel, Anand D Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in neural information processing systems*, 36:20695–20728, 2023.

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.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd, 2018. URL https://arxiv.org/abs/1802.08770.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

A Classical BBP Transition: Wigner + Rank-One Spike

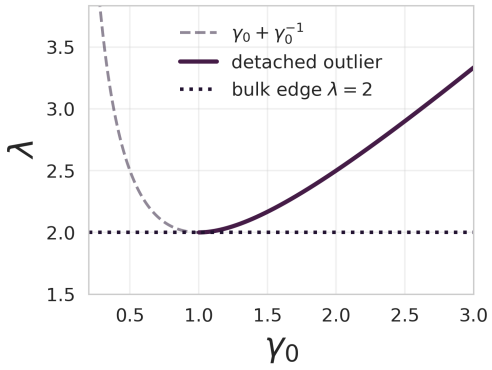


Figure 9: Classical BBP transition in the rank-one spiked Wigner model. The outlier location at $\lambda = \gamma_0 + \gamma_0^{-1}$ detaches from the Wigner bulk edge $\lambda = 2$ for $\gamma_0 > 1$.

As a warm-up, we consider the standard rank-one spiked Wigner model

$$\boldsymbol{M} = \boldsymbol{M}_0 + \frac{\gamma_0}{D}\boldsymbol{v}\boldsymbol{v}^\top, \qquad \|\boldsymbol{v}\|^2 = D \tag{17}$$

where $\boldsymbol{M}_0$ is a GOE matrix. The goal is to recover the classical BBP outlier condition using the same response-function method that we later apply to training-induced spikes. This example illustrates the main idea in the simplest setting: the normalized trace response gives the random-matrix bulk, while the projection of the linear probe onto the spike direction detects isolated eigenvalues.

We introduce an auxiliary spectral-time dynamics

$$\begin{aligned}\partial_\tau \boldsymbol{h}(\tau) &= -\boldsymbol{M}\boldsymbol{h}(\tau) \\ &= -\boldsymbol{M}_0\boldsymbol{h}(\tau) - \gamma_0 C_{vh}(\tau)\boldsymbol{v}\end{aligned} \tag{18}$$

where

$$C_{vh}(\tau) = \frac{1}{D}\boldsymbol{v}\cdot\boldsymbol{h}(\tau). \tag{19}$$

The trace response of this linear system is

$$R_h(\tau,\tau') = -\frac{i}{D}\boldsymbol{h}(\tau)\cdot\hat{\boldsymbol{h}}(\tau') \tag{20}$$

and its Fourier transform gives the resolvent of the Wigner bulk.

Averaging over the GOE disorder gives the usual single-site DMFT equation

$$\partial_\tau h(\tau) = u_h(\tau) + \int d\tau'\, R_h(\tau,\tau')h(\tau') - \gamma_0 C_{vh}(\tau)v \tag{21}$$

with Gaussian noise

$$u_h \sim \mathcal{N}(0, C_h). \tag{22}$$

In Fourier space

$$h(\omega) = \left[i\omega - R_h(\omega)\right]^{-1}\left(u_h(\omega) - \gamma_0 C_{vh}(\omega)v\right). \tag{23}$$

The bulk response is therefore the GOE resolvent

$$R_h(\omega) = \frac{i\omega - \sqrt{(i\omega)^2 - 4}}{2} \tag{24}$$

which captures the continuous Wigner bulk.

To detect outliers, we project the probe dynamics onto the spike direction. The projected correlation $C_{vh}(\omega)$ satisfies

$$C_{vh}(\omega) = -\gamma_0\left[i\omega - R_h(\omega)\right]^{-1} C_{vh}(\omega). \tag{25}$$

Equivalently, allowing for an initial overlap C_0,

$$C_{vh}(\omega) = \frac{C_0}{i\omega + \gamma_0 - R_h(\omega)}. \tag{26}$$

Thus isolated eigenvalues correspond to poles of the projected resolvent, i.e. to zeros of

$$i\omega + \gamma_0 - R_h(\omega) = 0. \tag{27}$$

Analytically continuing to the real spectral axis with $i\omega = -\lambda - i0^+$, the outlier condition becomes

$$-\lambda + \gamma_0 = \frac{-\lambda + \sqrt{\lambda^2 - 4}}{2}. \tag{28}$$

Solving gives the classical spiked-Wigner outlier location

$$\lambda_{\text{out}} = \gamma_0 + \frac{1}{\gamma_0}. \tag{29}$$

This solution lies outside the Wigner bulk only when $\gamma_0 > 1$, giving the usual BBP threshold. Thus the toy model demonstrates the two ingredients used throughout the paper: the trace response determines the bulk, while the spike-probe correlation reveals outliers through their poles.

B ANTI-HEBBIAN WIGNER DYNAMICS

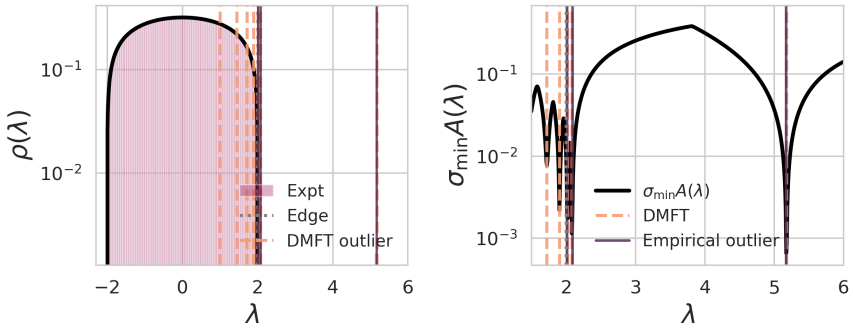


Figure 10: Anti-Hebbian Wigner toy model with dynamically generated spike directions, shown after $T = 200$ discretized physical-time steps with step size $\Delta t = 0.05$. Left: empirical eigenvalue density of $\boldsymbol{M}(T)$ compared with the GOE semicircle bulk; vertical lines mark the empirical outliers and the DMFT predictions from the projected resolvent condition. Right: minimum singular value of the finite-dimensional matrix $\boldsymbol{A}(\lambda)$, whose zeros give the outlier condition. The dips align with the empirical detached eigenvalues, showing that the two-level DMFT captures outliers generated by spike directions statistically coupled to the random bulk.

We next consider a toy model in which the spike directions are not planted independently of the random matrix, but are generated dynamically by the random matrix itself. This makes explicit the statistical dependence between the random bulk and the learned spike directions, which is the key difficulty in trained neural networks. Consider D degrees of freedom $v_i(t)$ with $i \in \{D\}$, evolving under a time-dependent symmetric coupling matrix $\boldsymbol{M}(t)$

$$\partial_t \boldsymbol{v}(t) = -\boldsymbol{M}(t)\boldsymbol{v}(t). \tag{30}$$

The coupling matrix $\boldsymbol{M}(t)$ is decomposed into a quenched random part and a plastic part

$$\boldsymbol{M}(t) = \boldsymbol{M}_0 + \boldsymbol{A}(t) \tag{31}$$

where $\boldsymbol{M}_0$ is drawn from the GOE ensemble $\langle (\boldsymbol{M}_0)_{ij} \rangle = 0$ and $\langle (\boldsymbol{M}_0)_{ij}(\boldsymbol{M}_0)_{kl} \rangle = \frac{1}{D}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$, while the plastic component is generated by the trajectory itself through the local anti-Hebbian rule (similarly to Clark & Abbott (2024))

$$\partial_t \boldsymbol{A}(t) = \frac{\gamma_0}{D}\boldsymbol{v}(t)\boldsymbol{v}(t)^\top. \tag{32}$$

At a fixed snapshot of time t, we want to characterize the spectrum of $\boldsymbol{M}(t)$. This has two parts: the $\mathcal{O}(D)$-dimensional bulk, which is described by the normalized resolvent $G_t(z) = \frac{1}{D}\mathrm{Tr}(z - \boldsymbol{M}(t))^{-1}$, and a rank$(t)$ possible isolated eigenvalues generated by the finite-rank deformation. Given $t = \mathcal{O}(1)$, the contribution of these eigenvalues to the normalized trace has weights $\mathcal{O}(D^{-1})$, and disappears from the large D limit of $G_t(z)$. To see the outliers, we must resolve the finite-dimensional subspace spanned by the learned directions $\{\boldsymbol{v}(s) : 0 \le s \le t\}$.

As before, we thus introduce a fictitious *spectral* time τ and a linear probe $\boldsymbol{h}(\tau)$ evolving under the random matrix

$$\partial_\tau \boldsymbol{h}(\tau) = -\boldsymbol{M}_0\boldsymbol{h}(\tau) - \gamma_0 \int_0^t dt' C_{vh}(t', \tau)\boldsymbol{v}(t') + \boldsymbol{j}(\tau). \tag{33}$$

The two objects we wish to track are

$$C_{vh}(t, \tau) = \frac{1}{D}\boldsymbol{v}(t) \cdot \boldsymbol{h}(\tau), \quad R_h(\tau, \tau') = \frac{1}{D}\mathrm{Tr}\frac{\partial \boldsymbol{h}(\tau)}{\partial \boldsymbol{j}(\tau')^\top} \tag{34}$$

the cross correlation $C_{vh}(t, \tau)$ between the structured direction $\boldsymbol{v}(t)$ and the linear probe $\boldsymbol{h}(\tau)$, which has the role of a resolvent projected onto the learned subspace, and $R_h(\tau, \tau')$ which is the trace

response of the spectral probe, hence the Stieltjes transform of the empirical spectrum. In the large D limit, $R_h(\tau, \tau')$ encodes the bulk, while the poles of $C_{vh}(t, \tau)$ give the outliers.

We thus end up with a two-level dynamical mean field theory (DMFT) description. The first level describes the training dynamics of $\boldsymbol{v}(t)$ evolving under the physical time t

$$\partial_t \boldsymbol{v}(t) = -\boldsymbol{M}_0 \boldsymbol{v}(t) - \gamma_0 \int_0^t C_v(t, t') \boldsymbol{v}(t') dt \tag{35}$$

which generates the spike subspace, while the second level refers to the linear system of Eq. equation 33 which probes the spectrum of the random matrix $\boldsymbol{M}(t)$.

B.1 Average over the quenched disorder

Computing the average over the GOE $\boldsymbol{M}_0$ matrix

$$\begin{aligned}
&\left\langle \exp\left(i \mathrm{Tr} \boldsymbol{M}_0 \left(\int dt \boldsymbol{v}(t) \hat{\boldsymbol{v}}(t)^\top + \int d\tau \boldsymbol{h}(\tau) \hat{\boldsymbol{h}}(\tau)^\top \right) \right) \right\rangle_{\boldsymbol{M}_0} = \\
&\exp\left(-\frac{1}{2} \int dt dt' C_v(t, t') \hat{\boldsymbol{v}}(t) \cdot \hat{\boldsymbol{v}}(t') + \frac{D}{2} \int dt dt' R_v(t, t') R_v(t', t) \right) \\
&\times \exp\left(-\frac{1}{2} \int d\tau d\tau' C_h(\tau, \tau') \hat{\boldsymbol{h}}(\tau) \cdot \hat{\boldsymbol{h}}(\tau') + \frac{D}{2} \int d\tau d\tau' R_h(\tau, \tau') R_h(\tau', \tau) \right) \\
&\times \exp\left(-\int dt d\tau C_{vh}(t, \tau) \hat{\boldsymbol{v}}(t) \cdot \hat{\boldsymbol{h}}(\tau) + D \int dt d\tau R_{hv}(\tau, t) R_{vh}(t, \tau) \right)
\end{aligned} \tag{36}$$

with definitions

$$C_v(t, t') = \frac{1}{D} \boldsymbol{v}(t) \cdot \boldsymbol{v}(t') \tag{37}$$

$$R_v(t, t') = -\frac{i}{D} \boldsymbol{v}(t) \cdot \hat{\boldsymbol{v}}(t') \tag{38}$$

$$C_h(\tau, \tau') = \frac{1}{D} \boldsymbol{h}(\tau) \cdot \boldsymbol{h}(\tau') \tag{39}$$

$$R_h(\tau, \tau') = -\frac{i}{D} \boldsymbol{h}(\tau) \cdot \hat{\boldsymbol{h}}(\tau') \tag{40}$$

$$C_{vh}(t, \tau) = \frac{1}{D} \boldsymbol{v}(t) \cdot \boldsymbol{h}(\tau) \tag{41}$$

$$R_{hv}(\tau, t) = -\frac{i}{D} \boldsymbol{h}(\tau) \cdot \hat{\boldsymbol{v}}(t) \tag{42}$$

that we can enforce through Fourier representation of the Dirac delta function

$$1 \equiv \int \frac{dR_v(t, t') d\hat{R}_v(t, t')}{2\pi} \exp\left(-D \int dt dt' R_v(t, t') \hat{R}_v(t, t') - i \int dt dt' \hat{R}_v(t, t') \boldsymbol{v}(t) \cdot \hat{\boldsymbol{v}}(t') \right) \tag{43}$$

$$1 \equiv \int \frac{dR_h(\tau, \tau') d\hat{R}_h(\tau, \tau')}{2\pi} \exp\left(-D \int d\tau d\tau' R_h(\tau, \tau') \hat{R}_h(\tau, \tau') - i \int d\tau d\tau' \hat{R}_h(\tau, \tau') \boldsymbol{h}(\tau) \cdot \hat{\boldsymbol{h}}(\tau') \right) \tag{44}$$

$$1 \equiv \int \frac{dR_{hv}(\tau, t) d\hat{R}_{hv}(\tau, t)}{2\pi} \exp\left(-D \int d\tau dt R_{hv}(\tau, t) \hat{R}_{hv}(\tau, t) - i \int d\tau dt \hat{R}_{hv}(\tau, t) \boldsymbol{h}(\tau) \cdot \hat{\boldsymbol{v}}(t) \right) \tag{45}$$

$$1 \equiv \int \frac{dR_{vh}(t, \tau) d\hat{R}_{vh}(t, \tau)}{2\pi} \exp\left(-D \int dt d\tau R_{vh}(t, \tau) \hat{R}_{vh}(t, \tau) - i \int dt d\tau \hat{R}_{vh}(t, \tau) \boldsymbol{v}(t) \cdot \hat{\boldsymbol{h}}(\tau) \right). \tag{46}$$

After averaging over the GOE disorder, its effect is replaced by a Gaussian dynamical noise and by retarded self-interactions. At the saddle point, the physical training trajectory obeys the usual single-site DMFT equation

$$\partial_t v(t) = u_v(t) - \gamma_0 \int_0^t dt' C_v(t',t) v(t') + \int dt' R_v(t,t') v(t') \tag{47}$$

where u_v is the Gaussian process with covariance $\langle u_v(t) u_v(t') \rangle = C_v(t,t')$. The order parameters C_v, R_v are fixed self-consistently from this training-level DMFT. Once this first DMFT problem has been solved, the spectral problem can be simplified as well as

$$\partial_\tau h = u_h(\tau) - \gamma_0 \int_0^t dt' C_{vh}(t',\tau) v(t') + \int d\tau' R_h(\tau,\tau') h(\tau') + \int dt' R_{hv}(\tau,t') v(t') \tag{48}$$

with joint Gaussian noise $\begin{pmatrix} u_v \\ u_h \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} C_v & C_{vh} \\ C_{vh}^\top & C_h \end{pmatrix}\right)$. Now, the bulk response of the probe is still the GOE resolvent. Therefore in Fourier space

$$R_h = \frac{1}{2}\Big[i\omega - \sqrt{(i\omega)^2 - 4}\Big] \tag{49}$$

with the physical branch fixed by $R_h(\omega) \sim \frac{1}{i\omega}$ at large $|\omega|$. The only spectral information not contained in the normalized trace is carried by the projected correlations

$$C_{vh}(t,\omega) = \frac{1}{D}\boldsymbol{v}(t) \cdot \boldsymbol{h}(\omega), \quad R_{vh}(t,\omega) = -\frac{i}{D}\boldsymbol{h}(\omega) \cdot \hat{\boldsymbol{v}}(t). \tag{50}$$

Taking the Fourier transform of the probe equation gives

$$h(\omega) = \Big[i\omega - R_h(\omega)\Big]^{-1}\Big(h_0 + u_h(\omega) + \int_0^t dt' \Big(R_{hv}(\omega,t') - \gamma_0 C_{vh}(t',\omega)\Big) v(t')\Big) \tag{51}$$

from which

$$C_{vh}(t'',\omega) = R_h(\omega)\Big(\langle v(t'') h_0 \rangle + \langle v(t'') u_h(\omega) \rangle + \int_0^t dt' \Big(R_{hv}(\omega,t') - \gamma_0 C_{vh}(t',\omega)\Big) C_v(t'',t')\Big) \tag{52}$$

$$R_{hv}(\omega,t'') = R_h(\omega)\Big(\int_{t''}^t dt' \Big(R_{hv}(\omega,t') - \gamma_0 C_{vh}(t',\omega)\Big) R_v(t',t'')\Big) \tag{53}$$

given the initial condition $\boldsymbol{c}_0 = \frac{1}{D}\boldsymbol{v}(t'') \cdot \boldsymbol{h}_0$ and

$$\langle v(t'') u_h(\omega) \rangle = \int_0^{t''} ds C_{vh}(s,\omega) R_v(t'',s). \tag{54}$$

After discretizing the physical time t, these two equations become a closed $2T \times 2T$ system

$$\underbrace{\begin{pmatrix} \boldsymbol{I} - R_h(\omega)(\boldsymbol{R}_v \odot \boldsymbol{\Theta}) + \gamma_0 \Delta t R_h(\omega) \boldsymbol{C}_v & -\Delta t R_h(\omega) \boldsymbol{C}_v \\ \gamma_0 R_h(\omega)(\boldsymbol{R}_v \odot \boldsymbol{\Theta})^\top & \boldsymbol{I} - R_h(\omega)(\boldsymbol{R}_v \odot \boldsymbol{\Theta})^\top \end{pmatrix}}_{\boldsymbol{\mathcal{A}}(\omega)} \begin{pmatrix} \boldsymbol{C}_{vh}(\omega) \\ \boldsymbol{R}_{hv}(\omega) \end{pmatrix} = \begin{pmatrix} R_h(\omega)\boldsymbol{c}_0 \\ 0 \end{pmatrix}. \tag{55}$$

Here $\odot$ denotes elementwise multiplication. The outliers correspond to poles of the projected resolvent, equivalently to the zeros of

$$\det \boldsymbol{\mathcal{A}}(\omega) = 0. \tag{56}$$

This toy model makes the role of the two-level construction explicit: the first DMFT level generates the statistically dependent spike directions, while the second level probes the spectrum of the resulting random matrix. The classical BBP pole condition of Appendix A is recovered when the spike direction is fixed.

C Derivation of Outlier Dynamics for Singular Values of Matrices with Statistically Dependent Spikes

Coupled Dynamics between System and Probe We consider an evolution equation for a probe vector $\boldsymbol{\psi}_0(\tau)$ as $\partial_\tau \boldsymbol{\psi}(\tau) = \boldsymbol{M}(t)\boldsymbol{\psi}(\tau)$. We decompose these dynamics into two components that are each linear in $\boldsymbol{W}(0)$

$$\textbf{Probe Dynamics}: \quad \boldsymbol{\psi}_1(\tau) = \frac{1}{\sqrt{N_0}}\boldsymbol{W}(0)\boldsymbol{\psi}(\tau) + \sum_s c_0(s)\boldsymbol{g}(s) \tag{57}$$

$$\partial_\tau \boldsymbol{\psi}_0(\tau) = \frac{\sqrt{N_0}}{N_1}\boldsymbol{W}(0)^\top \boldsymbol{\psi}_1(\tau) + \sum_s c_1(s)\boldsymbol{\phi}(s) \tag{58}$$

where we introduced the overlaps

$$\boldsymbol{c}_0(\tau, s) \equiv \frac{1}{N_0}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\phi}(s)\ ,\ \boldsymbol{c}_1(\tau, s) \equiv \frac{1}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\boldsymbol{g}(s) \tag{59}$$

These equations are combined with the defining equations for $\boldsymbol{\chi}(t)$ and $\boldsymbol{\xi}(t)$

$$\textbf{System Dynamics}: \quad \boldsymbol{\chi}(t) \equiv \frac{1}{\sqrt{N_0}}\boldsymbol{W}(0)\boldsymbol{\phi}(t)\ ,\ \boldsymbol{\xi}(t) \equiv \frac{\sqrt{N_0}}{N_1}\boldsymbol{W}(0)^\top \boldsymbol{g}(t) \tag{60}$$

$$\boldsymbol{\phi}(t) = \phi_t(\{\boldsymbol{\xi}(s)\})\ ,\ \boldsymbol{g}(t) = g_t(\{\boldsymbol{\chi}(s)\}) \tag{61}$$

where ϕ_t and g_t act elementwise over the entries of the vectors. We assume that the entries of $\boldsymbol{W}(0)$ are iid random variables with variance σ^2. For concreteness, we can assume Gaussian entries

$$W_{ij}(0) \sim \mathcal{N}(0, \sigma^2). \tag{62}$$

This condition defines a distribution over the random vectors $\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\chi}, \boldsymbol{\xi}$. We now aim to characterize this distribution using a formal path integral approach.

Martin–Siggia–Rose path integral. We set $N = N_0$ and $\alpha = N_1/N_0$. The probe and system variables obey

$$\boldsymbol{\psi}_1(\tau) = \frac{1}{\sqrt{N}}\boldsymbol{W}(0)\boldsymbol{\psi}_0(\tau) + \sum_s c_0(\tau, s)\boldsymbol{g}(s), \tag{63}$$

$$\partial_\tau \boldsymbol{\psi}_0(\tau) = \frac{\sqrt{N}}{N_1}\boldsymbol{W}(0)^\top \boldsymbol{\psi}_1(\tau) + \sum_s c_1(\tau, s)\boldsymbol{\phi}(s) + \boldsymbol{j}_\psi(\tau), \tag{64}$$

$$\boldsymbol{\chi}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}(0)\boldsymbol{\phi}(t), \tag{65}$$

$$\boldsymbol{\xi}(t) = \frac{\sqrt{N}}{N_1}\boldsymbol{W}(0)^\top \boldsymbol{g}(t). \tag{66}$$

Here

$$c_0(\tau, s) = \frac{1}{N}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\phi}(s), \qquad c_1(\tau, s) = \frac{1}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\boldsymbol{g}(s).$$

The MSR generating functional is

$$Z[\boldsymbol{j}_\psi, \boldsymbol{j}_\chi, \boldsymbol{j}_\xi] = \left\langle \exp\left(\int d\tau\, \boldsymbol{j}_\psi(\tau)\cdot\boldsymbol{\psi}_0(\tau) + \sum_t \boldsymbol{j}_\chi(t)\cdot\boldsymbol{\chi}(t) + \sum_t \boldsymbol{j}_\xi(t)\cdot\boldsymbol{\xi}(t) \right) \right\rangle_{\boldsymbol{W}(0)}. \tag{67}$$

Equivalently, imposing equation 63–equation 66 with MSR response fields gives

$$Z = \int \mathcal{D}\boldsymbol{\psi}_0 \mathcal{D}\boldsymbol{\psi}_1 \mathcal{D}\boldsymbol{\chi} \mathcal{D}\boldsymbol{\xi} \mathcal{D}\hat{\boldsymbol{\psi}}_0 \mathcal{D}\hat{\boldsymbol{\psi}}_1 \mathcal{D}\hat{\boldsymbol{\chi}} \mathcal{D}\hat{\boldsymbol{\xi}}\ \exp\{S_{\mathrm{MSR}}\}, \tag{68}$$

where

$$S_{\mathrm{MSR}} = i \int d\tau\, \hat{\boldsymbol{\psi}}_1(\tau) \cdot \left[\boldsymbol{\psi}_1(\tau) - \sum_s c_0(\tau, s) \boldsymbol{g}(s) \right] \tag{69}$$

$$+ i \int d\tau\, \hat{\boldsymbol{\psi}}_0(\tau) \cdot \left[\partial_\tau \boldsymbol{\psi}_0(\tau) - \sum_s c_1(\tau, s) \boldsymbol{\phi}(s) - \boldsymbol{j}_\psi(\tau) \right]$$

$$+ i \sum_t \hat{\boldsymbol{\chi}}(t) \cdot \boldsymbol{\chi}(t) + i \sum_t \hat{\boldsymbol{\xi}}(t) \cdot \boldsymbol{\xi}(t) - i \sum_{ij} W_{ij}(0) \mathcal{X}_{ij}. \tag{70}$$

The source conjugate to $W_{ij}(0)$ is

$$\mathcal{X}_{ij} = \frac{1}{\sqrt{N}} \left[\int d\tau\, \hat{\psi}_{1,i}(\tau) \psi_{0,j}(\tau) + \sum_t \hat{\chi}_i(t) \phi_j(t) \right] + \frac{\sqrt{N}}{N_1} \left[\int d\tau\, \psi_{1,i}(\tau) \hat{\psi}_{0,j}(\tau) + \sum_t g_i(t) \hat{\xi}_j(t) \right]. \tag{71}$$

Averaging over the initialization. Since $W_{ij}(0) \sim \mathcal{N}(0, \sigma^2)$,

$$\left\langle \exp\left(-i \sum_{ij} W_{ij}(0) \mathcal{X}_{ij} \right) \right\rangle_{\boldsymbol{W}(0)} = \exp\left(-\frac{\sigma^2}{2} \sum_{ij} \mathcal{X}_{ij}^2 \right). \tag{72}$$

Thus

$$Z = \int \mathcal{D}(\cdots) \exp\{S_0 - G\}, \qquad G = \frac{\sigma^2}{2} \sum_{ij} \mathcal{X}_{ij}^2.$$

Expanding G generates all quadratic contractions between the row-side variables

$$\hat{\boldsymbol{\psi}}_1, \quad \boldsymbol{\psi}_1, \quad \hat{\boldsymbol{\chi}}, \quad \boldsymbol{g}$$

and the column-side variables

$$\boldsymbol{\psi}_0, \quad \hat{\boldsymbol{\psi}}_0, \quad \boldsymbol{\phi}, \quad \hat{\boldsymbol{\xi}}.$$

Explicitly,

$$\begin{aligned} G = \frac{\sigma^2}{2} \sum_{ij} \Bigg\{ & \frac{1}{N} \int d\tau d\tau'\, \hat{\psi}_{1,i}(\tau) \hat{\psi}_{1,i}(\tau') \psi_{0,j}(\tau) \psi_{0,j}(\tau') \\ & + \frac{1}{\alpha^2 N} \int d\tau d\tau'\, \psi_{1,i}(\tau) \psi_{1,i}(\tau') \hat{\psi}_{0,j}(\tau) \hat{\psi}_{0,j}(\tau') \\ & + \frac{1}{N} \sum_{t,s} \hat{\chi}_i(t) \hat{\chi}_i(s) \phi_j(t) \phi_j(s) \\ & + \frac{1}{\alpha^2 N} \sum_{t,s} g_i(t) g_i(s) \hat{\xi}_j(t) \hat{\xi}_j(s) \\ & + \frac{2}{\alpha N} \int d\tau d\tau'\, \hat{\psi}_{1,i}(\tau) \psi_{1,i}(\tau') \psi_{0,j}(\tau) \hat{\psi}_{0,j}(\tau') \\ & + \frac{2}{N} \int d\tau \sum_t \hat{\psi}_{1,i}(\tau) \hat{\chi}_i(t) \psi_{0,j}(\tau) \phi_j(t) \\ & + \frac{2}{\alpha N} \int d\tau \sum_t \hat{\psi}_{1,i}(\tau) g_i(t) \psi_{0,j}(\tau) \hat{\xi}_j(t) \\ & + \frac{2}{\alpha N} \int d\tau \sum_t \psi_{1,i}(\tau) \hat{\chi}_i(t) \hat{\psi}_{0,j}(\tau) \phi_j(t) \\ & + \frac{2}{\alpha^2 N} \int d\tau \sum_t \psi_{1,i}(\tau) g_i(t) \hat{\psi}_{0,j}(\tau) \hat{\xi}_j(t) \\ & + \frac{2}{\alpha N} \sum_{t,s} \hat{\chi}_i(t) g_i(s) \phi_j(t) \hat{\xi}_j(s) \Bigg\}. \end{aligned} \tag{73}$$

Order parameters. The contractions in equation 73 are expressed in terms of the following macroscopic order parameters. For the probe,

$$C^{\psi_0}(\tau,\tau') = \frac{1}{N}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\psi}_0(\tau'), \qquad C^{\psi_1}(\tau,\tau') = \frac{1}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\boldsymbol{\psi}_1(\tau'), \tag{74}$$

$$R^{\psi_0}(\tau,\tau') = -\frac{i}{N}\boldsymbol{\psi}_0(\tau)\cdot\hat{\boldsymbol{\psi}}_0(\tau'), \qquad R^{\psi_1}(\tau,\tau') = -\frac{i}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\hat{\boldsymbol{\psi}}_1(\tau'). \tag{75}$$

For the training-level system,

$$C^{\phi}(t,s) = \frac{1}{N}\boldsymbol{\phi}(t)\cdot\boldsymbol{\phi}(s), \qquad C^{g}(t,s) = \frac{1}{N_1}\boldsymbol{g}(t)\cdot\boldsymbol{g}(s), \tag{76}$$

$$R^{\phi}(t,s) = -\frac{i}{N}\boldsymbol{\phi}(t)\cdot\hat{\boldsymbol{\xi}}(s), \qquad R^{g}(t,s) = -\frac{i}{N_1}\boldsymbol{g}(t)\cdot\hat{\boldsymbol{\chi}}(s). \tag{77}$$

For the mixed probe–system quantities,

$$c_0(\tau,t) = \frac{1}{N}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\phi}(t), \qquad c_1(\tau,t) = \frac{1}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\boldsymbol{g}(t), \tag{78}$$

$$\rho_0(\tau,t) = -\frac{i}{N}\boldsymbol{\psi}_0(\tau)\cdot\hat{\boldsymbol{\xi}}(t), \qquad \rho_1(\tau,t) = -\frac{i}{N_1}\boldsymbol{\psi}_1(\tau)\cdot\hat{\boldsymbol{\chi}}(t). \tag{79}$$

Conjugate order parameters. Each order parameter is enforced by a Fourier representation of a delta function. For example,

$$1 = \int dC^{\psi_0} d\widehat{C}^{\psi_0} \exp\left\{-N\int d\tau d\tau'\, \widehat{C}^{\psi_0}(\tau,\tau')\left[C^{\psi_0}(\tau,\tau') - \frac{1}{N}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\psi}_0(\tau')\right]\right\}, \tag{80}$$

and

$$1 = \int dR^{\psi_0} d\widehat{R}^{\psi_0} \exp\left\{-N\int d\tau d\tau'\, \widehat{R}^{\psi_0}(\tau,\tau')\left[R^{\psi_0}(\tau,\tau') + \frac{i}{N}\boldsymbol{\psi}_0(\tau)\cdot\hat{\boldsymbol{\psi}}_0(\tau')\right]\right\}. \tag{81}$$

Similarly, we insert identities for

$$C^{\psi_1}, C^{\phi}, C^{g}, R^{\psi_1}, R^{\phi}, R^{g}, c_0, c_1, \rho_0, \rho_1.$$

We collect all order parameters and conjugate fields into

$$Q = \left\{C, R, c, \rho, \widehat{C}, \widehat{R}, \widehat{c}, \widehat{\rho}\right\}.$$

Large-N effective action. After inserting the order-parameter identities, the microscopic variables factorize over row and column indices. Therefore the generating functional has the large-deviation form

$$Z = \int dQ\, \exp\left\{-N S_{\text{eff}}[Q]\right\}, \tag{82}$$

where

$$S_{\text{eff}}[Q] = S_{\text{op}}[Q] - \log Z_0[Q] - \alpha \log Z_1[Q]. \tag{83}$$

Here $Z_0[Q]$ is the single-site path integral for a representative column-side site (ψ_0, ϕ, ξ), $Z_1[Q]$ is the single-site path integral for a representative row-side site (ψ_1, g, χ), and S_{op} contains the order-parameter and conjugate-order-parameter terms.

In the proportional limit, $N, N_1 \to \infty$ with $N_1/N = \alpha$, the integral equation 82 is evaluated by saddle point:

$$\frac{\delta S_{\text{eff}}}{\delta Q} = 0.$$

Variation with respect to conjugate order parameters recovers the definitions of C, R, c, ρ. Variation with respect to the order parameters fixes the conjugates, or equivalently the covariances and retarded self-interactions of the effective single-site processes.

Saddle point in (τ, t)-space. At the saddle point, the disorder-averaged theory is equivalent to the following self-consistent stochastic process:

$$\partial_\tau \psi_0(\tau) = u^{\psi_0}(\tau) + \sigma^2 \int d\tau'\, R^{\psi_1}(\tau,\tau')\psi_0(\tau') + \sigma^2 \sum_s \rho_1(\tau,s)\phi(s) + \sum_s c_1(\tau,s)\phi(s) + j_\psi(\tau), \tag{84}$$

$$\psi_1(\tau) = u^{\psi_1}(\tau) + \frac{\sigma^2}{\alpha} \int d\tau'\, R^{\psi_0}(\tau,\tau')\psi_1(\tau') + \frac{\sigma^2}{\alpha} \sum_s \rho_0(\tau,s)g(s) + \sum_s c_0(\tau,s)g(s), \tag{85}$$

$$\chi(t) = u^{\chi}(t) + \frac{\sigma^2}{\alpha} \sum_s R^{\phi}(t,s)g(s), \tag{86}$$

$$\xi(t) = u^{\xi}(t) + \sigma^2 \sum_s R^{g}(t,s)\phi(s). \tag{87}$$

The functions $\phi(t)$ and $g(t)$ are then obtained from the elementwise structural relations

$$\phi(t) = \phi_t(\{\chi(s), \xi(s)\}_{s<T}), \qquad g(t) = g_t(\{\chi(s), \xi(s)\}_{s<T}).$$

The Gaussian noises in equation 84–equation 87 have zero mean and covariances

$$\left\langle u^{\psi_0}(\tau)u^{\psi_0}(\tau')\right\rangle = \frac{\sigma^2}{\alpha} C^{\psi_1}(\tau,\tau'), \qquad \left\langle u^{\psi_1}(\tau)u^{\psi_1}(\tau')\right\rangle = \sigma^2 C^{\psi_0}(\tau,\tau'), \tag{88}$$

$$\left\langle u^{\chi}(t)u^{\chi}(s)\right\rangle = \sigma^2 C^{\phi}(t,s), \qquad \left\langle u^{\xi}(t)u^{\xi}(s)\right\rangle = \frac{\sigma^2}{\alpha} C^{g}(t,s), \tag{89}$$

$$\left\langle u^{\psi_0}(\tau)u^{\xi}(s)\right\rangle = \frac{\sigma^2}{\alpha} c_1(\tau,s), \qquad \left\langle u^{\psi_1}(\tau)u^{\chi}(s)\right\rangle = \sigma^2 c_0(\tau,s). \tag{90}$$

The saddle-point values of the order parameters are deterministic and satisfy

$$C^{\psi_0}(\tau,\tau') = \left\langle \psi_0(\tau)\psi_0(\tau')\right\rangle, \qquad C^{\psi_1}(\tau,\tau') = \left\langle \psi_1(\tau)\psi_1(\tau')\right\rangle, \tag{91}$$

$$R^{\psi_0}(\tau,\tau') = \left\langle \frac{\delta\psi_0(\tau)}{\delta u^{\psi_0}(\tau')} \right\rangle, \qquad R^{\psi_1}(\tau,\tau') = \left\langle \frac{\delta\psi_1(\tau)}{\delta u^{\psi_1}(\tau')} \right\rangle, \tag{92}$$

$$C^{\phi}(t,s) = \left\langle \phi(t)\phi(s)\right\rangle, \qquad C^{g}(t,s) = \left\langle g(t)g(s)\right\rangle, \tag{93}$$

$$R^{\phi}(t,s) = \left\langle \frac{\partial\phi(t)}{\partial u^{\xi}(s)} \right\rangle, \qquad R^{g}(t,s) = \left\langle \frac{\partial g(t)}{\partial u^{\chi}(s)} \right\rangle, \tag{94}$$

$$c_0(\tau,t) = \left\langle \psi_0(\tau)\phi(t)\right\rangle, \qquad c_1(\tau,t) = \left\langle \psi_1(\tau)g(t)\right\rangle, \tag{95}$$

$$\rho_0(\tau,t) = \left\langle \frac{\partial\psi_0(\tau)}{\partial u^{\xi}(t)} \right\rangle, \qquad \rho_1(\tau,t) = \left\langle \frac{\partial\psi_1(\tau)}{\partial u^{\chi}(t)} \right\rangle. \tag{96}$$

Projected saddle-point equations in (τ, t)-space. We now derive the closed equations for the mixed order parameters before taking any Laplace transform. Multiplying equation 84 by $\phi(t)$, averaging, and using Stein's lemma gives

$$\partial_\tau c_0(\tau,t) = \sigma^2 \int d\tau'\, R^{\psi_1}(\tau,\tau')c_0(\tau',t) + \frac{\sigma^2}{\alpha} \sum_s R^{\phi}(t,s)c_1(\tau,s) + \sigma^2 \sum_s C^{\phi}(t,s)\rho_1(\tau,s) + \sum_s C^{\phi}(t,s)c_1(\tau,s). \tag{97}$$

Here we used

$$\left\langle u^{\psi_0}(\tau)\phi(t)\right\rangle = \frac{\sigma^2}{\alpha} \sum_s R^{\phi}(t,s)c_1(\tau,s).$$

Differentiating equation 84 with respect to $u^{\xi}(t)$ gives

$$\partial_\tau \rho_0(\tau,t) = \sigma^2 \int d\tau'\, R^{\psi_1}(\tau,\tau')\rho_0(\tau',t) + \sigma^2 \sum_s R^{\phi}(s,t)\rho_1(\tau,s) + \sum_s R^{\phi}(s,t)c_1(\tau,s). \tag{98}$$

Similarly, multiplying equation 85 by $g(t)$, averaging, and using

$$\left\langle u^{\psi_1}(\tau) g(t) \right\rangle = \sigma^2 \sum_s R^g(t,s) c_0(\tau, s),$$

gives

$$c_1(\tau,t) = \frac{\sigma^2}{\alpha} \int d\tau' \, R^{\psi_0}(\tau,\tau') c_1(\tau',t) + \sigma^2 \sum_s R^g(t,s) c_0(\tau,s) + \frac{\sigma^2}{\alpha} \sum_s C^g(t,s) \rho_0(\tau,s) + \sum_s C^g(t,s) c_0(\tau,s). \tag{99}$$

Finally, differentiating equation 85 with respect to $u^\chi(t)$ gives

$$\rho_1(\tau,t) = \frac{\sigma^2}{\alpha} \int d\tau' \, R^{\psi_0}(\tau,\tau') \rho_1(\tau',t) + \frac{\sigma^2}{\alpha} \sum_s R^g(s,t) \rho_0(\tau,s) + \sum_s R^g(s,t) c_0(\tau,s). \tag{100}$$

Laplace transform in the auxiliary spectral time. Only after obtaining the saddle-point equations equation 97–equation 100 do we Laplace transform in τ. For any probe-time function $f(\tau)$, define

$$f(z) = \int_0^\infty d\tau \, e^{-z\tau} f(\tau), \qquad \Re z > 0.$$

Boundary terms from the Laplace transform encode the initial probe condition and do not affect the pole condition for outliers, so we suppress them below.

The pure probe responses obey, in (τ, τ')-space,

$$\left[\partial_\tau - \sigma^2 R^{\psi_1}\right] \circ R^{\psi_0} = \delta, \qquad \left[I - \frac{\sigma^2}{\alpha} R^{\psi_0}\right] \circ R^{\psi_1} = \delta, \tag{101}$$

where $\circ$ denotes convolution in the auxiliary time. After Laplace transformation,

$$R^{\psi_0}(z) = \frac{1}{z - \sigma^2 R^{\psi_1}(z)}, \tag{102}$$

$$R^{\psi_1}(z) = \frac{1}{1 - \frac{\sigma^2}{\alpha} R^{\psi_0}(z)}. \tag{103}$$

We identify

$$\mathcal{G}(z) := R^{\psi_0}(z),$$

the Stieltjes transform of the Marchenko–Pastur bulk. Hence

$$z - \sigma^2 R^{\psi_1}(z) = \mathcal{G}(z)^{-1}, \qquad 1 - \frac{\sigma^2}{\alpha} \mathcal{G}(z) = R^{\psi_1}(z)^{-1}.$$

Laplace transforming equation 97–equation 100 gives

$$\mathcal{G}(z)^{-1} c_0(t) = \frac{\sigma^2}{\alpha} \sum_s R^\phi(t,s) c_1(s) + \sigma^2 \sum_s C^\phi(t,s) \rho_1(s) + \sum_s C^\phi(t,s) c_1(s), \tag{104}$$

$$\mathcal{G}(z)^{-1} \rho_0(t) = \sigma^2 \sum_s R^\phi(s,t) \rho_1(s) + \sum_s R^\phi(s,t) c_1(s), \tag{105}$$

$$\left(1 - \frac{\sigma^2}{\alpha} \mathcal{G}(z)\right) c_1(t) = \sigma^2 \sum_s R^g(t,s) c_0(s) + \frac{\sigma^2}{\alpha} \sum_s C^g(t,s) \rho_0(s) + \sum_s C^g(t,s) c_0(s), \tag{106}$$

$$\left(1 - \frac{\sigma^2}{\alpha} \mathcal{G}(z)\right) \rho_1(t) = \frac{\sigma^2}{\alpha} \sum_s R^g(s,t) \rho_0(s) + \sum_s R^g(s,t) c_0(s). \tag{107}$$

Outlier determinant condition. Collecting the Laplace-domain mixed order parameters into

$$\boldsymbol{q}(z) = \begin{pmatrix} \boldsymbol{c}_0(z) \\ \boldsymbol{\rho}_0(z) \\ \boldsymbol{c}_1(z) \\ \boldsymbol{\rho}_1(z) \end{pmatrix},$$

equations equation 104–equation 107 can be written as

$$\boldsymbol{A}(z)\boldsymbol{q}(z) = \boldsymbol{q}_0,$$

where $\boldsymbol{q}_0$ is an initial condition and $\boldsymbol{A}$ has the form

$$\boldsymbol{A}(z) = \begin{bmatrix} \mathcal{G}(z)^{-1}\boldsymbol{I} & \mathbf{0} & -\left(\sigma^2\alpha^{-1}\boldsymbol{R}^\phi + \boldsymbol{C}^\phi\right) & -\sigma^2\boldsymbol{C}^\phi \\ \mathbf{0} & \mathcal{G}(z)^{-1}\boldsymbol{I} & -(\boldsymbol{R}^\phi)^\top & -\sigma^2(\boldsymbol{R}^\phi)^\top \\ -(\sigma^2\boldsymbol{R}^g + \boldsymbol{C}^g) & -\sigma^2\alpha^{-1}\boldsymbol{C}^g & \left(1 - \frac{\sigma^2}{\alpha}\mathcal{G}(z)\right)\boldsymbol{I} & \mathbf{0} \\ -(\boldsymbol{R}^g)^\top & -\sigma^2\alpha^{-1}(\boldsymbol{R}^g)^\top & \mathbf{0} & \left(1 - \frac{\sigma^2}{\alpha}\mathcal{G}(z)\right)\boldsymbol{I} \end{bmatrix}.$$

Nontrivial projected resolvents exist when

$$\det \boldsymbol{A}(z) = 0.$$

The isolated outlier eigenvalues are the solutions outside the Marchenko–Pastur support.

C.1 Closed-form example: one population GD step

For the sake of clarity, in this short subsection we illustrate the two-step DMFT in a deliberately simple setting admitting a closed-form solution. Of course, the general theory extends beyond this example to multi-step dynamics, nonlinear networks, and data-dependent updates.

Consider a one-hidden-layer linear network in the population limit. After one gradient step,

$$\boldsymbol{W}_1 = \boldsymbol{W}_0 + \frac{\eta\gamma_0}{\sqrt{D}}\boldsymbol{w}_0\boldsymbol{v}_0^\top, \tag{108}$$

$$\boldsymbol{v}_0 = \boldsymbol{\beta}_\star - \frac{\sqrt{D}}{N\gamma_0}\boldsymbol{W}_0^\top\boldsymbol{w}_0 = \boldsymbol{\beta}_\star - \frac{\boldsymbol{\xi}}{\gamma_0}. \tag{109}$$

This has exactly the form of Eq. equation 1, since $\boldsymbol{W}_1 = \boldsymbol{W}_0 + D^{-1/2}\boldsymbol{g}\boldsymbol{\phi}^\top$ with $\boldsymbol{g} = \boldsymbol{w}_0$ and $\boldsymbol{\phi} = \eta\gamma_0\boldsymbol{v}_0$. Assuming independent standard Gaussian entries in $\boldsymbol{W}_0$, $\boldsymbol{w}_0$, and $\boldsymbol{\beta}_\star$, in the proportional limit $N, D \to \infty$ with $\nu = N/D$ fixed,

$$\frac{1}{N}||\mathbf{g}||^2 = 1, \quad \xi_i = \mathcal{N}(0, \frac{1}{\nu}). \tag{110}$$

The four scalar DMFT order parameters entering Result 1 from Lemma 1 are therefore

$$C^g = 1, \qquad R^g = 0, \qquad C^\phi = \eta^2(\gamma_0^2 + \nu^{-1}), \qquad R^\phi = \left\langle \frac{\partial\phi}{\partial u^\xi} \right\rangle = -\eta. \tag{111}$$

Substituting these into the determinant condition of Result 1 (with $\alpha = \nu$ and $\sigma^2 = 1$) gives

$$\det \boldsymbol{A}(z) = H(z)[H(z) - \kappa], \qquad H(z) = \frac{1}{\mathcal{G}(z)} - \frac{1}{\nu}, \qquad \kappa = C^\phi + \frac{2}{\nu}R^\phi, \tag{112}$$

where $\mathcal{G}(z)$ is the Stieltjes transform of the Wishart bulk given in Result 1. The learned outlier therefore comes from $H(z) = \kappa$, since $H(z) = 0$ does not give an outlier outside the MP bulk. Substituting $\mathcal{G}(z)$ yields the upper outlier of $\boldsymbol{W}_1^\top\boldsymbol{W}_1/N$

$$\lambda_{\text{out}} = \frac{(\kappa+1)(\kappa+\nu^{-1})}{\kappa}, \qquad \kappa = \eta^2\gamma_0^2 + \frac{\eta^2 - 2\eta}{\nu}. \tag{113}$$

This upper outlier detaches from the MP bulk only when $\kappa > \nu^{-1/2}$.

D NONLINEAR NETWORKS

In this section, we summarize the results of Bordelon & Pehlevan '22 Bordelon & Pehlevan (2022) which provides the evolution of hidden features and outputs of randomly initialized neural networks.

Setting and Super-Wide Scaling Assumptions We consider models in mean field parameterization (output scaling $\frac{1}{N\gamma}$ on the last layer) and consider SGD training dynamics from random initialization for all parameters $\boldsymbol{\theta} = \{\boldsymbol{W}^\ell\}$. Let the hidden widths be N and let depth L be fixed but arbitrary. We consider SGD dynamics with an arbitrary mini-batch stream $\mathcal{B}_t$

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta\gamma^2 N \, \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}_t} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) \tag{114}$$

These dynamics translate into each hidden weight matrix as

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\eta\gamma}{\sqrt{N}} \sum_{s<t} \mathbb{E}_{\boldsymbol{x}_\mu\in\mathcal{B}_s} \, \Delta_\mu(s) \boldsymbol{g}_\mu(s) \boldsymbol{\phi}_\mu(s)^\top \tag{115}$$

where $\boldsymbol{g}_\mu^\ell(t) \equiv N\gamma \frac{\partial f_\mu(t)}{\partial \boldsymbol{h}_\mu^\ell(t)}$ and $\Delta_\mu(t) = -\frac{\partial}{\partial f_\mu(t)} \mathcal{L}(f_\mu(t))$. This leads to the following dynamics for variables $\boldsymbol{h}_\mu^\ell(t)$ and $\boldsymbol{z}_\mu^\ell(t)$

$$\begin{aligned}
\boldsymbol{h}_\mu^\ell(t) &= \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(t) \boldsymbol{\phi}_\mu^{\ell-1}(t) = \boldsymbol{\chi}_\mu^\ell(t) + \eta\gamma \sum_{s<t} \mathbb{E}_{\boldsymbol{x}_\nu\in\mathcal{B}_s} \Delta_\nu(s) C_{\mu\nu}^{\phi^{\ell-1}}(t,s) \boldsymbol{g}_\nu^\ell(s) \\
\boldsymbol{z}_\mu^\ell(t) &= \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(t)^\top \boldsymbol{g}_\mu^{\ell+1}(t) = \boldsymbol{\xi}_\mu^\ell(t) + \eta\gamma \sum_{s<t} \mathbb{E}_{\boldsymbol{x}_\nu\in\mathcal{B}_s} \Delta_\nu(s) C_{\mu\nu}^{g^{\ell+1}}(t,s) \boldsymbol{\phi}_\mu^\ell(t) \\
\boldsymbol{\phi}_\mu^\ell(t) &= \phi(\boldsymbol{h}_\mu^\ell(t)) \, , \; \boldsymbol{g}_\mu^\ell(t) = \boldsymbol{z}_\mu^\ell(t) \odot \dot{\phi}(\boldsymbol{h}_\mu^\ell(t))
\end{aligned} \tag{116}$$

where we introduced the variables $\boldsymbol{\chi}$ and $\boldsymbol{\xi}$ that depend on the random matrix

$$\boldsymbol{\chi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \boldsymbol{\phi}_\mu^{\ell-1}(t) \, , \; \boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t). \tag{117}$$

The main claim to utilize our theory is that this dynamical system satisfies the requirements of the SVD result. We now will argue that this holds under the following assumptions

1. **Random Init**: The weights $W_{ij}^\ell(0)$ are initialized with iid random entries.
2. **Super-wide Scaling**: The network width $N \to \infty$ with batch size B, number of update steps T, and input-dimension D are held fixed.

Under these assumptions, one can show using DMFT Bordelon & Pehlevan (2022) that the feature structural condition holds since we have

1. **Concentration of Correlations and Errors**: The quantities C^ϕ and C^g and Δ concentrate due to a law of large numbers effect (they are averages over neurons in each layer).
2. **Decoupling of Neurons**: The neurons effectively decouple in their dynamics and become iid random processes throughout training.

As a consequence, we can indeed view the single-site equations as defining the features $\boldsymbol{\phi}$ and $\boldsymbol{g}$ elementwise in terms of $\boldsymbol{\chi}$ and $\boldsymbol{\xi}$

$$\boldsymbol{\phi}_\mu^\ell(t) = \phi_{\mu,t}^\ell\left(\{\boldsymbol{\xi}^\ell\}\right) \, , \; \boldsymbol{g}_\mu^{\ell+1}(t) = g_{\mu,t}^{\ell+1}\left(\{\boldsymbol{\chi}^{\ell+1}\}\right), \tag{118}$$

which verifies that our structural assumptions for Result 1 are satisfied.

Connection to Result 1 For each hidden layer ℓ, the trained weight matrix has the form

$$\boldsymbol{W}^\ell(T) = \boldsymbol{W}^\ell(0) + \frac{\eta\gamma}{\sqrt{N}} \sum_{s<T} \mathbb{E}_{\boldsymbol{x}_\nu\in\mathcal{B}_s} \bar{\Delta}_\nu(s) \, \boldsymbol{g}_\nu^{\ell+1}(s) \boldsymbol{\phi}_\nu^\ell(s)^\top.$$

This is precisely a finite-rank deformation of the initialization. Moreover, the spike vectors

$$\boldsymbol{\phi}_\nu^\ell(s), \qquad \boldsymbol{g}_\nu^{\ell+1}(s)$$

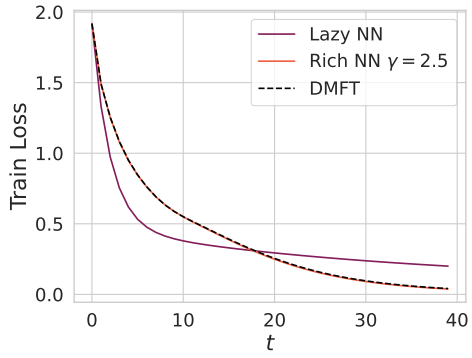


Figure 11: Loss Dynamics accurately predicted in the rich regime $\gamma > 0$.

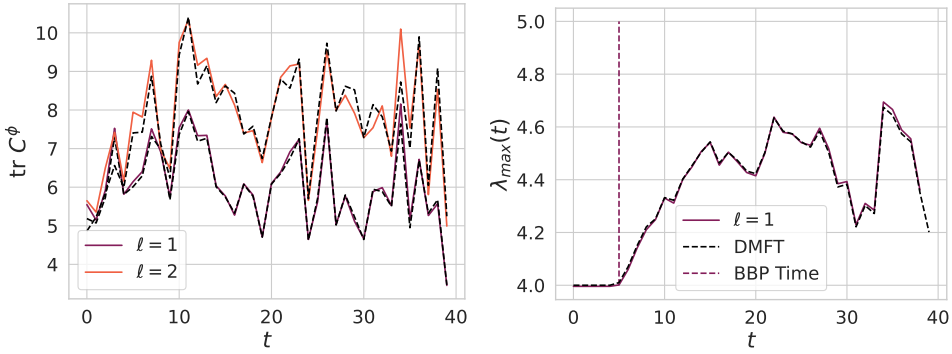


Figure 12: Our theory can also describe networks trained with minibatch SGD. (a) Hidden correlation functions in a nonlinear $L = 2$ hidden layer network trained with online SGD. (b) The maximum eigenvalue obeys a stochastic trajectory predicted by our outlier condition from Result 1.

have asymptotically iid entries whose single-site values are deterministic causal functions of the Gaussian fields (χ, ξ) generated by the same initial matrix $\boldsymbol{W}^\ell(0)$. Hence the feature structural condition required by Result 2.1 is satisfied. The matrices

$$\boldsymbol{C}^\phi, \quad \boldsymbol{C}^g, \quad \boldsymbol{R}^\phi, \quad \boldsymbol{R}^g$$

appearing in the outlier determinant are the corresponding single-site correlations and responses:

$$C^\phi_{ab} = \langle \phi_a \phi_b \rangle, \qquad C^g_{ab} = \langle g_a g_b \rangle,$$

and

$$R^\phi_{ab} = \left\langle \frac{\partial \phi_a}{\partial \xi_b} \right\rangle, \qquad R^g_{ab} = \left\langle \frac{\partial g_a}{\partial \chi_b} \right\rangle,$$

where the combined spike index $a = (s, \nu)$ ranges over all training times and mini-batch examples contributing to the finite-rank update.

E Two-Level DMFT for Spectral Dynamics in Deep Linear Networks

In this section, we extend the proportional-limit DMFT for deep linear networks to compute the spectrum of trained weight matrices. As usual, the first level of the theory describes the training dynamics of the fields generated by gradient descent, while the second level introduces a linear spectral probe whose response gives the resolvent of the trained weight covariance. This two-level construction is still accurate because the directions generated by the gradient updates are both low-rank and statistically coupled to the random initial weights.

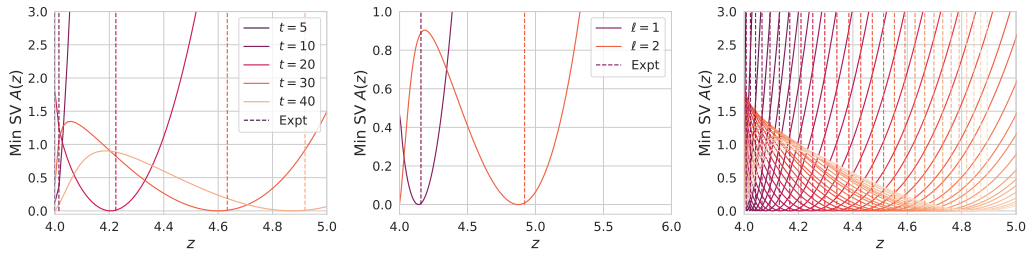


Figure 13: Demonstration of the root finding procedure for $\det \boldsymbol{A}(z) = 0$. In practice, we compute the minimum singular value for $\boldsymbol{A}(z)$. When the minimum singular value at z is lower than a specified threshold, we declare that value of z a solution in the set O_S.

E.1 Model definition and high dimensional limit

We consider a depth-L linear network with input dimension D, hidden width N, and scalar output

$$f(\boldsymbol{x}) = \frac{\sqrt{D}}{N\gamma_0}(\boldsymbol{w}^L)^\top \left[\prod_{\ell=1}^{L-1}\left(\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell\right)\right]\left(\frac{1}{\sqrt{D}}\boldsymbol{W}^0\right)\boldsymbol{x}. \tag{119}$$

The first-layer weight matrix is $\boldsymbol{W}^0 \in \mathbb{R}^{N\times D}$, the hidden-layer weights are $\boldsymbol{W}^\ell \in \mathbb{R}^{N\times N}$ for $\ell = 1, \dots, L-1$, and the readout is $\boldsymbol{w}^L \in \mathbb{R}^N$. The parameter γ_0 controls the feature-learning scale.

The training data are $\mathcal{D} = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^P$, with isotropic random covariates and a linear teacher

$$y_\mu = \frac{1}{\sqrt{D}}\boldsymbol{w}_\star \cdot \boldsymbol{x}_\mu. \tag{120}$$

We study the proportional high-dimensional asymptotic limit

$$P, N, D \to \infty, \qquad \alpha = \frac{P}{D}, \qquad \nu = \frac{N}{D} \tag{121}$$

and given a fixed number of training steps $T = \mathcal{O}(1)$.

Our goal is to characterize the spectrum of the trained weight covariances

$$\boldsymbol{M}^\ell(t) = \frac{1}{N}\boldsymbol{W}^\ell(t)^\top \boldsymbol{W}^\ell(t) \in \mathbb{R}^{N\times N} \qquad \forall \ell = 1, \dots, L-1 \tag{122}$$

and similarly for the first layer

$$\boldsymbol{M}^0(t) = \frac{1}{N}\boldsymbol{W}^0(t)^\top \boldsymbol{W}^0(t) \in \mathbb{R}^{D\times D}. \tag{123}$$

E.2 Training fields and finite-rank weight updates

Let $\Delta_\mu(t) = y_\mu - f(\boldsymbol{x}_\mu, t)$ denote the training error at step t. We define the training-set averaged forward field

$$\boldsymbol{h}^0(t) = \frac{\sqrt{D}}{P}\sum_{\mu=1}^P \Delta_\mu(t)\boldsymbol{x}_\mu, \tag{124}$$

and the hidden forward and backward fields recursively by

$$\boldsymbol{h}^{\ell+1}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(t)\boldsymbol{h}^\ell(t) \qquad \ell = 0, \dots, L-1 \tag{125}$$

$$\boldsymbol{g}^L(t) = \boldsymbol{w}^L(t) \tag{126}$$

$$\boldsymbol{g}^\ell(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(t)^\top \boldsymbol{g}^{\ell+1}(t), \qquad \ell = 1, \dots, L-1. \tag{127}$$

The gradient updates can be written as finite-rank deformations of the initial random Gaussian weights $W_{ij}^{\ell}(0) \sim \mathcal{N}(0,1)$

$$\boldsymbol{W}^0(t) = \boldsymbol{W}^0(0) + \frac{\eta\gamma_0}{\sqrt{D}} \sum_{s<t} \boldsymbol{g}^1(s)\boldsymbol{h}^0(s)^\top \tag{128}$$

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\eta\gamma_0}{\sqrt{N}} \sum_{s<t} \boldsymbol{g}^{\ell+1}(s)\boldsymbol{h}^\ell(s)^\top \qquad 1 \le \ell \le L-1 \tag{129}$$

$$\boldsymbol{w}^L(t) = \boldsymbol{w}^L(0) + \eta\gamma_0 \sum_{s<t} \boldsymbol{h}^L(s). \tag{130}$$

which at fixed training time t is spanned by the training fields. To isolate the dependence on the initialized weights, we define

$$\boldsymbol{\xi}^0(t) = \frac{\sqrt{D}}{N\gamma_0}\boldsymbol{W}^0(0)^\top \boldsymbol{g}^1(t) \qquad \boldsymbol{\chi}^1(t) = \frac{1}{\sqrt{D}}\boldsymbol{W}^0(0)\boldsymbol{h}^0(t) \tag{131}$$

$$\boldsymbol{\xi}^\ell(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)^\top \boldsymbol{g}^{\ell+1}(t), \qquad \boldsymbol{\chi}^{\ell+1}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\boldsymbol{h}^\ell(t) \qquad 1 \le \ell \le L-1. \tag{132}$$

Substituting the finite-rank weight updates into the forward and backward recursions gives

$$\boldsymbol{v}(t) = \boldsymbol{w}_\star - \boldsymbol{\xi}^0(t) - \eta \sum_{s<t} C_g^1(t,s)\boldsymbol{h}^0(s) \tag{133}$$

$$\boldsymbol{h}^{\ell+1}(t) = \boldsymbol{\chi}^{\ell+1}(t) + \eta\gamma_0 \sum_{s<t} C_h^\ell(t,s)\boldsymbol{g}^{\ell+1}(s) \tag{134}$$

$$\boldsymbol{g}^\ell(t) = \boldsymbol{\xi}^\ell(t) + \eta\gamma_0 \sum_{s<t} C_g^{\ell+1}(t,s)\boldsymbol{h}^\ell(s). \tag{135}$$

Here

$$C_h^\ell(t,s) = \frac{1}{N}\boldsymbol{h}^\ell(t)\cdot\boldsymbol{h}^\ell(s), \qquad C_g^\ell(t,s) = \frac{1}{N}\boldsymbol{g}^\ell(t)\cdot\boldsymbol{g}^\ell(s) \tag{136}$$

with the convention that C_h^0 is normalized by D when $\boldsymbol{h}^0 \in \mathbb{R}^D$.

The first-level DMFT determines the joint law of the fields

$$\boldsymbol{v}(t), \quad \Delta(t), \quad \boldsymbol{h}^\ell(t), \quad \boldsymbol{g}^\ell(t), \tag{137}$$

through their correlation and response functions. These are the same training-level order parameters used to characterize the loss dynamics in the proportional limit (see Bordelon & Pehlevan (2025) for more details).

E.3 Spectral probe for the trained weight covariance

We now introduce a second-level probe to compute the spectrum of $\boldsymbol{M}^\ell(t)$. For a fixed layer ℓ and fixed training time t, consider the linear response dynamics

$$\partial_\tau \boldsymbol{\psi}_0(\tau) = \boldsymbol{M}^\ell(t)\boldsymbol{\psi}_0(\tau) + \boldsymbol{j}(\tau). \tag{138}$$

The response of $\boldsymbol{\psi}_0$ to the source $\boldsymbol{j}$ gives the resolvent of $\boldsymbol{M}^\ell(t)$

$$\mathcal{H}^\ell(\omega) = \frac{1}{d_\ell}\operatorname{Tr}\left(i\omega\boldsymbol{I} - \boldsymbol{M}^\ell(t)\right)^{-1} \tag{139}$$

where $d_0 = D$ for the first layer and $d_\ell = N$ for the hidden layers.

The key step is to express the action of $\boldsymbol{M}^\ell(t)$ on the probe using the same initialized weights and training fields that appear in the first-level DMFT.

For the first layer, $\ell = 0$, define

$$\boldsymbol{\psi}_1(\tau) = \underbrace{\frac{1}{\sqrt{D}}\boldsymbol{W}^0(0)\boldsymbol{\psi}_0(\tau)}_{\zeta_1(\tau)} + \eta\gamma_0 \sum_{s<t} \boldsymbol{g}^1(s) C^{h^0\psi_0}(s,\tau) \tag{140}$$

$$\boldsymbol{\psi}_2(\tau) = \underbrace{\frac{\sqrt{D}}{N}\boldsymbol{W}^0(0)^\top\boldsymbol{\psi}_1(\tau)}_{\zeta_2(\tau)} + \eta\gamma_0 \sum_{s<t} \boldsymbol{h}^0(s) C^{g^1\psi_1}(s,\tau) \tag{141}$$

where

$$C^{h^0\psi_0}(s,\tau) = \frac{1}{D}\boldsymbol{h}^0(s)\cdot\boldsymbol{\psi}_0(\tau), \qquad C^{g^1\psi_1}(s,\tau) = \frac{1}{N}\boldsymbol{g}^1(s)\cdot\boldsymbol{\psi}_1(\tau). \tag{142}$$

By construction

$$\boldsymbol{\psi}_2(\tau) = \boldsymbol{M}^0(t)\boldsymbol{\psi}_0(\tau). \tag{143}$$

For hidden layers $1 \le \ell \le L-1$, define instead

$$\boldsymbol{\psi}_1(\tau) = \underbrace{\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\boldsymbol{\psi}_0(\tau)}_{\zeta_1(\tau)} + \eta\gamma_0 \sum_{s<t} \boldsymbol{g}^{\ell+1}(s) C^{h^\ell\psi_0}(s,\tau) \tag{144}$$

$$\boldsymbol{\psi}_2(\tau) = \underbrace{\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)^\top\boldsymbol{\psi}_1(\tau)}_{\zeta_2(\tau)} + \eta\gamma_0 \sum_{s<t} \boldsymbol{h}^\ell(s) C^{g^{\ell+1}\psi_1}(s,\tau) \tag{145}$$

with

$$C^{h^\ell\psi_0}(s,\tau) = \frac{1}{N}\boldsymbol{h}^\ell(s)\cdot\boldsymbol{\psi}_0(\tau), \qquad C^{g^{\ell+1}\psi_1}(s,\tau) = \frac{1}{N}\boldsymbol{g}^{\ell+1}(s)\cdot\boldsymbol{\psi}_1(\tau). \tag{146}$$

Again

$$\boldsymbol{\psi}_2(\tau) = \boldsymbol{M}^\ell(t)\boldsymbol{\psi}_0(\tau). \tag{147}$$

The spectral problem therefore introduces new probe fields $\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ and new mixed order parameters between the probe and the training fields. These mixed order parameters play the role of projected resolvents onto the finite-dimensional subspace generated by the training dynamics.

E.4 Path-Integral (MGF)

We now aim to characterize the joint distribution of our training fields and spectral probe $\{\boldsymbol{v}(t), \boldsymbol{\Delta}(t), \boldsymbol{h}^0(t), \{\boldsymbol{h}^\ell(t), \boldsymbol{g}^\ell(t)\}_{\ell=1}^L, \boldsymbol{\psi}_0\}$ over draws of the quenched disorder $\mathcal{D} \equiv \{\boldsymbol{X}, \{\boldsymbol{W}^\ell(0)\}_{\ell=0}^L\}$ represented by data and initialization. To do so, we will start by calculating the moment generating function of these fields with a Martin-Siggia-Rose integral over the trajectories

$$Z\left[\boldsymbol{j}_v, \boldsymbol{j}_\Delta, \{\boldsymbol{j}_{h^\ell}, \boldsymbol{j}_{\boldsymbol{g}^\ell}\}_{\ell\in[L]}, \boldsymbol{j}_\psi\right]$$
$$= \left\langle \exp\left(\sum_t (\boldsymbol{j}_v(t)\cdot\boldsymbol{v}(t) + \boldsymbol{j}_\Delta(t)\cdot\Delta(t) + \sum_\ell \boldsymbol{j}_{h^\ell}(t)\cdot\boldsymbol{h}^\ell(t) + \sum_\ell \boldsymbol{j}_{g^\ell}(t)\cdot\boldsymbol{g}^\ell(t) + \sum_\tau \boldsymbol{j}_\psi(\tau)\cdot\boldsymbol{\psi}_0(\tau) \right) \right\rangle. \tag{148}$$

As in the previous section, imposing all the fields definitions that appear in equation 148 leads to isolate each disorder contribution to then perform easily the averages.

E.5 Data average

The relevant terms involving $\boldsymbol{X}$ are

$$\left\langle \exp\left[-i \operatorname{Tr} \boldsymbol{X}^\top \sum_t \left(\frac{\sqrt{D}}{P}\boldsymbol{\Delta}(t)\hat{\boldsymbol{h}}^0(t)^\top + \frac{1}{\sqrt{D}}\hat{\boldsymbol{\Delta}}(t)\boldsymbol{v}(t)^\top \right) \right] \right\rangle_{\boldsymbol{X}}. \tag{149}$$

Averaging over the Gaussian data gives

$$\exp\left[-\frac{1}{2}\sum_{tt'} C_v(t,t')\hat{\boldsymbol{\Delta}}(t)\cdot\hat{\boldsymbol{\Delta}}(t') - \frac{1}{2\alpha}\sum_{tt'} C_\Delta(t,t')\hat{\boldsymbol{h}}^0(t)\cdot\hat{\boldsymbol{h}}^0(t') \right]$$
$$\times \exp\left[D\sum_{tt'} R_\Delta(t',t) R^0_{vu}(t,t') \right] \tag{150}$$

where

$$C_v(t,t') = \frac{1}{D}\boldsymbol{v}(t)\cdot\boldsymbol{v}(t') \qquad C_\Delta(t,t') = \frac{1}{P}\boldsymbol{\Delta}(t)\cdot\boldsymbol{\Delta}(t') \tag{151}$$

$$R_\Delta(t,t') = -\frac{i}{P}\boldsymbol{\Delta}(t)\cdot\hat{\boldsymbol{\Delta}}(t') \qquad R^0_{vu}(t,t') = -\frac{i}{D}\boldsymbol{v}(t)\cdot\hat{\boldsymbol{h}}^0(t'). \tag{152}$$

To enforce these definitions, we use Fourier representation of Dirac delta function, for instance

$$1 \equiv \int \frac{dR_\Delta(t,t')d\hat{R}_\Delta(t,t')}{2\pi} \exp\left(- P\sum_{tt'} R_\Delta(t,t')\hat{R}_\Delta(t,t') - i\sum_{tt'} \hat{R}_\Delta(t,t')\boldsymbol{\Delta}(t)\cdot\hat{\boldsymbol{\Delta}}(t')\right) \tag{153}$$

$$1 \equiv \int \frac{dR^0_{vu}(t,t')d\hat{R}^0_{vu}(t,t')}{2\pi} \exp\left(- D\sum_{tt'} R^0_{vu}(t,t')\hat{R}^0_{vu}(t,t') - i\sum_{tt'} \hat{R}^0_{vu}(t,t')\boldsymbol{v}(t)\cdot\hat{\boldsymbol{h}}^0(t')\right) \tag{154}$$

At the saddle point, differentiating with respect to R_Δ and R^0_{vu} gives

$$\hat{R}_\Delta(t',t) = \frac{1}{\alpha}R^0_{vu}(t,t') \tag{155}$$

$$\hat{R}^0_{vu}(t,t') = R_\Delta(t',t). \tag{156}$$

These terms encode the memory induced by reusing the same random dataset in full-batch gradient descent.

E.6 First-layer weight average

We now average over the first-layer initialization $\boldsymbol{W}^0(0)$

$$\begin{aligned}
&\left\langle \exp\left(-\frac{i}{\sqrt{D}}\mathrm{Tr}\boldsymbol{W}^0(0)^\top\left(\sum_t \hat{\boldsymbol{\chi}}^1(t)\boldsymbol{h}^0(t)^\top + \frac{1}{\gamma_0\nu}\sum_t \boldsymbol{g}^1(t)\hat{\boldsymbol{\xi}}^0(t)^\top + \sum_\tau \hat{\boldsymbol{\zeta}}_1(\tau)\boldsymbol{\psi}_0(\tau)^\top + \frac{1}{\nu}\sum_\tau \boldsymbol{\psi}_1(\tau)\hat{\boldsymbol{\zeta}}_2(\tau)^\top\right)\right)\right\rangle \\
&= \exp\left(- \frac{1}{2}\sum_{tt'} C^0_h(t,t')\hat{\boldsymbol{\chi}}^1(t)\cdot\hat{\boldsymbol{\chi}}^1(t') - \frac{1}{2\gamma_0^2\nu}\sum_{tt'} C^1_g(t,t')\hat{\boldsymbol{\xi}}^0(t)\cdot\hat{\boldsymbol{\xi}}^0(t')\right) \\
&\times \exp\left(- \frac{1}{2}\sum_{\tau\tau'} C^{\psi_0}(\tau,\tau')\hat{\boldsymbol{\zeta}}_1(\tau)\cdot\hat{\boldsymbol{\zeta}}_1(\tau') - \frac{1}{2\nu}\sum_{\tau\tau'} C^{\psi_1}(\tau,\tau')\hat{\boldsymbol{\zeta}}_2(\tau)\cdot\hat{\boldsymbol{\zeta}}_2(\tau')\right) \\
&\times \exp\left(+ \frac{D}{\gamma_0}\sum_{t,t'} R^1_{gu}(t',t)R^0_{hr}(t,t') - \sum_{t\tau} C^{h^0\psi_0}(t,\tau)\hat{\boldsymbol{\chi}}^1(t)\cdot\hat{\boldsymbol{\zeta}}_1(\tau)\right) \\
&\times \exp\left(+ D\sum_{t\tau} R^{\psi_1 u^1}(\tau,t)R^{h_0\zeta_2}(t,\tau) + \frac{D}{\gamma_0}\sum_{t\tau} R^{\psi_0 r^0}(\tau,t)R^{g^1\zeta_1}(t,\tau)\right) \\
&\times \exp\left(- \frac{1}{\gamma_0\nu}\sum_{t\tau} C^{g^1\psi_1}(t,\tau)\hat{\boldsymbol{\xi}}^0(t)\cdot\hat{\boldsymbol{\zeta}}_2(\tau) + D\sum_{\tau\tau'} R^{\psi_1\zeta_1}(\tau,\tau')R^{\psi_0\zeta_2}(\tau',\tau)\right).
\end{aligned} \tag{157}$$

This Gaussian average over $\boldsymbol{W}^0(0)$ generates covariance terms for each pair of fields, together with mixed training–probe response terms. Define

$$C^0_h(s,s') = \frac{1}{D}\boldsymbol{h}^0(s)\cdot\boldsymbol{h}^0(s'), \qquad C^1_g(s,s') = \frac{1}{N}\boldsymbol{g}^1(s)\cdot\boldsymbol{g}^1(s'), \tag{158}$$

$$C^{\psi_0}(\tau,\tau') = \frac{1}{D}\boldsymbol{\psi}_0(\tau)\cdot\boldsymbol{\psi}_0(\tau'), \qquad C^{\psi_1}(\tau,\tau') = \frac{1}{N}\boldsymbol{\psi}_1(\tau)\cdot\boldsymbol{\psi}_1(\tau'), \tag{159}$$

$$C^{h^0\psi_0}(s,\tau) = \frac{1}{D}\boldsymbol{h}^0(s)\cdot\boldsymbol{\psi}_0(\tau), \qquad C^{g^1\psi_1}(s,\tau) = \frac{1}{N}\boldsymbol{g}^1(s)\cdot\boldsymbol{\psi}_1(\tau). \tag{160}$$

The response-type order parameters generated by the same average are

$$R^0_{hr}(s,s') = -\frac{i}{D}\boldsymbol{h}^0(s)\cdot\hat{\boldsymbol{\xi}}^0(s'), \qquad R^1_{gu}(s,s') = -\frac{i}{N}\boldsymbol{g}^1(s)\cdot\hat{\boldsymbol{\chi}}^1(s'), \tag{161}$$

$$R^{h^0\zeta_2}(s,\tau) = -\frac{i}{D}\boldsymbol{h}^0(s)\cdot\hat{\boldsymbol{\zeta}}_2(\tau), \qquad R^{g^1\zeta_1}(s,\tau) = -\frac{i}{N}\boldsymbol{g}^1(s)\cdot\hat{\boldsymbol{\zeta}}_1(\tau), \tag{162}$$

$$R^{\psi_0 r^0}(\tau,s) = -\frac{i}{D}\boldsymbol{\psi}_0(\tau)\cdot\hat{\boldsymbol{\xi}}^0(s), \qquad R^{\psi_1 u^1}(\tau,s) = -\frac{i}{N}\boldsymbol{\psi}_1(\tau)\cdot\hat{\boldsymbol{\chi}}^1(s), \tag{163}$$

$$R^{\psi_0\zeta_2}(\tau,\tau') = -\frac{i}{D}\boldsymbol{\psi}_0(\tau)\cdot\hat{\boldsymbol{\zeta}}_2(\tau'), \qquad R^{\psi_1\zeta_1}(\tau,\tau') = -\frac{i}{N}\boldsymbol{\psi}_1(\tau)\cdot\hat{\boldsymbol{\zeta}}_1(\tau'). \tag{164}$$

E.7 HIDDEN LAYERS WEIGHT AVERAGE

For $1 \le \ell \le L-1$ the weight matrix $\boldsymbol{W}^\ell$ is square. The Gaussian average is

$$\left\langle \exp\left(-\frac{i}{\sqrt{N}} \mathrm{Tr} \boldsymbol{W}^\ell(0)^\top \left(\sum_t \hat{\boldsymbol{\chi}}^{\ell+1}(t) \boldsymbol{h}^\ell(t)^\top + \sum_t \boldsymbol{g}^{\ell+1}(t) \hat{\boldsymbol{\xi}}^\ell(t)^\top + \sum_\tau \hat{\boldsymbol{\zeta}}_1(\tau) \boldsymbol{\psi}_0(\tau)^\top + \sum_\tau \boldsymbol{\psi}_1(\tau) \hat{\boldsymbol{\zeta}}_2(\tau)^\top \right) \right) \right\rangle \tag{165}$$

$$= \exp\left(-\frac{1}{2} \sum_{tt'} C_h^\ell(t,t') \hat{\boldsymbol{\chi}}^{\ell+1}(t) \cdot \hat{\boldsymbol{\chi}}^{\ell+1}(t') - \frac{1}{2} \sum_{tt'} C_g^{\ell+1}(t,t') \hat{\boldsymbol{\xi}}^\ell(t) \cdot \hat{\boldsymbol{\xi}}^\ell(t') \right) \tag{166}$$

$$\times \exp\left(-\frac{1}{2} \sum_{\tau\tau'} C^{\psi_0}(\tau,\tau') \hat{\boldsymbol{\zeta}}_1(\tau) \cdot \hat{\boldsymbol{\zeta}}_1(\tau') - \frac{1}{2} \sum_{\tau\tau'} C^{\psi_1}(\tau,\tau') \hat{\boldsymbol{\zeta}}_2(\tau) \cdot \hat{\boldsymbol{\zeta}}_2(\tau') \right) \tag{167}$$

$$\times \exp\left(N \sum_{tt'} R_{gu}^{\ell+1}(t',t) R_{hr}^\ell(t,t') - \sum_{t\tau} C^{h^\ell \psi_0}(t,\tau) \hat{\boldsymbol{\chi}}^{\ell+1}(t) \cdot \hat{\boldsymbol{\zeta}}_1(\tau) \right) \tag{168}$$

$$\times \exp\left(N \sum_{t\tau} R^{\psi_1 u^{\ell+1}}(\tau,t) R^{h^\ell \zeta_2}(t,\tau) + N \sum_{t\tau} R^{\psi_0 r^\ell}(\tau,t) R^{g^{\ell+1}\zeta_1}(t,\tau) \right) \tag{169}$$

$$\times \exp\left(-\sum_{t\tau} C^{g^{\ell+1}\psi_1}(t,\tau) \hat{\boldsymbol{\xi}}^\ell(t) \cdot \hat{\boldsymbol{\zeta}}_2(\tau) + N \sum_{\tau\tau'} R^{\psi_1\zeta_1}(\tau,\tau') R^{\psi_0\zeta_2}(\tau',\tau) \right) \tag{170}$$

which yields the same structure as for the first layer, but without the aspect-ratio ν which comes from having a rectangular first layer $\boldsymbol{W}^0$. The hidden-layer weight average generates the following correlation functions

$$C_h^\ell(t,t') = \frac{1}{N} \boldsymbol{h}^\ell(t) \cdot \boldsymbol{h}^\ell(t'), \qquad C_g^{\ell+1}(t,t') = \frac{1}{N} \boldsymbol{g}^{\ell+1}(t) \cdot \boldsymbol{g}^{\ell+1}(t'),$$

$$C^{\psi_0}(\tau,\tau') = \frac{1}{N} \boldsymbol{\psi}_0(\tau) \cdot \boldsymbol{\psi}_0(\tau'), \qquad C^{\psi_1}(\tau,\tau') = \frac{1}{N} \boldsymbol{\psi}_1(\tau) \cdot \boldsymbol{\psi}_1(\tau'),$$

and the mixed training-probe correlations

$$C^{h^\ell \psi_0}(t,\tau) = \frac{1}{N} \boldsymbol{h}^\ell(t) \cdot \boldsymbol{\psi}_0(\tau), \qquad C^{g^{\ell+1}\psi_1}(t,\tau) = \frac{1}{N} \boldsymbol{g}^{\ell+1}(t) \cdot \boldsymbol{\psi}_1(\tau).$$

The corresponding response functions are defined by contractions with the MSR conjugate fields. The training response functions are

$$R_{hr}^\ell(t,t') = -\frac{i}{N} \boldsymbol{h}^\ell(t) \cdot \widehat{\boldsymbol{\xi}}^\ell(t'), \qquad R_{gu}^{\ell+1}(t,t') = -\frac{i}{N} \boldsymbol{g}^{\ell+1}(t) \cdot \widehat{\boldsymbol{\chi}}^{\ell+1}(t').$$

The mixed probe-training response functions are

$$R^{\psi_0 r^\ell}(\tau,t) = -\frac{i}{N} \boldsymbol{\psi}_0(\tau) \cdot \widehat{\boldsymbol{\xi}}^\ell(t), \qquad R^{\psi_1 u^{\ell+1}}(\tau,t) = -\frac{i}{N} \boldsymbol{\psi}_1(\tau) \cdot \widehat{\boldsymbol{\chi}}^{\ell+1}(t),$$

$$R^{h^\ell \zeta_2}(t,\tau) = -\frac{i}{N} \boldsymbol{h}^\ell(t) \cdot \widehat{\boldsymbol{\zeta}}_2(\tau), \qquad R^{g^{\ell+1}\zeta_1}(t,\tau) = -\frac{i}{N} \boldsymbol{g}^{\ell+1}(t) \cdot \widehat{\boldsymbol{\zeta}}_1(\tau).$$

Finally, the pure probe response functions are

$$R^{\psi_1\zeta_1}(\tau,\tau') = -\frac{i}{N} \boldsymbol{\psi}_1(\tau) \cdot \widehat{\boldsymbol{\zeta}}_1(\tau'), \qquad R^{\psi_0\zeta_2}(\tau,\tau') = -\frac{i}{N} \boldsymbol{\psi}_0(\tau) \cdot \widehat{\boldsymbol{\zeta}}_2(\tau').$$

E.8 LAST-LAYER READOUT AVERAGE

The readout initialization only appears in

$$\exp\Big[-i \boldsymbol{w}^L(0) \cdot \sum_t \hat{\boldsymbol{g}}^L(t) \Big]. \tag{171}$$

Thus

$$\left\langle \exp\left(-i\boldsymbol{w}^L(0)\cdot\sum_t \hat{\boldsymbol{g}}^L(t)\right)\right\rangle_{\boldsymbol{w}^L(0)} = \exp\Big[-\frac{1}{2}\sum_{tt'}\hat{\boldsymbol{g}}^L(t)\cdot\hat{g}^L(t')\Big]. \tag{172}$$

Equivalently, after Hubbard linearization, the readout contributes as a Gaussian process $r^L(t)$ with covariance

$$\langle r^L(t)r^L(t')\rangle = 1. \tag{173}$$

The covariance is one for all pairs of times because the same readout initialization is reused at every training step.

E.9 From the averaged action to single-site processes

All correlations and responses after the averages can be now enforced by delta-function identities. After these insertions, the joint generating functional has the schematic form

$$Z[\mathcal{J}] = \int \mathrm{d}\mathcal{Q}\mathrm{d}\widehat{\mathcal{Q}}\,\exp\Big[-D\mathcal{S}(\mathcal{Q},\widehat{\mathcal{Q}};\mathcal{J})\Big] \tag{174}$$

where $\mathcal{Q}$ collects all training and probe order parameters.

The key factorization is as follows. Once the disorder has been averaged, distinct data samples μ, input coordinates k, and hidden neurons i no longer interact directly. They interact only through the macroscopic order parameters $\mathcal{Q}$. Therefore the microscopic path integral factorizes as

$$Z[\mathcal{J}] = \int \mathrm{d}\mathcal{Q}\mathrm{d}\widehat{\mathcal{Q}}\,e^{-D\boldsymbol{\Phi}(\mathcal{Q},\widehat{\mathcal{Q}})}\prod_{\mu=1}^{P} Z_\Delta^{(\mu)}[\mathcal{Q}]\prod_{k=1}^{D} Z_0^{(k)}[\mathcal{Q}]\prod_{\ell=1}^{L}\prod_{i=1}^{N} Z_\ell^{(i)}[\mathcal{Q}]\times Z_{\mathrm{pr}}^{(a)}[\mathcal{Q}]. \tag{175}$$

Here Z_Δ is the scalar process for one training sample, Z_0 is the scalar process for one input coordinate, and Z_ℓ is the scalar process for one hidden neuron at layer ℓ. The factor $Z_{\mathrm{pr}}^{(a)}$ denotes the same single-site factors augmented by probe variables on the probed layer. At zero source the sites are exchangeable, so

$$\frac{1}{D}\sum_{k=1}^{D} F_k \to \langle F\rangle_{Z_0}, \qquad \frac{1}{N}\sum_{i=1}^{N} F_i \to \langle F\rangle_{Z_\ell}, \qquad \frac{1}{P}\sum_{\mu=1}^{P} F_\mu \to \langle F\rangle_{Z_\Delta}. \tag{176}$$

At the saddle point the conjugates of the correlation order parameters vanish in the zero-source limit, while the response conjugates produce the retarded memory kernels. The remaining quadratic forms in the Fourier fields are linearized by Hubbard transformations, e.g.

$$\exp\left[-\frac{1}{2}\sum_{tt'}\widehat{\Delta}(t)C_v(t,t')\widehat{\Delta}(t')\right] = \left\langle e^{-i\sum_t \widehat{\Delta}(t)u_\Delta(t)}\right\rangle_{u_\Delta\sim GP(0,C_v)}, \tag{177}$$

$$\exp\left[-\frac{1}{2\alpha}\sum_{tt'}\widehat{h}^0(t)C_\Delta(t,t')\widehat{h}^0(t')\right] = \left\langle e^{-i\sum_t \widehat{h}^0(t)u^0(t)}\right\rangle_{u^0\sim GP(0,\alpha^{-1}C_\Delta)}, \tag{178}$$

$$\exp\left[-\frac{1}{2\gamma_0^2\nu}\sum_{tt'}\widehat{\xi}^0(t)C_g^1(t,t')\widehat{\xi}^0(t')\right] = \left\langle e^{-i\sum_t \widehat{\xi}^0(t)r^0(t)}\right\rangle_{r^0\sim GP(0,(\gamma_0^2\nu)^{-1}C_g^1)}, \tag{179}$$

$$\exp\left[-\frac{1}{2}\sum_{tt'}\widehat{\chi}^\ell(t)C_h^{\ell-1}(t,t')\widehat{\chi}^\ell(t')\right] = \left\langle e^{-i\sum_t \widehat{\chi}^\ell(t)u^\ell(t)}\right\rangle_{u^\ell\sim GP(0,C_h^{\ell-1})}, \tag{180}$$

$$\exp\left[-\frac{1}{2}\sum_{tt'}\widehat{\xi}^\ell(t)C_g^{\ell+1}(t,t')\widehat{\xi}^\ell(t')\right] = \left\langle e^{-i\sum_t \widehat{\xi}^\ell(t)r^\ell(t)}\right\rangle_{r^\ell\sim GP(0,C_g^{\ell+1})}. \tag{181}$$

Integrating over the Fourier variables then gives stochastic single-site equations. The response functions can equivalently be read as functional derivatives with respect to the Gaussian noises. For example,

$$R_\Delta(t,t') = \left\langle \frac{\partial\Delta(t)}{\partial u_\Delta(t')}\right\rangle, \qquad R_{vu}^0(t,t') = \left\langle \frac{\partial v(t)}{\partial u^0(t')}\right\rangle. \tag{182}$$

E.10 Deterministic limiting order parameters

For fixed L and fixed training horizon, the action is proportional to D. Hence the macroscopic order parameters concentrate at the saddle point. In particular,

$$\frac{1}{D}\boldsymbol{v}(t)\cdot\boldsymbol{v}(t') \xrightarrow[D\to\infty]{} C_v(t,t') = \langle v(t)v(t')\rangle, \tag{183}$$

$$\frac{1}{P}\boldsymbol{\Delta}(t)\cdot\boldsymbol{\Delta}(t') \xrightarrow[D\to\infty]{} C_\Delta(t,t') = \langle \Delta(t)\Delta(t')\rangle, \tag{184}$$

$$\frac{1}{D}\boldsymbol{h}^0(t)\cdot\boldsymbol{h}^0(t') \xrightarrow[D\to\infty]{} C_h^0(t,t') = \langle h^0(t)h^0(t')\rangle, \tag{185}$$

$$\frac{1}{N}\boldsymbol{h}^\ell(t)\cdot\boldsymbol{h}^\ell(t') \xrightarrow[D\to\infty]{} C_h^\ell(t,t') = \langle h^\ell(t)h^\ell(t')\rangle, \tag{186}$$

$$\frac{1}{N}\boldsymbol{g}^\ell(t)\cdot\boldsymbol{g}^\ell(t') \xrightarrow[D\to\infty]{} C_g^\ell(t,t') = \langle g^\ell(t)g^\ell(t')\rangle. \tag{187}$$

Similarly,

$$-\frac{i}{P}\boldsymbol{\Delta}(t)\cdot\widehat{\boldsymbol{\Delta}}(t') \to R_\Delta(t,t') = \left\langle \frac{\partial\Delta(t)}{\partial u_\Delta(t')}\right\rangle, \tag{188}$$

$$-\frac{i}{D}\boldsymbol{v}(t)\cdot\widehat{\boldsymbol{h}}^0(t') \to R_{vu}^0(t,t') = \left\langle \frac{\partial v(t)}{\partial u^0(t')}\right\rangle, \tag{189}$$

$$-\frac{i}{D}\boldsymbol{h}^0(t)\cdot\widehat{\boldsymbol{\xi}}^0(t') \to R_{hr}^0(t,t') = \left\langle \frac{\partial h^0(t)}{\partial r^0(t')}\right\rangle, \tag{190}$$

$$-\frac{i}{N}\boldsymbol{h}^\ell(t)\cdot\widehat{\boldsymbol{\xi}}^\ell(t') \to R_{hr}^\ell(t,t') = \left\langle \frac{\partial h^\ell(t)}{\partial r^\ell(t')}\right\rangle, \tag{191}$$

$$-\frac{i}{N}\boldsymbol{g}^\ell(t)\cdot\widehat{\boldsymbol{\chi}}^\ell(t') \to R_{gu}^\ell(t,t') = \left\langle \frac{\partial g^\ell(t)}{\partial u^\ell(t')}\right\rangle. \tag{192}$$

The mixed probe order parameters also converge to deterministic single-site averages. For a hidden probed layer a,

$$C^{h^a\psi_0}(t,\tau) \to \langle h^a(t)\psi_0(\tau)\rangle, \qquad C^{g^{a+1}\psi_1}(t,\tau) \to \langle g^{a+1}(t)\psi_1(\tau)\rangle, \tag{193}$$

$$R^{\psi_0 r^a}(\tau,t) \to \left\langle \frac{\delta\psi_0(\tau)}{\delta r^a(t)}\right\rangle, \qquad R^{\psi_1 u^{a+1}}(\tau,t) \to \left\langle \frac{\delta\psi_1(\tau)}{\delta u^{a+1}(t)}\right\rangle. \tag{194}$$

The first-layer version is identical except for the normalizations by D, N and the factors of ν, γ_0 already displayed in the first-layer average.

E.11 Training-level DMFT

The zero-source training DMFT is obtained from the joint saddle by setting the probe sources to zero after the disorder average has been performed. This gives a closed scalar stochastic process for one data point, one input coordinate, and one representative hidden neuron in each layer. The probe does not alter this training process; it only probes the solution obtained from it.

The data sector is

$$\Delta(t) = u_\Delta(t) + \frac{1}{\alpha}\sum_{s<t} R_{vu}^0(t,s)\Delta(s) + \sigma\epsilon, \qquad u_\Delta \sim GP(0, C_v), \quad \epsilon\sim\mathcal{N}(0,1), \tag{195}$$

$$h^0(t) = u^0(t) + \sum_{s<t} R_\Delta(t,s)v(s), \qquad u^0 \sim GP(0, \frac{1}{\alpha}C_\Delta). \tag{196}$$

The first-layer boundary is

$$h^1(t) = u^1(t) + \sum_{s<t}\left[\frac{1}{\gamma_0\nu}R_{hr}^0(t,s) + \eta\gamma_0 C_h^0(t,s)\right] g^1(s), \quad u^1(t)\sim GP(0, C_h^0), \tag{197}$$

$$v(t) = w_\star - r^0(t) - \sum_{s<t}\left[\frac{1}{\gamma_0}R_{gu}^1(t,s) + \eta C_g^1(t,s)\right] h^0(s), \quad r^0(t)\sim GP(0, \frac{1}{\gamma_0^2\nu}C_g^1). \tag{198}$$

For the remaining forward fields, $2 \le \ell \le L$,

$$h^\ell(t) = u^\ell(t) + \sum_{s<t} \left[R_{hr}^{\ell-1}(t,s) + \eta\gamma_0 C_h^{\ell-1}(t,s)\right] g^\ell(s), \quad u^\ell \sim GP(0, C_h^{\ell-1}). \tag{199}$$

For the backward fields below the readout, $1 \le \ell \le L-1$,

$$g^\ell(t) = r^\ell(t) + \sum_{s<t} \left[R_{gu}^{\ell+1}(t,s) + \eta\gamma_0 C_g^{\ell+1}(t,s)\right] h^\ell(s), \quad r^\ell \sim GP(0, C_g^{\ell+1}). \tag{200}$$

The readout boundary is

$$g^L(t) = r^L + \eta\gamma_0 \sum_{s<t} \left[u^L(s) + \sum_{r<s} \left(R_{hr}^{L-1}(s,r) + \eta\gamma_0 C_h^{L-1}(s,r)\right) g^L(r)\right], \quad \langle r^L(t) r^L(t')\rangle = 1. \tag{201}$$

The correlation fixed points are

$$C_v(t,t') = \langle v(t)v(t')\rangle, \qquad C_\Delta(t,t') = \langle \Delta(t)\Delta(t')\rangle, \tag{202}$$

$$C_h^\ell(t,t') = \langle h^\ell(t)h^\ell(t')\rangle, \qquad C_g^\ell(t,t') = \langle g^\ell(t)g^\ell(t')\rangle. \tag{203}$$

The response fixed points are

$$R_\Delta(t,t') = \left\langle \frac{\partial \Delta(t)}{\partial u_\Delta(t')} \right\rangle, \qquad R_{vu}^0(t,t') = \left\langle \frac{\partial v(t)}{\partial u^0(t')} \right\rangle, \tag{204}$$

$$R_{hr}^0(t,t') = \left\langle \frac{\partial h^0(t)}{\partial r^0(t')} \right\rangle, \qquad R_{hr}^\ell(t,t') = \left\langle \frac{\partial h^\ell(t)}{\partial r^\ell(t')} \right\rangle, \tag{205}$$

$$R_{gu}^\ell(t,t') = \left\langle \frac{\partial g^\ell(t)}{\partial u^\ell(t')} \right\rangle, \qquad 1 \le \ell \le L. \tag{206}$$

Finally,

$$\mathcal{L}_{\text{test}}(t) = C_v(t,t) + \sigma^2, \qquad \widehat{\mathcal{L}}_{\text{train}}(t) = C_\Delta(t,t). \tag{207}$$

E.12 Probe-level DMFT fixed points

The probe equations are solved after the training-level DMFT has produced $C_h^\ell, C_g^\ell, R_{hr}^\ell, R_{gu}^\ell$. Thus the training process and the spectral probe are separated at the final stage: the training order parameters are inputs to the probe fixed point.

E.12.1 First-layer probe

For $\ell = 0$, the effective first-layer probe process is

$$\begin{aligned}\psi_1(\tau) = u_{\zeta_1}(\tau) &+ \frac{1}{\nu}\int d\tau'\, R^{\psi_0\zeta_2}(\tau,\tau')\psi_1(\tau') \\ &+ \sum_{s<t}\left[\eta\gamma_0 C^{h^0\psi_0}(s,\tau) + \frac{1}{\gamma_0\nu} R^{\psi_0 r^0}(\tau,s)\right] g^1(s)\end{aligned} \tag{208}$$

$$\begin{aligned}\psi_2(\tau) = u_{\zeta_2}(\tau) &+ \int d\tau'\, R^{\psi_1\zeta_1}(\tau,\tau')\psi_0(\tau') \\ &+ \sum_{s<t}\left[\eta\gamma_0 C^{g^1\psi_1}(s,\tau) + R^{\psi_1 u^1}(\tau,s)\right] h^0(s).\end{aligned} \tag{209}$$

The Gaussian noises have covariances

$$\langle u_{\zeta_1}(\tau) u_{\zeta_1}(\tau')\rangle = C^{\psi_0}(\tau,\tau'), \qquad \langle u_{\zeta_2}(\tau) u_{\zeta_2}(\tau')\rangle = \frac{1}{\nu} C^{\psi_1}(\tau,\tau'), \tag{210}$$

and mixed covariances

$$\left\langle u_{\zeta_1}(\tau) u^1(t)\right\rangle = C^{h^0\psi_0}(t,\tau), \qquad \left\langle u_{\zeta_2}(\tau) r^0(t)\right\rangle = \frac{1}{\gamma_0\nu} C^{g^1\psi_1}(t,\tau). \tag{211}$$

These equations are the spectral-probe analogue of the first-layer single-site training equations.

To obtain the outlier condition, Fourier transform in the spectral time τ and set

$$z = i\omega, \qquad \mathcal{H}(z) = \left[z - R^{\psi_1\zeta_1}(z)\right]^{-1}.$$

The ψ_2 equation gives

$$\left[i\omega - R^{\psi_1\zeta_1}(\omega)\right]\psi_0(\omega) = u_{\zeta_2}(\omega) + \sum_{s<t}\left[\eta\gamma_0 C^{g^1\psi_1}(s,\omega) + R^{\psi_1 u^1}(\omega,s)\right]h^0(s). \tag{212}$$

Projecting onto h^0 once you stack the training time T in T-dimensional vectors, and differentiating with respect to r^0 gives

$$\Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)\boldsymbol{C}^{h^0\psi_0}(\omega) = \langle\zeta_2(\omega)\boldsymbol{h}^0\rangle + \boldsymbol{C}_h^0\Big(\eta\gamma_0\boldsymbol{C}^{g^1\psi_1}(\omega) + \boldsymbol{R}^{\psi_1,u^1}(\omega)\Big) \tag{213}$$

$$\Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)\boldsymbol{R}^{\psi_0 r^0}(\omega) = (\boldsymbol{R}_{hr}^0)^\top\Big(\eta\gamma_0\boldsymbol{C}^{g^1\psi_1}(\omega) + \boldsymbol{R}^{\psi_1,u^1}(\omega)\Big) \tag{214}$$

where we use Stein's lemma to write

$$\left\langle u_{\zeta_2}(z)\boldsymbol{h}^0\right\rangle = \frac{1}{\gamma_0\nu}\boldsymbol{R}_{hr}^0\boldsymbol{C}^{g^1\psi_1}(z). \tag{215}$$

Similarly, the ψ_1 equation gives

$$\Big(1+\nu^{-1}\mathcal{H}(\omega)\Big)\psi_1(\omega) = u_{\zeta_1}(\omega) + \sum_{s<t}\Big(\eta\gamma_0 C^{h^0,\psi_0}(s,\tau) + \frac{1}{\gamma_0\nu}R^{\psi_0,r^0}(\tau,s)\Big)g^1(s) \tag{216}$$

where

$$\mathcal{H}(\omega) = -R_{0,2}(\omega) = \Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)^{-1}. \tag{217}$$

Projecting onto g^1 and differentiating with respect to u^1 gives

$$\Big(1+\nu^{-1}\mathcal{H}(\omega)\Big)\boldsymbol{C}^{g^1\psi_1}(\omega) = \langle u_{\zeta_1}(\omega)\boldsymbol{g}^1\rangle + \boldsymbol{C}_g^1\Big(\eta\gamma_0\boldsymbol{C}^{h^0,\psi_0}(\omega) + \frac{1}{\gamma_0\nu}\boldsymbol{R}^{\psi_0,r^0}(\omega)\Big) \tag{218}$$

$$\Big(1+\nu^{-1}\mathcal{H}(\omega)\Big)\boldsymbol{R}^{\psi_1 u^1}(\omega) = (\boldsymbol{R}_{gu}^1)^\top\Big(\eta\gamma_0\boldsymbol{C}^{h^0,\psi_0}(\omega) + \frac{1}{\gamma_0\nu}\boldsymbol{R}^{\psi_0,r^0}(\omega)\Big). \tag{219}$$

Here, again, we used Stein's lemma to express

$$\left\langle u_{\zeta_1}(\omega)\boldsymbol{g}^1\right\rangle = \boldsymbol{R}_{gu}^1\boldsymbol{C}^{h^0\psi_0}(\omega). \tag{220}$$

Collecting the four projected quantities and with the analytic continuation from $z = i\omega$ to the real spectral axis one gets

$$\boldsymbol{q}(z) = \begin{pmatrix}\boldsymbol{C}^{h^0\psi_0}(z)\\ \boldsymbol{R}^{\psi_0 r^0}(z)\\ \boldsymbol{C}^{g^1\psi_1}(z)\\ \boldsymbol{R}^{\psi_1 u^1}(z)\end{pmatrix} \tag{221}$$

from which we obtain the finite-dimensional linear system

$$\boldsymbol{\mathcal{A}}_0(z)\boldsymbol{q}(z) = 0 \tag{222}$$

where

$$\boldsymbol{\mathcal{A}}_0(z) = \begin{pmatrix}(z - R^{\psi_1,\zeta_1}(z))\boldsymbol{I} & \boldsymbol{0} & -\frac{1}{\gamma_0\nu}\boldsymbol{R}_{hr}^0 - \eta\gamma_0\boldsymbol{C}_h^0 & -\boldsymbol{C}_h^0\\ \boldsymbol{0} & (z - R^{\psi_1,\zeta_1}(z))\boldsymbol{I} & -\eta\gamma_0(\boldsymbol{R}_{hr}^0)^\top & -(\boldsymbol{R}_{hr}^0)^\top\\ -\boldsymbol{R}_{gu}^1 - \eta\gamma_0\boldsymbol{C}_g^1 & -\frac{1}{\gamma_0\nu}\boldsymbol{C}_g^1 & (R^{\psi_1,\zeta_1}(z))^{-1}\boldsymbol{I} & \boldsymbol{0}\\ -\eta\gamma_0(\boldsymbol{R}_{gu}^1)^\top & -\frac{1}{\gamma_0\nu}(\boldsymbol{R}_{gu}^1)^\top & \boldsymbol{0} & (R^{\psi_1,\zeta_1}(z))^{-1}\boldsymbol{I}\end{pmatrix} \tag{223}$$

where we remember

$$R^{\psi_1\zeta_1}(z) = \left[1+\nu^{-1}\mathcal{H}(z)\right]^{-1}. \tag{224}$$

The first-layer outliers are therefore obtained from the pole condition

$$\det\boldsymbol{\mathcal{A}}_0(z) = 0. \tag{225}$$

E.12.2 HIDDEN-LAYER PROBE

As before, for $\ell = 1, \dots, L-1$ the probe process generates

$$\psi_1(\tau) = u_{\zeta_1}(\tau) + \int R^{\psi_0,\zeta_2}(\tau,\tau')\psi_1(\tau') + \sum_{s<t}\Big(\eta\gamma_0 C^{h^\ell\psi_0}(s,\tau) + R^{\psi_0,r^\ell}(\tau,s)\Big)g^{\ell+1}(s) \quad (226)$$

$$\psi_2(\tau) = u_{\zeta_2}(\tau) + \int R^{\psi_1,\zeta_1}(\tau,\tau')\psi_0(\tau') + \sum_{s<t}\Big(\eta\gamma_0 C^{g^{\ell+1}\psi_1}(s,\tau) + R^{\psi_1,u^{\ell+1}}(\tau,s)\Big)h^\ell(s) \quad (227)$$

with

$$\langle u_{\zeta_1}(\tau)u_{\zeta_1}(\tau')\rangle = C^{\psi_0}(\tau,\tau'), \quad \langle\zeta_1(\tau)u^{\ell+1}(t)\rangle = C^{h^\ell\psi_0}(t,\tau) \quad (228)$$

$$\langle u_{\zeta_2}(\tau)u_{\zeta_2}(\tau')\rangle = C^{\psi_1}(\tau,\tau'), \quad \langle\zeta_2(\tau)r^\ell(t)\rangle = C^{g^{\ell+1}\psi_1}(t,\tau). \quad (229)$$

Thus, the variable of interest has the form

$$\Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)\psi_0(\omega) = u_{\zeta_2}(\omega) + \sum_{s<t}\Big(\eta\gamma_0 C^{g^{\ell+1}\psi_1}(s,\omega) + R^{\psi_1,u^{\ell+1}}(\omega,s)\Big)h^\ell(s) \quad (230)$$

from which by projecting onto h^ℓ and differentiating with respect to r^ℓ one gets

$$\Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)\boldsymbol{C}^{h^\ell\psi_0}(\omega) = \langle\zeta_2(\omega)\boldsymbol{h}^\ell\rangle + \boldsymbol{C}_h^\ell\Big(\eta\gamma_0\boldsymbol{C}^{g^{\ell+1}\psi_1}(\omega) + \boldsymbol{R}^{\psi_1,u^{\ell+1}}(\omega)\Big) \quad (231)$$

$$\Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)\boldsymbol{R}^{\psi_0 r^\ell}(\omega) = (\boldsymbol{R}_{hr}^\ell)^\top\Big(\eta\gamma_0\boldsymbol{C}^{g^{\ell+1}\psi_1}(\omega) + \boldsymbol{R}^{\psi_1,u^{\ell+1}}(\omega)\Big). \quad (232)$$

We can again apply Stein's Lemma for

$$\langle\zeta_2(\omega)\boldsymbol{h}^\ell\rangle = \boldsymbol{R}_{hr}^\ell\boldsymbol{C}^{g^{\ell+1}\psi_1}(\omega). \quad (233)$$

Similarly for $\psi_1(\omega)$

$$\Big(1 + \mathcal{H}(\omega)\Big)\psi_1(\omega) = u_{\zeta_1}(\omega) + \sum_{s<t}\Big(\eta\gamma_0 C^{h^\ell,\psi_0}(s,\omega) + R^{\psi_0,r^\ell}(\omega,s)\Big)g^{\ell+1}(s) \quad (234)$$

with $\mathcal{H}(\omega) = -R_{0,2}(\omega) = \Big(i\omega - R^{\psi_1,\zeta_1}(\omega)\Big)^{-1}$. From this,

$$\Big(1 + \mathcal{H}(\omega)\Big)\boldsymbol{C}^{g^{\ell+1}\psi_1}(\omega) = \langle u_{\zeta_1}(\omega)\boldsymbol{g}^{\ell+1}\rangle + \boldsymbol{C}_g^{\ell+1}\Big(\eta\gamma_0\boldsymbol{C}^{h^\ell,\psi_0}(\omega) + \boldsymbol{R}^{\psi_0,r^\ell}(\omega)\Big) \quad (235)$$

$$\Big(1 + \mathcal{H}(\omega)\Big)\boldsymbol{R}^{\psi_1 u^{\ell+1}}(\omega) = (\boldsymbol{R}_{gu}^{\ell+1})^\top\Big(\eta\gamma_0\boldsymbol{C}^{h^\ell,\psi_0}(\omega) + \boldsymbol{R}^{\psi_0,r^\ell}(\omega)\Big). \quad (236)$$

By collecting all the four projected quantities as for the first layer and setting $z = i\omega$, then

$$\boldsymbol{q}^\ell(z) = \begin{pmatrix} \boldsymbol{C}^{h^\ell\psi_0}(z) \\ \boldsymbol{R}^{\psi_0 r^\ell}(z) \\ \boldsymbol{C}^{g^{\ell+1}\psi_1}(z) \\ \boldsymbol{R}^{\psi_1 u^{\ell+1}}(z) \end{pmatrix} \quad (237)$$

by giving the poles condition of

$$\det\boldsymbol{A}^\ell(z) = 0 \quad (238)$$

with

$$\boldsymbol{A}^\ell(z) = \begin{pmatrix} (z - R^{\psi_1,\zeta_1}(z))\boldsymbol{I} & \boldsymbol{0} & -\boldsymbol{R}_{hr}^\ell - \eta\gamma_0\boldsymbol{C}_h^\ell & -\boldsymbol{C}_h^\ell \\ \boldsymbol{0} & (z - R^{\psi_1,\zeta_1}(z))\boldsymbol{I} & -\eta\gamma_0(\boldsymbol{R}_{hr}^\ell)^\top & -(\boldsymbol{R}_{hr}^\ell)^\top \\ -\boldsymbol{R}_{gu}^{\ell+1} - \eta\gamma_0\boldsymbol{C}_g^{\ell+1} & -\boldsymbol{C}_g^{\ell+1} & (R^{\psi_1,\zeta_1}(z))^{-1}\boldsymbol{I} & \boldsymbol{0} \\ -\eta\gamma_0(\boldsymbol{R}_{gu}^{\ell+1})^\top & -(\boldsymbol{R}_{gu}^{\ell+1})^\top & \boldsymbol{0} & (R^{\psi_1,\zeta_1}(z))^{-1}\boldsymbol{I} \end{pmatrix}. \quad (239)$$

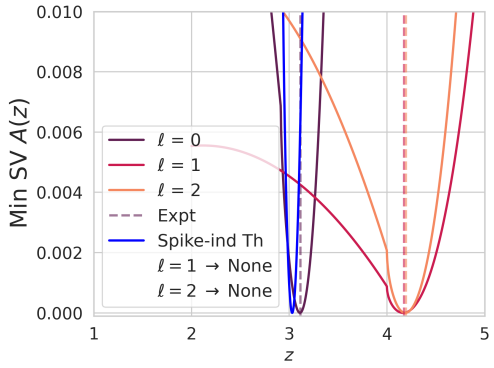


Figure 14: Effect of spike–initialization correlations in a three-hidden-layer linear network. Colored curves show the minimum singular value of the full DMFT outlier matrix for each layer; their zeros agree with the empirical outliers (dashed vertical lines). The independent-spike approximation (blue) underestimates the outlier for $\ell = 0$ and predicts no outliers for $\ell = 1, 2$, demonstrating the importance of retaining these correlations.

E.12.3 Comparison with an independent-spike approximation

To assess the role of spike–initialization correlations, we compare the full theory with an independent-spike approximation. We retain the covariance matrices $\boldsymbol{C}_h^\ell, \boldsymbol{C}_g^{\ell+1}$ obtained from the full training DMFT, but set

$$\boldsymbol{R}_{hr}^\ell = 0, \qquad \boldsymbol{R}_{gu}^{\ell+1} = 0 \tag{240}$$

only in the spectral-probe equations. This corresponds to replacing $\boldsymbol{W}^\ell(0)$ by an independent Gaussian matrix while preserving the learned finite-rank update and its temporal correlations.

The mixed Stein terms then vanish, and the projected responses satisfy

$$\boldsymbol{R}^{\psi_0 r^\ell} = 0, \qquad \boldsymbol{R}^{\psi_1 u^{\ell+1}} = 0. \tag{241}$$

Eliminating these sectors reduces the nontrivial outlier condition to

$$\det\left[a_\ell(z) b_\ell(z) \boldsymbol{I}_T - (\eta\gamma_0)^2 \boldsymbol{C}_g^{\ell+1} \boldsymbol{C}_h^\ell\right] = 0, \tag{242}$$

where

$$a_\ell(z) = z - R^{\psi_1 \zeta_1}(z), \qquad b_\ell(z) = \left[R^{\psi_1 \zeta_1}(z)\right]^{-1}. \tag{243}$$

The bulk responses remain unchanged, and only roots outside the MP support on the physical resolvent branch represent isolated outliers.

Figure 14 compares this approximation with the full theory for a three-hidden-layer network with $\alpha = 4$, $\gamma_0 = 8$, and $\nu = 2$. In this example, the independent-spike approximation underestimates the outlier for $\ell = 0$ and misses the outliers observed for $\ell = 1, 2$.

E.13 Multi-Class setting

The extension to multiple output classes is straightforward when the number of classes C remains $\mathcal{O}(1)$ as

$$D, N, P \to \infty, \qquad \alpha = \frac{P}{D}, \qquad \nu = \frac{N}{D}.$$

The important point is that the class index is not averaged away in the proportional limit. Instead, every training-time order parameter becomes a block kernel indexed by both time and class. Thus the scalar-output DMFT is lifted to a class-time DMFT by replacing

$$t \quad \longrightarrow \quad (c, t), \qquad c \in \{1, \dots, C\}.$$

Equivalently, every $T \times T$ correlation or response kernel becomes an $(CT) \times (CT)$ block kernel.

For each class $c \in \{1, \dots, C\}$, introduce a teacher vector $\boldsymbol{w}_{\star,c}$, a readout vector $\boldsymbol{w}_c^L$, an error vector $\boldsymbol{\Delta}_c(t)$, and class-resolved training fields

$$\boldsymbol{v}_c(t), \qquad \boldsymbol{h}_c^\ell(t), \qquad \boldsymbol{g}_c^\ell(t).$$

The hidden weights are shared across classes, while the readout vectors are class-specific. Keeping the same normalization as in the scalar-output case, the finite-rank weight updates become

$$\boldsymbol{W}^0(t) = \boldsymbol{W}^0(0) + \frac{\eta\gamma_0}{\sqrt{D}} \sum_{s<t} \sum_{c=1}^{C} \boldsymbol{g}_c^1(s) \boldsymbol{h}_c^0(s)^\top, \tag{244}$$

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\eta\gamma_0}{\sqrt{N}} \sum_{s<t} \sum_{c=1}^{C} \boldsymbol{g}_c^{\ell+1}(s) \boldsymbol{h}_c^\ell(s)^\top, \qquad 1 \le \ell \le L-1, \tag{245}$$

$$\boldsymbol{w}_c^L(t) = \boldsymbol{w}_c^L(0) + \eta\gamma_0 \sum_{s<t} \boldsymbol{h}_c^L(s). \tag{246}$$

The initialized-weight fields are now class-dependent as well

$$\boldsymbol{\xi}_c^0(t) = \frac{\sqrt{D}}{N\gamma_0} \boldsymbol{W}^0(0)^\top \boldsymbol{g}_c^1(t), \qquad \boldsymbol{\chi}_c^1(t) = \frac{1}{\sqrt{D}} \boldsymbol{W}^0(0) \boldsymbol{h}_c^0(t), \tag{247}$$

$$\boldsymbol{\xi}_c^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_c^{\ell+1}(t), \quad \boldsymbol{\chi}_c^{\ell+1}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \boldsymbol{h}_c^\ell(t), \qquad 1 \le \ell \le L-1. \tag{248}$$

Expanding the finite-rank updates gives

$$\boldsymbol{v}_c(t) = \boldsymbol{w}_{\star,c} - \boldsymbol{\xi}_c^0(t) - \eta \sum_{s<t} \sum_{c'=1}^{C} C_{g,cc'}^1(t,s) \boldsymbol{h}_{c'}^0(s), \tag{249}$$

$$\boldsymbol{h}_c^{\ell+1}(t) = \boldsymbol{\chi}_c^{\ell+1}(t) + \eta\gamma_0 \sum_{s<t} \sum_{c'=1}^{C} C_{h,cc'}^\ell(t,s) \boldsymbol{g}_{c'}^{\ell+1}(s), \tag{250}$$

$$\boldsymbol{g}_c^\ell(t) = \boldsymbol{\xi}_c^\ell(t) + \eta\gamma_0 \sum_{s<t} \sum_{c'=1}^{C} C_{g,cc'}^{\ell+1}(t,s) \boldsymbol{h}_{c'}^\ell(s). \tag{251}$$

The only structural change in the full-batch DMFT is that every correlation and response function acquires two class indices. The class-resolved correlations are

$$C_{v,cc'}(t,t') = \frac{1}{D} \boldsymbol{v}_c(t) \cdot \boldsymbol{v}_{c'}(t'), \qquad C_{\Delta,cc'}(t,t') = \frac{1}{P} \boldsymbol{\Delta}_c(t) \cdot \boldsymbol{\Delta}_{c'}(t'), \tag{252}$$

$$C_{h,cc'}^0(t,t') = \frac{1}{D} \boldsymbol{h}_c^0(t) \cdot \boldsymbol{h}_{c'}^0(t'), \qquad C_{h,cc'}^\ell(t,t') = \frac{1}{N} \boldsymbol{h}_c^\ell(t) \cdot \boldsymbol{h}_{c'}^\ell(t'), \tag{253}$$

$$C_{g,cc'}^\ell(t,t') = \frac{1}{N} \boldsymbol{g}_c^\ell(t) \cdot \boldsymbol{g}_{c'}^\ell(t'). \tag{254}$$

The corresponding response functions are

$$R_{\Delta,cc'}(t,t') = -\frac{i}{P} \boldsymbol{\Delta}_c(t) \cdot \widehat{\boldsymbol{\Delta}}_{c'}(t'), \qquad R_{vu,cc'}^0(t,t') = -\frac{i}{D} \boldsymbol{v}_c(t) \cdot \widehat{\boldsymbol{h}}_{c'}^0(t'), \tag{255}$$

$$R_{hr,cc'}^0(t,t') = -\frac{i}{D} \boldsymbol{h}_c^0(t) \cdot \widehat{\boldsymbol{\xi}}_{c'}^0(t'), \qquad R_{gu,cc'}^1(t,t') = -\frac{i}{N} \boldsymbol{g}_c^1(t) \cdot \widehat{\boldsymbol{\chi}}_{c'}^1(t'), \tag{256}$$

$$R_{hr,cc'}^\ell(t,t') = -\frac{i}{N} \boldsymbol{h}_c^\ell(t) \cdot \widehat{\boldsymbol{\xi}}_{c'}^\ell(t'), \qquad R_{gu,cc'}^{\ell+1}(t,t') = -\frac{i}{N} \boldsymbol{g}_c^{\ell+1}(t) \cdot \widehat{\boldsymbol{\chi}}_{c'}^{\ell+1}(t'). \tag{257}$$

Equivalently, in the single-site description, the scalar Gaussian noises become C-dimensional Gaussian processes over class index. For instance

$$\langle u_{\Delta,c}(t) u_{\Delta,c'}(t') \rangle = C_{v,cc'}(t,t'), \qquad \langle u_c^0(t) u_{c'}^0(t') \rangle = \frac{1}{\alpha} C_{\Delta,cc'}(t,t'), \tag{258}$$

$$\langle u_c^\ell(t) u_{c'}^\ell(t') \rangle = C_{h,cc'}^{\ell-1}(t,t'), \qquad \langle r_c^\ell(t) r_{c'}^\ell(t') \rangle = C_{g,cc'}^{\ell+1}(t,t'). \tag{259}$$

The first-layer backward noise has covariance

$$\left\langle r_c^0(t) r_{c'}^0(t') \right\rangle = \frac{1}{\nu \gamma_0^2} C^1_{g,cc'}(t,t').$$

The readout boundary becomes

$$g_c^L(t) = r_c^L + \eta\gamma_0 \sum_{s<t} h_c^L(s), \qquad \left\langle r_c^L r_{c'}^L \right\rangle = \delta_{cc'}.$$

Thus the training DMFT equations are the scalar-output equations with the replacement $t \mapsto (c,t)$, together with sums over repeated class indices. Explicitly

$$\Delta_c(t) = u_{\Delta,c}(t) + \frac{1}{\alpha} \sum_{s<t} \sum_{c'=1}^{C} R^0_{vu,cc'}(t,s) \Delta_{c'}(s) + \sigma \epsilon_c(t), \tag{260}$$

$$h_c^0(t) = u_c^0(t) + \sum_{s<t} \sum_{c'=1}^{C} R_{\Delta,cc'}(t,s) v_{c'}(s), \tag{261}$$

$$v_c(t) = w_{\star,c} - r_c^0(t) - \sum_{s<t} \sum_{c'=1}^{C} \left[\frac{1}{\gamma_0} R^1_{gu,cc'}(t,s) + \eta C^1_{g,cc'}(t,s) \right] h^0_{c'}(s), \tag{262}$$

$$h_c^1(t) = u_c^1(t) + \sum_{s<t} \sum_{c'=1}^{C} \left[\frac{1}{\nu\gamma_0} R^0_{hr,cc'}(t,s) + \eta\gamma_0 C^0_{h,cc'}(t,s) \right] g^1_{c'}(s), \tag{263}$$

$$h_c^\ell(t) = u_c^\ell(t) + \sum_{s<t} \sum_{c'=1}^{C} \left[R^{\ell-1}_{hr,cc'}(t,s) + \eta\gamma_0 C^{\ell-1}_{h,cc'}(t,s) \right] g^\ell_{c'}(s), \qquad 2 \le \ell \le L, \tag{264}$$

$$g_c^\ell(t) = r_c^\ell(t) + \sum_{s<t} \sum_{c'=1}^{C} \left[R^{\ell+1}_{gu,cc'}(t,s) + \eta\gamma_0 C^{\ell+1}_{g,cc'}(t,s) \right] h^\ell_{c'}(s), \qquad 1 \le \ell \le L-1. \tag{265}$$

The responses are again functional derivatives of the single-site process

$$R_{\Delta,cc'}(t,t') = \left\langle \frac{\partial \Delta_c(t)}{\partial u_{\Delta,c'}(t')} \right\rangle, \qquad R^0_{vu,cc'}(t,t') = \left\langle \frac{\partial v_c(t)}{\partial u^0_{c'}(t')} \right\rangle,$$

$$R^\ell_{hr,cc'}(t,t') = \left\langle \frac{\partial h^\ell_c(t)}{\partial r^\ell_{c'}(t')} \right\rangle, \qquad R^\ell_{gu,cc'}(t,t') = \left\langle \frac{\partial g^\ell_c(t)}{\partial u^\ell_{c'}(t')} \right\rangle.$$

The deterministic self-consistency conditions are

$$C_{v,cc'}(t,t') = \left\langle v_c(t) v_{c'}(t') \right\rangle, \qquad C_{\Delta,cc'}(t,t') = \left\langle \Delta_c(t) \Delta_{c'}(t') \right\rangle, \tag{266}$$

$$C^\ell_{h,cc'}(t,t') = \left\langle h^\ell_c(t) h^\ell_{c'}(t') \right\rangle, \qquad C^\ell_{g,cc'}(t,t') = \left\langle g^\ell_c(t) g^\ell_{c'}(t') \right\rangle. \tag{267}$$

The spectral calculation changes in the same way. The probe vector itself has no class index because it probes the shared weight matrix. However, the finite-rank update directions are now the collection

$$\left\{ \boldsymbol{h}^\ell_c(s), \boldsymbol{g}^{\ell+1}_c(s) \right\}_{c=1}^{C,\, s<t}.$$

For example, for a hidden layer ℓ

$$\boldsymbol{\psi}_1(\tau) = \boldsymbol{\zeta}_1(\tau) + \eta\gamma_0 \sum_{s<t} \sum_{c=1}^{C} \boldsymbol{g}^{\ell+1}_c(s) C^{h^\ell_c \psi_0}(s,\tau), \tag{268}$$

$$\boldsymbol{\psi}_2(\tau) = \boldsymbol{\zeta}_2(\tau) + \eta\gamma_0 \sum_{s<t} \sum_{c=1}^{C} \boldsymbol{h}^\ell_c(s) C^{g^{\ell+1}_c \psi_1}(s,\tau). \tag{269}$$

Therefore the pole matrix derived above is unchanged in structure, but its blocks are enlarged from $T \times T$ matrices to $(CT) \times (CT)$ block matrices. If

$$I = (c,t), \qquad J = (c',t'),$$

we define

$$[\mathsf{C}_h^\ell]_{I,J} = C_{h,cc'}^\ell(t,t'), \qquad [\mathsf{C}_g^{\ell+1}]_{I,J} = C_{g,cc'}^{\ell+1}(t,t'),$$

and similarly for

$$\mathsf{R}_{hr}^\ell, \qquad \mathsf{R}_{gu}^{\ell+1}.$$

Then the hidden-layer spectral condition is obtained by the substitutions

$$\boldsymbol{C}_h^\ell \to \mathsf{C}_h^\ell, \qquad \boldsymbol{C}_g^{\ell+1} \to \mathsf{C}_g^{\ell+1}, \qquad \boldsymbol{R}_{hr}^\ell \to \mathsf{R}_{hr}^\ell, \qquad \boldsymbol{R}_{gu}^{\ell+1} \to \mathsf{R}_{gu}^{\ell+1}, \qquad \boldsymbol{I}_T \to \boldsymbol{I}_{CT}.$$

Thus

$$\det \boldsymbol{\mathcal{A}}_\ell^{\mathrm{multi-class}}(z) = 0 \tag{270}$$

has exactly the same form as in the scalar-output case.

If the teacher/readout initialization is class-symmetric and the targets are orthogonal or independent across classes, the saddle can often be restricted to a class-diagonal or permutation-symmetric ansatz. In the most general case, however, one should keep all cross-class order parameters.

E.14 Online Dynamics

The full-batch calculation above assumes that the same dataset $\boldsymbol{X} \in \mathbb{R}^{P\times D}$ is reused at every training step. In online SGD, this data block is replaced by a fresh mini-batch $\boldsymbol{X}(t) \in \mathbb{R}^{B\times D}$ at each step t, with

$$B = \alpha_B D, \qquad D, N, B \to \infty, \qquad \nu = \frac{N}{D}, \qquad \alpha_B = \frac{B}{D}.$$

The input-layer data fields are now

$$\boldsymbol{h}^0(t) = \frac{\sqrt{D}}{B}\boldsymbol{X}(t)^\top \boldsymbol{\Delta}(t), \qquad \boldsymbol{\Delta}(t) = \frac{1}{\sqrt{D}}\boldsymbol{X}(t)\boldsymbol{v}(t) + \sigma\boldsymbol{\epsilon}(t), \tag{271}$$

where $\boldsymbol{X}(t)$ and $\boldsymbol{\epsilon}(t)$ are resampled independently at each training step. All weight-dependent fields and all finite-rank weight updates are unchanged.

The only modification to the DMFT is therefore in the data average. Since the data are fresh at each time step, the average over $\boldsymbol{X}(t)$ factorizes over t. Thus the nonlocal data-response sector of the full-batch theory is removed. Equivalently, in the full-batch data block one makes the replacement

$$R_\Delta(t,t') \longrightarrow \delta_{tt'}, \tag{272}$$

and the error covariance becomes

$$C_\Delta(t,t') = C_v(t,t') + \sigma^2\delta_{tt'}. \tag{273}$$

The stochastic input-layer process is therefore

$$\Delta(t) = u_\Delta(t) + \sigma\epsilon(t), \qquad u_\Delta \sim \mathrm{GP}(0, C_v), \qquad \langle\epsilon(t)\epsilon(t')\rangle = \delta_{tt'}, \tag{274}$$

and

$$h^0(t) = v(t) + u^0(t), \qquad u^0 \sim \mathrm{GP}\left(0, \frac{1}{\alpha_B}C_\Delta(t,t')\delta_{tt'}\right). \tag{275}$$

In particular, the full-batch memory term

$$\frac{1}{\alpha}\sum_{t'<t} R_{vu}^0(t,t')\Delta(t')$$

in the Δ-equation is absent in online SGD. The finite-batch effect instead appears as a time-local stochastic-gradient noise in h^0. On the diagonal,

$$\langle h^0(t)^2\rangle = C_v(t,t) + \frac{1}{\alpha_B}C_\Delta(t,t), \tag{276}$$

so smaller batches increase the effective input noise.

The remaining single-site training equations are unchanged, except that they are driven by the online-SGD input process above. Thus

$$v(t) = w_\star - r^0(t) - \sum_{s<t}\left[\frac{1}{\gamma_0}R_{gu}^1(t,s) + \eta C_g^1(t,s)\right]h^0(s), \qquad r^0 \sim \mathrm{GP}\left(0, \frac{1}{\nu\gamma_0^2}C_g^1\right), \tag{277}$$

$$h^1(t) = u^1(t) + \sum_{s<t} \left[\frac{1}{\nu\gamma_0} R^0_{hr}(t,s) + \eta\gamma_0 C^0_h(t,s) \right] g^1(s), \qquad u^1 \sim \mathrm{GP}(0, C^0_h), \tag{278}$$

and, for $1 < \ell \leq L$,

$$h^\ell(t) = u^\ell(t) + \sum_{s<t} \left[R^{\ell-1}_{hr}(t,s) + \eta\gamma_0 C^{\ell-1}_h(t,s) \right] g^\ell(s), \qquad u^\ell \sim \mathrm{GP}(0, C^{\ell-1}_h). \tag{279}$$

Similarly, for $1 \leq \ell \leq L-1$,

$$g^\ell(t) = r^\ell(t) + \sum_{s<t} \left[R^{\ell+1}_{gu}(t,s) + \eta\gamma_0 C^{\ell+1}_g(t,s) \right] h^\ell(s), \qquad r^\ell \sim \mathrm{GP}(0, C^{\ell+1}_g), \tag{280}$$

with the same readout boundary condition

$$g^L(t) = r^L + \eta\gamma_0 \sum_{s<t} h^L(s), \qquad r^L \sim \mathcal{N}(0,1). \tag{281}$$

The self-consistency equations for the correlations and responses are also unchanged in form:

$$C_v(t,t') = \langle v(t)v(t') \rangle, \qquad C_\Delta(t,t') = \langle \Delta(t)\Delta(t') \rangle, \tag{282}$$

$$C^\ell_h(t,t') = \langle h^\ell(t)h^\ell(t') \rangle, \qquad C^\ell_g(t,t') = \langle g^\ell(t)g^\ell(t') \rangle, \tag{283}$$

and

$$R^0_{vu}(t,t') = \left\langle \frac{\partial v(t)}{\partial u^0(t')} \right\rangle, \qquad R^\ell_{hr}(t,t') = \left\langle \frac{\partial h^\ell(t)}{\partial r^\ell(t')} \right\rangle, \tag{284}$$

$$R^\ell_{gu}(t,t') = \left\langle \frac{\partial g^\ell(t)}{\partial u^\ell(t')} \right\rangle. \tag{285}$$

Finally, the spectral-probe DMFT derived above is not modified at the level of the weight averages. The probe equations and the pole conditions

$$\det \boldsymbol{\mathcal{A}}_0(z) = 0, \qquad \det \boldsymbol{\mathcal{A}}_\ell(z) = 0$$

retain the same form. The only change is that the training order parameters

$$C^\ell_h, \quad C^\ell_g, \quad R^\ell_{hr}, \quad R^\ell_{gu}$$

that enter the probe fixed-point equations must now be computed from the online-SGD training DMFT rather than from the full-batch DMFT.

E.15 NTK SPECTRUM

For a two-layer linear network, i.e. $L = 1$, the model is

$$f_t(\boldsymbol{x}) = \frac{1}{N\gamma_0} \boldsymbol{w}(t)^\top \boldsymbol{W}(t) \boldsymbol{x}, \tag{286}$$

where $\boldsymbol{W}(t) \equiv \boldsymbol{W}^0(t) \in \mathbb{R}^{N \times D}$ and $\boldsymbol{w}(t) \equiv \boldsymbol{w}^1(t) \in \mathbb{R}^N$. In this case the backward field is simply

$$\boldsymbol{g}^1(t) = \boldsymbol{w}(t). \tag{287}$$

The population NTK can be written as a kernel on input space. Indeed,

$$\frac{\partial f_t(\boldsymbol{x})}{\partial \boldsymbol{w}} = \frac{1}{N\gamma_0} \boldsymbol{W}(t) \boldsymbol{x}, \qquad \frac{\partial f_t(\boldsymbol{x})}{\partial W_{ij}} = \frac{1}{N\gamma_0} w_i(t) x_j. \tag{288}$$

Therefore, up to the overall convention-dependent prefactor coming from the learning-rate normalization, the NTK has the

$$\Theta_t(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{D} \boldsymbol{x}^\top \boldsymbol{B}_t \boldsymbol{x}', \tag{289}$$

where the $D \times D$ population NTK core is

$$\boldsymbol{B}_t = \frac{1}{N} \boldsymbol{W}(t)^\top \boldsymbol{W}(t) + q_g(t) \boldsymbol{I}_D \tag{290}$$

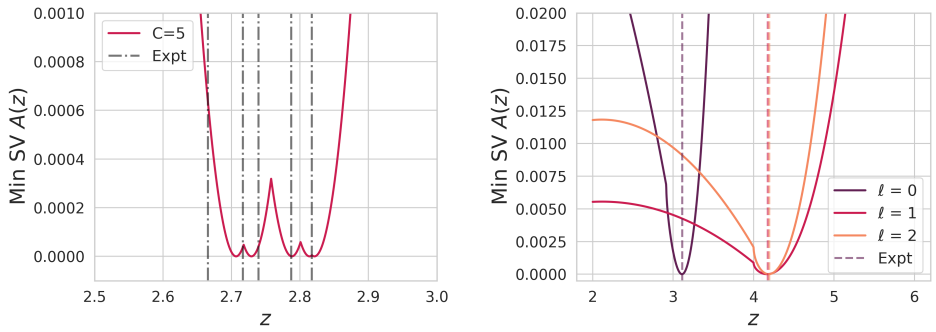


Figure 15: Root-finding procedure for $\det \boldsymbol{A}(z) = 0$ in a multiclass model ($C = 5$). We compute the minimum singular value of $\boldsymbol{A}(z)$ as a function of z; values where it approaches zero identify predicted spectral outliers. The presence of multiple minima reflects the rank-C structure of the problem. Dashed lines indicate empirical outlier eigenvalues of $\boldsymbol{W}^\top \boldsymbol{W}/N$.

with

$$q_g(t) = \frac{1}{N}\|\boldsymbol{w}(t)\|^2 = C_g^1(t,t) \tag{291}$$

in the usual notation. Thus, in the two-layer case, the NTK core is exactly the first-layer weight covariance plus an isotropic readout-gradient contribution. The scalar $q_g(t)$ is already one of the training DMFT order parameters, since $C_g^1(t,t) = N^{-1}\boldsymbol{g}^1(t) \cdot \boldsymbol{g}^1(t)$.

Consequently, the spectrum of the population NTK core is just a shift of the spectrum of the first-layer weight covariance

$$\boldsymbol{M}^0(t) = \frac{1}{N}\boldsymbol{W}(t)^\top \boldsymbol{W}(t). \tag{292}$$

If

$$\lambda_1^M(t), \ldots, \lambda_D^M(t) \tag{293}$$

are the eigenvalues of $\boldsymbol{M}^0(t)$, then the eigenvalues of $\boldsymbol{B}_t$ are

$$\lambda_a^{\mathrm{NTK}}(t) = \lambda_a^M(t) + q_g(t), \qquad a = 1, \ldots, D. \tag{294}$$

Equivalently, if $m_M(z;t)$ is the Stieltjes transform of the spectral measure of $\boldsymbol{M}^0(t)$, then the NTK-core resolvent is

$$m_{\mathrm{NTK}}(z;t) = m_M(z - q_g(t);t). \tag{295}$$

Thus the NTK spectral density satisfies

$$\rho_{\mathrm{NTK}}(\lambda;t) = \rho_M(\lambda - q_g(t);t). \tag{296}$$

Notice that computing the spectrum of $\boldsymbol{K}_t$ would require in principle an additional probe involving the data matrix $\boldsymbol{X}$. The simple shift relation above applies to the population limit only.

E.15.1 3-LAYER EXTENSION

In the three layer setting, the population NTK is

$$\boldsymbol{B} = q_1 \boldsymbol{I}_D + \frac{1}{N}(\boldsymbol{W}^0)^\top \left[q_2 \boldsymbol{I}_N + \frac{1}{N}(\boldsymbol{W}^1)^\top \boldsymbol{W}^1 \right] \boldsymbol{W}^0 \tag{297}$$

suppressing t on the weights, and where $\boldsymbol{W}^0 \in \mathbb{R}^{N \times D}$ and $\boldsymbol{W}^1 \in \mathbb{R}^{N \times N}$. As usual, let's introduce the spectral dynamics

$$\partial_\tau \boldsymbol{\psi}_0 = \boldsymbol{B}\boldsymbol{\psi}_0 + \boldsymbol{j} \tag{298}$$

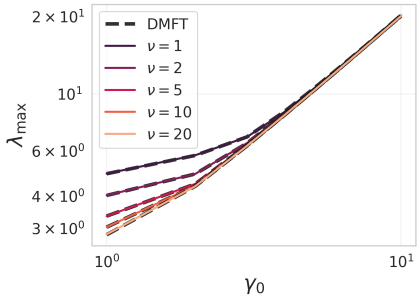


Figure 16: Leading NTK eigenvalue $\lambda_{\max}$ as a function of richness γ_0 for different aspect rations ν. As γ_0 increases, $\lambda_{\max}$ rapidly grows and the curves collapse, indicating fast convergence to a width-stable sharpening regime.

where τ is the auxiliary probe time. We define the following fields

$$\begin{aligned}
\boldsymbol{\psi}_1 &= \frac{1}{\sqrt{D}} \boldsymbol{W}^0 \boldsymbol{\psi}_0 \\
\boldsymbol{\psi}_2 &= \frac{1}{\sqrt{N}} \boldsymbol{W}^1 \boldsymbol{\psi}_1 \\
\boldsymbol{\psi}_3 &= q_2 \boldsymbol{\psi}_1 + \frac{1}{\sqrt{N}} (\boldsymbol{W}^1)^\top \boldsymbol{\psi}_2 \\
\boldsymbol{\psi}_4 &= \frac{\sqrt{D}}{N} (\boldsymbol{W}^0)^\top \boldsymbol{\psi}_3
\end{aligned} \tag{299}$$

which imply

$$\boldsymbol{\psi}_4 = (\boldsymbol{B} - q_1 \boldsymbol{I}_D) \boldsymbol{\psi}_0, \qquad \partial_\tau \boldsymbol{\psi}_0 = q_1 \boldsymbol{\psi}_0 + \boldsymbol{\psi}_4 + \boldsymbol{j}. \tag{300}$$

Here $\boldsymbol{\psi}_0, \boldsymbol{\psi}_4$ live in input space, $\boldsymbol{\psi}_1, \boldsymbol{\psi}_3$ in hidden layer 1, and $\boldsymbol{\psi}_2$ in hidden layer 2. The first-layer normalization is the same as in the two-layer probe.

By defining

$$\kappa = \eta \gamma_0, \qquad \chi = \frac{1}{\gamma_0 \nu}. \tag{301}$$

the finite-rank weight updates can be rewritten as

$$\begin{aligned}
\boldsymbol{W}^0(t) &= \boldsymbol{W}^0(0) + \frac{\kappa}{\sqrt{D}} \sum_{s<t} \boldsymbol{g}^1(s) \boldsymbol{h}^0(s)^\top \\
\boldsymbol{W}^1(t) &= \boldsymbol{W}^1(0) + \frac{\kappa}{\sqrt{N}} \sum_{s<t} \boldsymbol{g}^2(s) \boldsymbol{h}^1(s)^\top.
\end{aligned} \tag{302}$$

The updates remain statistically dependent on initialization. This dependence is retained in the mixed response sectors below. Substituting Eq. 302 into Eq. 299 gives

$$\begin{aligned}
\boldsymbol{\psi}_1 &= \boldsymbol{\zeta}_1 + \kappa \sum_{s<t} \boldsymbol{g}^1(s) C^{h^0 \psi_0}(s, \tau) \\
\boldsymbol{\psi}_2 &= \boldsymbol{\zeta}_2 + \kappa \sum_{s<t} \boldsymbol{g}^2(s) C^{h^1 \psi_1}(s, \tau) \\
\boldsymbol{\psi}_3 &= q_2 \boldsymbol{\psi}_1 + \boldsymbol{\zeta}_3 + \kappa \sum_{s<t} \boldsymbol{h}^1(s) C^{g^2 \psi_2}(s, \tau) \\
\boldsymbol{\psi}_4 &= \boldsymbol{\zeta}_4 + \kappa \sum_{s<t} \boldsymbol{h}^0(s) C^{g^1 \psi_3}(s, \tau)
\end{aligned} \tag{303}$$

where the initialized-weight fields are

$$\begin{aligned}
\boldsymbol{\zeta}_1 &= \frac{1}{\sqrt{D}} \boldsymbol{W}^0(0) \boldsymbol{\psi}_0, \quad \boldsymbol{\zeta}_4 = \frac{\sqrt{D}}{N} \boldsymbol{W}^0(0)^\top \boldsymbol{\psi}_3, \\
\boldsymbol{\zeta}_2 &= \frac{1}{\sqrt{N}} \boldsymbol{W}^1(0) \boldsymbol{\psi}_1, \quad \boldsymbol{\zeta}_3 = \frac{1}{\sqrt{N}} \boldsymbol{W}^1(0)^\top \boldsymbol{\psi}_2.
\end{aligned} \tag{304}$$

After taking a Fourier transform in τ, we stack the history $s = 0, \dots, t-1$ into the following t-dimensional vectors

$$\begin{aligned} \boldsymbol{c}_0 &: \quad C^{h^0\psi_0}(s,z) & \boldsymbol{\rho}_0 &: \quad R^{\psi_0 r^0}(z,s) \\ \boldsymbol{c}_1 &: \quad C^{g^1\psi_3}(s,z) & \boldsymbol{\rho}_1 &: \quad R^{\psi_3 u^1}(z,s) \\ \boldsymbol{c}_2 &: \quad C^{h^1\psi_1}(s,z) & \boldsymbol{\rho}_2 &: \quad R^{\psi_1 r^1}(z,s) \\ \boldsymbol{c}_3 &: \quad C^{g^2\psi_2}(s,z) & \boldsymbol{\rho}_3 &: \quad R^{\psi_2 u^2}(z,s). \end{aligned} \tag{305}$$

For convenience, we also define

$$\begin{aligned} \boldsymbol{b}_0 &= \kappa \boldsymbol{c}_0 + \chi \boldsymbol{\rho}_0, \quad & \boldsymbol{b}_1 &= \kappa \boldsymbol{c}_1 + \boldsymbol{\rho}_1, \\ \boldsymbol{b}_2 &= \kappa \boldsymbol{c}_2 + \boldsymbol{\rho}_2, \quad & \boldsymbol{b}_3 &= \kappa \boldsymbol{c}_3 + \boldsymbol{\rho}_3. \end{aligned} \tag{306}$$

These combinations collect the explicit update and the response contribution generated by averaging over initialization. Here, there are two independent disorder averages, over $\boldsymbol{W}^0(0)$ and $\boldsymbol{W}^1(0)$. In the generating functional each has the usual form

$$\left\langle e^{-i \operatorname{Tr}(\boldsymbol{W}^\top \boldsymbol{J})} \right\rangle_{\boldsymbol{W}} = e^{-\frac{1}{2} \operatorname{Tr}(\boldsymbol{J}^\top \boldsymbol{J})}. \tag{307}$$

The training–probe cross terms generate mixed responses and mixed noise covariances as in the previous derivations. The probe–probe terms generate the pure probe responses. After the Hubbard–Stratonovich step, the scalar spectral-domain equations are

$$\begin{aligned} \psi_1 &= u_{\zeta_1} + \frac{1}{\nu} R^{\psi_0 \zeta_4} \psi_3 + (\boldsymbol{g}^1)^\top \boldsymbol{b}_0, \\ \psi_2 &= u_{\zeta_2} + R^{\psi_1 \zeta_3} \psi_2 + (\boldsymbol{g}^2)^\top \boldsymbol{b}_2, \\ \psi_3 &= (q_2 + R^{\psi_2 \zeta_2}) \psi_1 + u_{\zeta_3} + (\boldsymbol{h}^1)^\top \boldsymbol{b}_3, \\ (z - q_1)\psi_0 &= R^{\psi_3 \zeta_1} \psi_0 + u_{\zeta_4} + (\boldsymbol{h}^0)^\top \boldsymbol{b}_1. \end{aligned} \tag{308}$$

Here $\boldsymbol{h}^\ell, \boldsymbol{g}^\ell$ are scalar single-site training histories.

The mixed Gaussian-noise covariances needed for the closure are

$$\begin{aligned} \langle u_{\zeta_1}(z) u^1(s) \rangle &= c_0(s,z), \quad & \langle u_{\zeta_4}(z) r^0(s) \rangle &= \chi c_1(s,z), \\ \langle u_{\zeta_2}(z) u^2(s) \rangle &= c_2(s,z), \quad & \langle u_{\zeta_3}(z) r^1(s) \rangle &= c_3(s,z). \end{aligned} \tag{309}$$

Before transforming, the pure noise covariances are instead

$$\begin{aligned} \langle u_{\zeta_1}(\tau) u_{\zeta_1}(\tau') \rangle &= C^{\psi_0}(\tau,\tau'), \quad & \langle u_{\zeta_4}(\tau) u_{\zeta_4}(\tau') \rangle &= \nu^{-1} C^{\psi_3}(\tau,\tau'), \\ \langle u_{\zeta_2}(\tau) u_{\zeta_2}(\tau') \rangle &= C^{\psi_1}(\tau,\tau'), \quad & \langle u_{\zeta_3}(\tau) u_{\zeta_3}(\tau') \rangle &= C^{\psi_2}(\tau,\tau'). \end{aligned} \tag{310}$$

Bulk spectrum At fixed training time t, replacing the trained weights by their initializations changes $\boldsymbol{B}$ by rank at most $4t$ and therefore preserves its limiting spectral density. The training dependence of the bulk enters through $q_1 = C_g^1(t,t)$ and $q_2 = C_g^2(t,t)$.

Define

$$\mathcal{H}(z) = \lim_{D \to \infty} \frac{1}{D} \operatorname{Tr}(z \boldsymbol{I}_D - \boldsymbol{B})^{-1}. \tag{311}$$

The pure probe responses satisfy

$$\mathcal{H}^{-1} = z - q_1 - bP, \qquad P^{-1} = 1 - ab, \qquad S^{-1} = 1 - aP, \qquad a = \frac{\mathcal{H}}{\nu}, \quad b = q_2 + S. \tag{312}$$

Eliminating a, b, S gives

$$P = 1 + \frac{(z - q_1)\mathcal{H} - 1}{\nu}, \qquad P\left(1 - \frac{q_2 \mathcal{H}}{\nu}\right)\left(1 - \frac{\mathcal{H} P}{\nu}\right) = 1. \tag{313}$$

Substitution of the first relation into the second yields a quartic equation for $\mathcal{H}$. Its physical solution is the branch analytic in the upper half-plane with $\mathcal{H}(z) \sim z^{-1}$ for $z \to \infty$ and $\operatorname{Im} \mathcal{H}(z) < 0$.

The limiting bulk density is therefore

$$\rho_{\mathrm{NTK}}^{\mathrm{bulk}}(\lambda) = (1 - \nu)_+ \delta(\lambda - q_1) + \rho_{\mathrm{cont}}(\lambda), \qquad \rho_{\mathrm{cont}}(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0} \operatorname{Im} \mathcal{H}(\lambda + i\epsilon) \tag{314}$$

where the continuous part is evaluated away from $\lambda = q_1$. For $\nu < 1$, a fraction $1 - \nu$ of the eigenvalues equals q_1 due to the null space of $\boldsymbol{W}^0$, producing the point mass above. This contribution vanishes for $\nu \geq 1$.

Training quantities required at the shared hidden layer. All training matrices below use the time ordering of Appendix E and are restricted to $s, s' < t$. Besides $\boldsymbol{C}_h^0, \boldsymbol{C}_h^1, \boldsymbol{C}_g^1, \boldsymbol{C}_g^2$ and $\boldsymbol{R}_{hr}^0, \boldsymbol{R}_{hr}^1, \boldsymbol{R}_{gu}^1, \boldsymbol{R}_{gu}^2$, the shared hidden layer requires the direct responses $\boldsymbol{R}_{h^1u^1}, \boldsymbol{R}_{g^1r^1}$ and the mixed covariance $\boldsymbol{C}^{h^1g^1} = \langle \boldsymbol{h}^1(\boldsymbol{g}^1)^\top \rangle$. They follow from the existing first-level solution.

Let keep the strictly lower-triangular part of a time matrix, and set

$$\boldsymbol{K}_h = \chi \boldsymbol{R}_{hr}^0 + \kappa(\boldsymbol{C}_h^0), \qquad \boldsymbol{K}_g = (\boldsymbol{R}_{gu}^2) + \kappa(\boldsymbol{C}_g^2). \tag{315}$$

The single-site training equations are

$$\boldsymbol{h}^1 = \boldsymbol{u}^1 + \boldsymbol{K}_h \boldsymbol{g}^1, \qquad \boldsymbol{g}^1 = \boldsymbol{r}^1 + \boldsymbol{K}_g \boldsymbol{h}^1, \tag{316}$$

with independent $\boldsymbol{u}^1 \sim \mathcal{N}(0, \boldsymbol{C}_h^0)$ and $\boldsymbol{r}^1 \sim \mathcal{N}(0, \boldsymbol{C}_g^2)$. Solving gives

$$\begin{aligned} \boldsymbol{R}_{h^1u^1} &= (\boldsymbol{I} - \boldsymbol{K}_h \boldsymbol{K}_g)^{-1}, \quad \boldsymbol{R}_{g^1r^1} = (\boldsymbol{I} - \boldsymbol{K}_g \boldsymbol{K}_h)^{-1}, \\ \boldsymbol{R}_{hr}^1 &= \boldsymbol{R}_{h^1u^1} \boldsymbol{K}_h, \qquad \boldsymbol{R}_{gu}^1 = \boldsymbol{R}_{g^1r^1} \boldsymbol{K}_g. \end{aligned} \tag{317}$$

Therefore

$$\boldsymbol{C}^{h^1g^1} = \boldsymbol{R}_{h^1u^1} \boldsymbol{C}_h^0 (\boldsymbol{R}_{gu}^1)^\top + \boldsymbol{R}_{hr}^1 \boldsymbol{C}_g^2 (\boldsymbol{R}_{g^1r^1})^\top. \tag{318}$$

No additional training fixed point is required. Retain equal-time entries of $\boldsymbol{R}_{hr}^0$ in Eq. 315; this response is not strictly masked. For uncentered population training, $R_{hr}^0(0,0) = -1$.

Project onto training fields and close the mixed sectors. Eliminating the middle probes gives

$$\begin{aligned} \psi_1 &= P\left[u_{\zeta_1} + a u_{\zeta_3} + (\boldsymbol{g}^1)^\top \boldsymbol{b}_0 + a(\boldsymbol{h}^1)^\top \boldsymbol{b}_3\right], \\ \psi_3 &= P\left[b u_{\zeta_1} + u_{\zeta_3} + b(\boldsymbol{g}^1)^\top \boldsymbol{b}_0 + (\boldsymbol{h}^1)^\top \boldsymbol{b}_3\right]. \end{aligned} \tag{319}$$

Stein's lemma and Eq. 309 imply

$$\begin{aligned} \langle \boldsymbol{g}^1 u_{\zeta_1} \rangle &= \boldsymbol{R}_{gu}^1 \boldsymbol{c}_0, \qquad \langle \boldsymbol{g}^1 u_{\zeta_3} \rangle = \boldsymbol{R}_{g^1r^1} \boldsymbol{c}_3, \\ \langle \boldsymbol{h}^1 u_{\zeta_1} \rangle &= \boldsymbol{R}_{h^1u^1} \boldsymbol{c}_0, \quad \langle \boldsymbol{h}^1 u_{\zeta_3} \rangle = \boldsymbol{R}_{hr}^1 \boldsymbol{c}_3. \end{aligned} \tag{320}$$

The remaining pairs follow by the same projection and response calculation. Collecting the eight projection equations, we get

$$\begin{aligned} \mathcal{H}^{-1} \boldsymbol{c}_0 &= \chi \boldsymbol{R}_{hr}^0 \boldsymbol{c}_1 + \boldsymbol{C}_h^0 \boldsymbol{b}_1, \\ \mathcal{H}^{-1} \boldsymbol{\rho}_0 &= (\boldsymbol{R}_{hr}^0)^\top \boldsymbol{b}_1, \\ P^{-1} \boldsymbol{c}_1 &= b \boldsymbol{R}_{gu}^1 \boldsymbol{c}_0 + \boldsymbol{R}_{g^1r^1} \boldsymbol{c}_3 + b \boldsymbol{C}_g^1 \boldsymbol{b}_0 + (\boldsymbol{C}^{h^1g^1})^\top \boldsymbol{b}_3, \\ P^{-1} \boldsymbol{\rho}_1 &= b(\boldsymbol{R}_{gu}^1)^\top \boldsymbol{b}_0 + (\boldsymbol{R}_{h^1u^1})^\top \boldsymbol{b}_3, \\ P^{-1} \boldsymbol{c}_2 &= \boldsymbol{R}_{h^1u^1} \boldsymbol{c}_0 + a \boldsymbol{R}_{hr}^1 \boldsymbol{c}_3 + \boldsymbol{C}^{h^1g^1} \boldsymbol{b}_0 + a \boldsymbol{C}_h^1 \boldsymbol{b}_3, \\ P^{-1} \boldsymbol{\rho}_2 &= (\boldsymbol{R}_{g^1r^1})^\top \boldsymbol{b}_0 + a(\boldsymbol{R}_{hr}^1)^\top \boldsymbol{b}_3, \\ S^{-1} \boldsymbol{c}_3 &= \boldsymbol{R}_{gu}^2 \boldsymbol{c}_2 + \boldsymbol{C}_g^2 \boldsymbol{b}_2, \\ S^{-1} \boldsymbol{\rho}_3 &= (\boldsymbol{R}_{gu}^2)^\top \boldsymbol{b}_2. \end{aligned} \tag{321}$$

The determinant matrix. Stack $\boldsymbol{x}_i = (\boldsymbol{c}_i^\top, \boldsymbol{\rho}_i^\top)^\top \in \mathbb{C}^{2t}$, real on the upper exterior branch, and let $\boldsymbol{x} = (\boldsymbol{x}_0^\top, \boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \boldsymbol{x}_3^\top)^\top$. Substitute Eq. 306 and define the following $2t \times 2t$ blocks:

$$\begin{aligned} \boldsymbol{K}_{01} &= \begin{pmatrix} \chi \boldsymbol{R}_{hr}^0 + \kappa \boldsymbol{C}_h^0 & \boldsymbol{C}_h^0 \\ \kappa(\boldsymbol{R}_{hr}^0)^\top & (\boldsymbol{R}_{hr}^0)^\top \end{pmatrix}, \qquad \boldsymbol{K}_{10} = \begin{pmatrix} \boldsymbol{R}_{gu}^1 + \kappa \boldsymbol{C}_g^1 & \chi \boldsymbol{C}_g^1 \\ \kappa(\boldsymbol{R}_{gu}^1)^\top & \chi(\boldsymbol{R}_{gu}^1)^\top \end{pmatrix}, \\ \boldsymbol{K}_{13} &= \begin{pmatrix} \boldsymbol{R}_{g^1r^1} + \kappa(\boldsymbol{C}^{h^1g^1})^\top & (\boldsymbol{C}^{h^1g^1})^\top \\ \kappa(\boldsymbol{R}_{h^1u^1})^\top & (\boldsymbol{R}_{h^1u^1})^\top \end{pmatrix}, \quad \boldsymbol{K}_{20} = \begin{pmatrix} \boldsymbol{R}_{h^1u^1} + \kappa \boldsymbol{C}^{h^1g^1} & \chi \boldsymbol{C}^{h^1g^1} \\ \kappa(\boldsymbol{R}_{g^1r^1})^\top & \chi(\boldsymbol{R}_{g^1r^1})^\top \end{pmatrix}, \\ \boldsymbol{K}_{23} &= \begin{pmatrix} \boldsymbol{R}_{hr}^1 + \kappa \boldsymbol{C}_h^1 & \boldsymbol{C}_h^1 \\ \kappa(\boldsymbol{R}_{hr}^1)^\top & (\boldsymbol{R}_{hr}^1)^\top \end{pmatrix}, \qquad \boldsymbol{K}_{32} = \begin{pmatrix} \boldsymbol{R}_{gu}^2 + \kappa \boldsymbol{C}_g^2 & \boldsymbol{C}_g^2 \\ \kappa(\boldsymbol{R}_{gu}^2)^\top & (\boldsymbol{R}_{gu}^2)^\top \end{pmatrix}. \end{aligned} \tag{322}$$

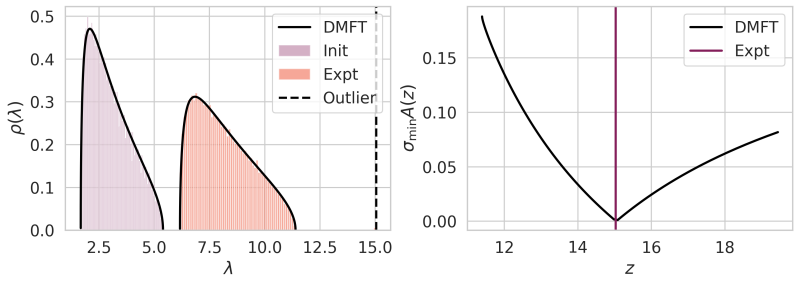


Figure 17: Population NTK spectrum of a two-hidden-layer linear network ($L = 2$). Left: empirical spectral densities at initialization and after $t = 20$ gradient steps, compared with the DMFT bulk predictions (black curves). The dashed line marks the predicted outlier. Right: minimum singular value of $\boldsymbol{\mathcal{A}}(z)$, whose zero locates the outlier and agrees with the empirical eigenvalue (vertical line). DMFT accurately captures both the bulk evolution and the isolated outlier.

Equations Eq. 321 become $\boldsymbol{\mathcal{A}}(z)\boldsymbol{x} = \boldsymbol{x}_{\mathrm{src}}$, with

$$\boldsymbol{\mathcal{A}}(z) = \begin{pmatrix} \mathcal{H}^{-1}\boldsymbol{I} & -\boldsymbol{K}_{01} & \mathbf{0} & \mathbf{0} \\ -b\boldsymbol{K}_{10} & P^{-1}\boldsymbol{I} & \mathbf{0} & -\boldsymbol{K}_{13} \\ -\boldsymbol{K}_{20} & \mathbf{0} & P^{-1}\boldsymbol{I} & -a\boldsymbol{K}_{23} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{K}_{32} & S^{-1}\boldsymbol{I} \end{pmatrix}, \qquad \det \boldsymbol{\mathcal{A}}(z) = 0. \tag{323}$$

Here every identity block is $2t \times 2t$, so $\boldsymbol{\mathcal{A}}$ is $8t \times 8t$. Real zeros on the physical branch outside the limiting bulk are corresponds to isolated eigenvalues; a nonzero residue in the corresponding projected resolvent selects observable poles.

F Large Output Channels

Consider a linear network with input dimension D, width N, and C output channels. Let $\boldsymbol{X} \in \mathbb{R}^{D\times P}$ be the data matrix, $\boldsymbol{W} \in \mathbb{R}^{N\times D}$ the hidden weight matrix, and $\boldsymbol{A} \in \mathbb{R}^{N\times C}$ the readout. We study the weight covariance

$$\boldsymbol{M}(1) = \frac{1}{N}\boldsymbol{W}(1)^\top \boldsymbol{W}(1) \tag{324}$$

as $D, N, C \to \infty$ at fixed positive ratios $\nu = N/D$ and $\chi = C/D$.

To make the dimension-dependent scaling explicit, let $\bar{\gamma}_0$ and $\bar{\eta}$ denote the unrescaled output-normalization parameter and learning-rate parameter. The network output and teacher are

$$\boldsymbol{F} = \frac{1}{\bar{\gamma}_0 N}\boldsymbol{A}^\top \boldsymbol{W}\boldsymbol{X}, \qquad \boldsymbol{Y} = \frac{1}{\sqrt{D}}\boldsymbol{B}^\star \boldsymbol{X}, \tag{325}$$

where $\boldsymbol{B}^\star \in \mathbb{R}^{C\times D}$. We take $\boldsymbol{W}(0)$, $\boldsymbol{A}(0)$, and $\boldsymbol{B}^\star$ to have independent $\mathcal{N}(0,1)$ entries, and write $\boldsymbol{A} = \boldsymbol{A}(0)$ below.

We consider the quadratic loss $\mathcal{L} = \frac{1}{2CP}\|\boldsymbol{F} - \boldsymbol{Y}\|^2$ in the population limit, where $\frac{1}{P}\boldsymbol{X}\boldsymbol{X}^\top \to \boldsymbol{I}_D$. Under μP update $\boldsymbol{W}(1) = \boldsymbol{W}(0) - \bar{\eta}\bar{\gamma}_0^2 N \nabla_{\boldsymbol{W}}\mathcal{L}$, the exact population step is

$$\boldsymbol{W}(1) = \boldsymbol{W}(0) + \frac{\bar{\eta}\bar{\gamma}_0}{C\sqrt{D}}\boldsymbol{A}\boldsymbol{B}^\star - \frac{\bar{\eta}}{CN}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{W}(0). \tag{326}$$

For the extensive-output regime, we choose the additional scaling

$$\bar{\eta} = \sqrt{C}\,\eta, \qquad \bar{\gamma}_0 = \sqrt{D}\,\gamma_0, \tag{327}$$

with η and γ_0 fixed. This gives

$$\boldsymbol{W}(1) = \boldsymbol{W}(0) + \frac{\eta\gamma_0}{\sqrt{C}}\boldsymbol{A}\boldsymbol{B}^\star - \frac{\eta}{\sqrt{C}\,N}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{W}(0). \tag{328}$$

Define the leading-order weight matrix

$$\widetilde{\boldsymbol{W}} = \boldsymbol{W}(0) + \frac{\eta\gamma_0}{\sqrt{C}}\boldsymbol{A}\boldsymbol{B}^\star. \tag{329}$$

At fixed ν and χ, its operator norm is $\|\widetilde{\boldsymbol{W}}\|_{\text{op}} = \mathcal{O}(\sqrt{D})$, whereas the omitted term has operator norm $\|\frac{AA^\top W(0)}{\sqrt{C}N}\|_{\text{op}} = \mathcal{O}(1)$. Consequently, in the asymptotic limit

$$\left\| \boldsymbol{M}(1) - \frac{1}{N}\widetilde{\boldsymbol{W}}^\top\widetilde{\boldsymbol{W}} \right\|_{\text{op}} = \mathcal{O}(D^{-1/2}) \longrightarrow 0. \tag{330}$$

Thus the omitted term does not affect the limiting spectral density. Below we therefore analyze the leading-order covariance, denoted simply by $\boldsymbol{M}$

$$\boldsymbol{M} = \frac{1}{N}\left(\boldsymbol{W}(0) + \frac{\eta\gamma_0}{\sqrt{C}}\boldsymbol{A}\boldsymbol{B}^\star\right)^\top\left(\boldsymbol{W}(0) + \frac{\eta\gamma_0}{\sqrt{C}}\boldsymbol{A}\boldsymbol{B}^\star\right). \tag{331}$$

At initialization, $\boldsymbol{W}(0)$, $\boldsymbol{A}$, and $\boldsymbol{B}^\star$ are independent random matrices, which makes the model tractable for the extensive output regime.

To study the spectrum of $\boldsymbol{M}$, we introduce an auxiliary linear response problem driven by a source $\boldsymbol{j}(t)$

$$\frac{d\boldsymbol{v}(t)}{dt} = -\boldsymbol{M}\boldsymbol{v}(t) + \boldsymbol{j}(t). \tag{332}$$

The reason for introducing this dynamics is that its response function is exactly the resolvent of $\boldsymbol{M}$. Indeed, the causal solution is $\boldsymbol{v}(t) = \int_{-\infty}^{t} ds\, e^{-\boldsymbol{M}(t-s)}\boldsymbol{j}(s)$, so that $\frac{\delta\boldsymbol{v}(t+\tau)}{\delta\boldsymbol{j}(t)} = \Theta(\tau)e^{-\boldsymbol{M}\tau}$. Taking the Fourier transform of the normalized trace of this response gives

$$\begin{aligned}\mathcal{H}(\omega) &= \frac{1}{D}\text{Tr}\int_0^\infty d\tau\, e^{-(i\omega\boldsymbol{I}_D+\boldsymbol{M})\tau} \\ &= \frac{1}{D}\text{Tr}\left(i\omega\boldsymbol{I}_D + \boldsymbol{M}\right)^{-1}.\end{aligned} \tag{333}$$

Thus the spectral density of $\boldsymbol{M}$ can be recovered from the analytic continuation of $\mathcal{H}(\omega)$.

We implement this response calculation using a Martin–Siggia–Rose generating functional (MGF)

$$\mathcal{Z}[\boldsymbol{j}] = \int D\boldsymbol{v}^0 D\hat{\boldsymbol{v}}^0 \exp\left[i\int dt\, \hat{\boldsymbol{v}}^0(t)^\top\left(\partial_t\boldsymbol{v}^0(t) + \boldsymbol{M}\boldsymbol{v}^0(t) - \boldsymbol{j}(t)\right)\right]. \tag{334}$$

The remaining task is to average this generating functional over the random matrices entering $\boldsymbol{M}$. To expose the random-matrix structure, we represent the action of $\boldsymbol{M}$ on $\boldsymbol{v}^0(t)$ through the auxiliary fields

$$\boldsymbol{v}^1(t) = \frac{1}{\sqrt{D}}\boldsymbol{B}\boldsymbol{v}^0(t), \tag{335}$$

$$\boldsymbol{v}^2(t) = \frac{1}{\sqrt{D}}\boldsymbol{W}(0)\boldsymbol{v}^0(t) + \frac{\gamma}{\sqrt{C}}\boldsymbol{A}\boldsymbol{v}^1(t), \tag{336}$$

$$\boldsymbol{v}^3(t) = \frac{\sqrt{C}}{N}\boldsymbol{A}^\top\boldsymbol{v}^2(t), \tag{337}$$

$$\boldsymbol{v}^4(t) = \frac{\sqrt{D}}{N}\boldsymbol{W}(0)^\top\boldsymbol{v}^2(t) + \gamma\frac{\sqrt{D}}{C}\boldsymbol{B}^\top\boldsymbol{v}^3(t). \tag{338}$$

With these definitions,

$$\boldsymbol{v}^4(t) = \boldsymbol{M}\boldsymbol{v}^0(t) \tag{339}$$

and resolvent is obtained from the response of the linear system

$$\partial_t\boldsymbol{v}^0(t) = -\boldsymbol{v}^4(t) + \boldsymbol{j}(t). \tag{340}$$

We enforce the above using the response fields $\hat{\boldsymbol{v}}^i$ through the integral representation of the Dirac delta function, e.g.

$$1 \equiv \int D\boldsymbol{v}^1 D\hat{\boldsymbol{v}}^1 \exp\left[i\int dt\, \hat{\boldsymbol{v}}^1(t)^\top\left(\boldsymbol{v}^1(t) - \frac{1}{\sqrt{D}}\boldsymbol{B}\boldsymbol{v}^0(t)\right)\right]. \tag{341}$$

Since $\boldsymbol{W}(0)$, $\boldsymbol{A}$, and $\boldsymbol{B}$ are independent Gaussian matrices, their averages generate only quadratic contractions between the dynamical fields. The resulting effective theory depends on the correlation order parameters

$$C_{00}(t,t') = \frac{1}{D}\boldsymbol{v}^0(t)\cdot\boldsymbol{v}^0(t'), \qquad C_{11}(t,t') = \frac{1}{C}\boldsymbol{v}^1(t)\cdot\boldsymbol{v}^1(t'), \tag{342}$$

$$C_{22}(t,t') = \frac{1}{N}\boldsymbol{v}^2(t)\cdot\boldsymbol{v}^2(t'), \qquad C_{33}(t,t') = \frac{1}{C}\boldsymbol{v}^3(t)\cdot\boldsymbol{v}^3(t'), \tag{343}$$

and response order parameters

$$R_{04}(t,t') = -\frac{i}{D}\boldsymbol{v}^0(t)\cdot\hat{\boldsymbol{v}}^4(t'), \qquad R_{13}(t,t') = -\frac{i}{C}\boldsymbol{v}^1(t)\cdot\hat{\boldsymbol{v}}^3(t'), \tag{344}$$

$$R_{22}(t,t') = -\frac{i}{N}\boldsymbol{v}^2(t)\cdot\hat{\boldsymbol{v}}^2(t'), \qquad R_{31}(t,t') = -\frac{i}{C}\boldsymbol{v}^3(t)\cdot\hat{\boldsymbol{v}}^1(t') \tag{345}$$

which converge to deterministic quantities in the high-dimensional limit.

F.1 DMFT ACTION

At large D, N, C, with $\nu = N/D$ and $\chi = C/D$ fixed, the disorder-averaged generating functional factorizes into independent single-site problems for the D-dimensional sector $(\boldsymbol{v}^0, \boldsymbol{v}^4)$, the N-dimensional sector $\boldsymbol{v}^2$, and the C-dimensional sector $(\boldsymbol{v}^1, \boldsymbol{v}^3)$

$$\mathcal{Z}[\boldsymbol{j}] = \int d\mathcal{Q}\, \exp\left\{DS_{\text{eff}}[\mathcal{Q};\boldsymbol{j}]\right\}, \tag{346}$$

where

$$\mathcal{Q} = \{C_{00}, C_{11}, C_{22}, C_{33}, R_{04}, R_{13}, R_{22}, R_{31}\}, \tag{347}$$

while the DMFT action can be compressed as

$$S_{\text{eff}} = S_{\text{op}}[\mathcal{Q}] + \log Z_{04}[\boldsymbol{j}] + \nu \log Z_2 + \chi \log Z_{13}. \tag{348}$$

Here, S_{op} collects the order-parameter terms generated by the Gaussian averages, while Z_{04}, Z_2, and Z_{13} are the corresponding single-site path integrals. The response-sector saddle point equations from $\frac{\partial S_{\text{eff}}}{\partial R} = 0$ are

$$\hat{R}_{04} = R_{22} + \gamma R_{31} \qquad \hat{R}_{13} = \gamma R_{22} \tag{349}$$

$$\hat{R}_{22} = \frac{1}{\nu}R_{04} + \gamma\frac{\chi}{\nu}R_{13} \qquad \hat{R}_{31} = \frac{\gamma}{\chi}R_{04}. \tag{350}$$

Substituting these back into the effective action gives three coupled single-site theories. Specifically, the D-sector, which contains the resolvent response, is

$$Z_{04}[\boldsymbol{j}] = \int Dv^0 D\hat{v}^0 Dv^4 D\hat{v}^4 \exp\left\{S_{04}[\boldsymbol{j}]\right\}, \tag{351}$$

with

$$\begin{aligned} S_{04}[\boldsymbol{j}] = i\int dt\, \hat{v}^0(t)\left(\partial_t v^0(t) + v^4(t) - j(t)\right) - \frac{1}{2}\int dtdt'\, \hat{C}_{00}(t,t')v^0(t)v^0(t') \\ - \frac{1}{2}\int dtdt'\, \left[\frac{1}{\nu}C_{22}(t,t') + \frac{\gamma^2}{\chi}C_{33}(t,t')\right]\hat{v}^4(t)\hat{v}^4(t') \\ - i\int dtdt'\, \left[R_{22}(t,t') + \gamma R_{31}(t,t')\right]v^0(t)\hat{v}^4(t') + i\int dt v_4(t)\hat{v}_4(t). \end{aligned} \tag{352}$$

The N-sector is

$$Z_2 = \int Dv^2 D\hat{v}^2 \exp\left\{S_2\right\}, \tag{353}$$

with

$$\begin{aligned} S_2 = -\frac{1}{2}\int dtdt'\, \hat{C}_{22}(t,t')v^2(t)v^2(t') - \frac{1}{2}\int dtdt'\, \left[C_{00}(t,t') + \gamma^2 C_{11}(t,t')\right]\hat{v}^2(t)\hat{v}^2(t') \\ - i\int dtdt'\, \left[\frac{1}{\nu}R_{04}(t,t') + \gamma\frac{\chi}{\nu}R_{13}(t,t')\right]v^2(t)\hat{v}^2(t') + i\int dt\, v^2(t)\hat{v}^2(t). \end{aligned} \tag{354}$$

Finally, the C-sector is

$$Z_{13} = \int Dv^1 D\hat{v}^1 Dv^3 D\hat{v}^3 \exp\{S_{13}\}, \tag{355}$$

with

$$\begin{aligned} S_{13} = &-\frac{1}{2}\int dtdt'\, \hat{C}_{11}(t,t')v^1(t)v^1(t') - \frac{1}{2}\int dtdt'\, \hat{C}_{33}(t,t')v^3(t)v^3(t') \\ &- \frac{1}{2}\frac{\chi}{\nu}\int dtdt'\, C_{22}(t,t')\hat{v}^3(t)\hat{v}^3(t') - \frac{1}{2}\int dtdt'\, C_{00}(t,t')\hat{v}^1(t)\hat{v}^1(t') \\ &- i\gamma\int dtdt'\, R_{22}(t,t')v^1(t)\hat{v}^3(t') - i\frac{\gamma}{\chi}\int dtdt'\, R_{04}(t,t')v^3(t)\hat{v}^1(t') \\ &+ i\int dt\, v^1(t)\hat{v}^1(t) + i\int dt\, v^3(t)\hat{v}^3(t). \end{aligned} \tag{356}$$

F.1.1 Saddle point equations

The saddle point of the effective action identifies the order parameters with the corresponding single-site correlations and responses. We start by differentiating with respect to the conjugate order parameters for the correlation functions which give

$$C_{00}(t,t') = \left\langle v^0(t)v^0(t')\right\rangle_{04} \qquad C_{11}(t,t') = \left\langle v^1(t)v^1(t')\right\rangle_{13} \tag{357}$$

$$C_{22}(t,t') = \left\langle v^2(t)v^2(t')\right\rangle_{2} \qquad C_{33}(t,t') = \left\langle v^3(t)v^3(t')\right\rangle_{13}. \tag{358}$$

while the response functions are

$$R_{04}(t,t') = \left\langle \frac{\delta v^0(t)}{\delta u_4(t')}\right\rangle_{04} \qquad R_{22}(t,t') = \left\langle \frac{\delta v^2(t)}{\delta u_2(t')}\right\rangle_{2} \tag{359}$$

$$R_{13}(t,t') = \left\langle \frac{\delta v^1(t)}{\delta u_3(t')}\right\rangle_{13} \qquad R_{31}(t,t') = \left\langle \frac{\delta v^3(t)}{\delta u_1(t')}\right\rangle_{13}. \tag{360}$$

For causality conditions, the conjugate correlation fields vanish at the saddle

$$\hat{C}_{00} = \hat{C}_{11} = \hat{C}_{22} = \hat{C}_{33} = 0. \tag{361}$$

The final DMFT equations are then

$$v^2(t) = u_2(t) + \sum_{t'}\left[\frac{1}{\nu}R_{04}(t,t') + \gamma\frac{\chi}{\nu}R_{13}(t,t')\right]v^2(t') \tag{362}$$

where

$$u_2 \sim \mathcal{N}\left(0,\, C_{00} + \gamma^2 C_{11}\right). \tag{363}$$

The C-sector gives the coupled equations

$$v^1(t) = u_1(t) + \frac{\gamma}{\chi}\sum_{t'} R_{04}(t,t')v^3(t') \tag{364}$$

$$v^3(t) = u_3(t) + \gamma\sum_{t'} R_{22}(t,t')v^1(t') \tag{365}$$

where

$$u_1 \sim \mathcal{N}(0, C_{00}) \qquad u_3 \sim \mathcal{N}\left(0,\, \frac{\chi}{\nu}C_{22}\right) \tag{366}$$

and the D-sector gives the resolvent probe

$$v^4(t) = u_4(t) + \sum_{t'}\left[R_{22}(t,t') + \gamma R_{31}(t,t')\right]v^0(t') \tag{367}$$

$$\partial_t v^0(t) = -u_4(t) - \sum_{t'}\left[R_{22}(t,t') + \gamma R_{31}(t,t')\right]v^0(t') + j(t) \tag{368}$$

where

$$u_4 \sim \mathcal{N}\left(0,\, \frac{1}{\nu}C_{22} + \frac{\gamma^2}{\chi}C_{33}\right). \tag{369}$$

These equations form the DMFT closure: the correlation functions are computed from the stochastic processes above, while the response functions are obtained by differentiating the same processes with respect to their Gaussian noises.

Since the linear system is time-translation-invariant (T.T.I.), in Fourier space

$$\int dt\, e^{-i\omega t}\left(\frac{\partial}{\partial t'}v^0(t') + u_4(t) + \sum_{t'}\Big(R_{2,2}(t,t') + \gamma R_{3,1}(t,t')\Big)v^0(t') - j(t)\right) \tag{370}$$

$$\Rightarrow \Big(i\omega + R_{2,2}(\omega) + \gamma R_{3,1}(\omega)\Big)v_0(\omega) + u_4(\omega) - j(\omega) = 0 \tag{371}$$

$$\Rightarrow \mathcal{H}(\omega) = \frac{1}{i\omega + R_{2,2}(\omega) + \gamma R_{3,1}(\omega)} \tag{372}$$

and since R_{04} is the response of v^0 to u_4, one has

$$R_{04}(\omega) = -\mathcal{H}(\omega). \tag{373}$$

The remaining response functions can be solved algebraically from the single-site equations. Defining

$$\Delta(\omega) = 1 + \frac{\gamma^2}{\chi}\mathcal{H}(\omega)R_{22}(\omega) \tag{374}$$

we obtain

$$R_{31}(\omega) = \frac{\gamma R_{22}(\omega)}{\Delta(\omega)} \tag{375}$$

$$R_{13}(\omega) = -\frac{\gamma}{\chi}\frac{\mathcal{H}(\omega)}{\Delta(\omega)} \tag{376}$$

$$R_{22}(\omega) = \left[1 + \frac{1}{\nu}\mathcal{H}(\omega) + \frac{\gamma^2}{\nu}\frac{\mathcal{H}(\omega)}{\Delta(\omega)}\right]^{-1}. \tag{377}$$

Together with

$$\mathcal{H}(\omega) = \left[i\omega + R_{22}(\omega) + \frac{\gamma^2 R_{22}(\omega)}{\Delta(\omega)}\right]^{-1} \tag{378}$$

these equations give a closed self-consistency problem for the resolvent.

The spectral density of $\boldsymbol{M}$ is obtained by analytic continuation,

$$\rho(\lambda) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\operatorname{Im}\mathcal{H}(\omega = i\lambda - \epsilon). \tag{379}$$

This set of equations shows explicitly why the extensive-output regime differs from the finite-rank spike setting: when $\chi = \mathcal{O}(1)$, the readout-induced update contributes at the level of the full single-site, modifying the resolvent self-consistently and restructuring the bulk spectrum rather than only producing isolated outliers.

F.2 Edge of the Bulk

Spectral edges correspond to poles of the Stieltjes transform. Suppose we have $G(z) = \operatorname{tr}\left[z - \boldsymbol{M}\right]^{-1} = \frac{1}{z - R(G)}$. We can re-express this as

$$z = \frac{1}{G} + R(G). \tag{380}$$

The derivative $\frac{dz}{dG} = \left(\frac{dG}{dz}\right)^{-1}$ should be zero at the spectral edge since $\frac{dG}{dz}$ diverges there. In our notation, the resolvent is

$$\mathcal{H}(z) = \left[z + R_{22}(z) + \frac{\gamma^2 R_{22}(z)}{1 + \frac{\gamma^2}{\chi}\mathcal{H}(z)R_{22}(z)}\right]^{-1} \tag{381}$$

together with

$$R_{22}(z) = \left[1 + \frac{1}{\nu}\mathcal{H}(z) + \frac{\gamma^2}{\nu}\frac{\mathcal{H}(z)}{1+\frac{\gamma^2}{\chi}\mathcal{H}(z)R_{22}(z)}\right]^{-1}. \tag{382}$$

In general, the bulk edge is obtained by solving these equations together with the condition that the Jacobian of the map from $(\mathcal{H}, R_{22})$ to z becomes singular. A simple closed form is obtained in the limit $\nu \to \infty$. In this case $R_{22}(z) \to 1$, and the resolvent equation reduces to

$$\mathcal{H}(z) = \left[z + 1 + \frac{\gamma^2}{1+\frac{\gamma^2}{\chi}\mathcal{H}(z)}\right]^{-1}. \tag{383}$$

Equivalently, using the physical spectral variable $\lambda = -z$,

$$\lambda(\mathcal{H}) = -\frac{1}{\mathcal{H}} + 1 + \frac{\gamma^2}{1+\frac{\gamma^2}{\chi}\mathcal{H}}. \tag{384}$$

The bulk edges are therefore determined by

$$\frac{d\lambda}{d\mathcal{H}} = \frac{1}{\mathcal{H}^2} - \frac{\gamma^4/\chi}{\left(1+\frac{\gamma^2}{\chi}\mathcal{H}\right)^2} = 0. \tag{385}$$

Solving this equation and substituting back into $\lambda(\mathcal{H})$ gives a shifted Marchenko–Pastur bulk with upper edge

$$\lambda_{\text{edge}} = 1 + \gamma^2\left(1 + \chi^{-1/2}\right)^2. \tag{386}$$

This is consistent with the intuition that, when $N/D \to \infty$, the initialization covariance concentrates as

$$\frac{1}{N}\boldsymbol{W}(0)^\top \boldsymbol{W}(0) \to \boldsymbol{I}_D, \tag{387}$$

while the learned extensive-output term contributes a Wishart-like component proportional to $\boldsymbol{B}^\top\boldsymbol{B}/C$.

F.2.1 Expansion in the $\nu \to \infty$ limit

In the large ν limit, we expect the corrections term to the response function to be

$$R_{22} = 1 + \frac{1}{\nu}R^{(1)} + \mathcal{O}(\nu^{-2}) \tag{388}$$

with

$$R^{(1)} = -\mathcal{H}\Big[1 + \frac{\gamma^2}{1+\frac{\gamma^2}{\chi}\mathcal{H}}\Big]. \tag{389}$$

By expanding the edge to first order in $1/\nu$ as well, one gets

$$\lambda(\mathcal{H}) = -\frac{1}{\mathcal{H}} + 1 + \frac{\gamma^2}{1+\frac{\gamma^2}{\chi}\mathcal{H}} + \frac{1}{\nu}R^{(1)}\Big[1 + \frac{\gamma^2}{(1+\frac{\gamma^2}{\chi}\mathcal{H})^2}\Big] + \mathcal{O}(\nu^{-2}) \tag{390}$$

and therefore $\lambda(\mathcal{H}) = \lambda_0(\mathcal{H}) + \frac{1}{\nu}\lambda_1(\mathcal{H}) + \mathcal{O}(\nu^{-2})$ with

$$\lambda_1(\mathcal{H}) = -\mathcal{H}\Big[1 + \frac{\gamma^2}{1+\frac{\gamma^2}{\chi}\mathcal{H}}\Big]\Big[1 + \frac{\gamma^2}{(1+\frac{\gamma^2}{\chi}\mathcal{H})^2}\Big]. \tag{391}$$

The edge is stable across ν when the relative correction to the bulk edge is small

$$\frac{1}{\nu}\frac{\lambda_1(\mathcal{H})}{\lambda_{\text{edge}}} \ll 1 \tag{392}$$

therefore the condition for the upper edge to have stabilized is

$$\nu \gg \nu_\star = \sqrt{\chi} + \frac{\chi}{\gamma^2(1+\sqrt{\chi})}. \tag{393}$$

Figure 21 verifies this scaling collapse: when the relative edge shift is plotted against $\nu/\nu_\star(\chi,\gamma)$, curves at different χ collapse onto the predicted $\nu_\star/\nu$ behavior at large rescaled width.

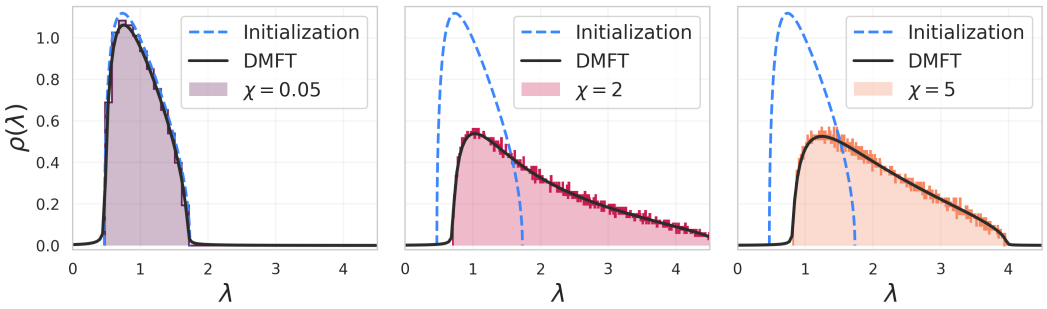


Figure 18: Weight-covariance spectrum after one gradient step for a deep ($L = 2$) linear network in the proportional limit $D, N, C \to \infty$ with fixed $\nu = N/D$ and varying $\chi = C/D$. For small χ, the bulk remains close to the MP initialization spectrum; for larger χ, the bulk shifts and broadens, qualitatively resembling spectra of trained language models.

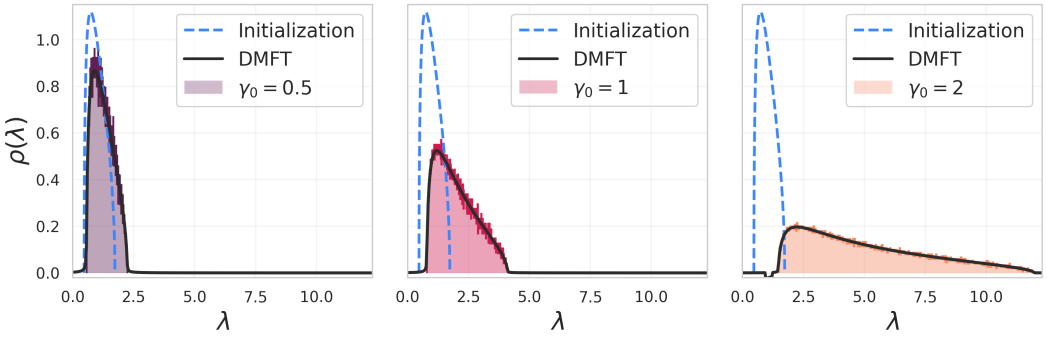
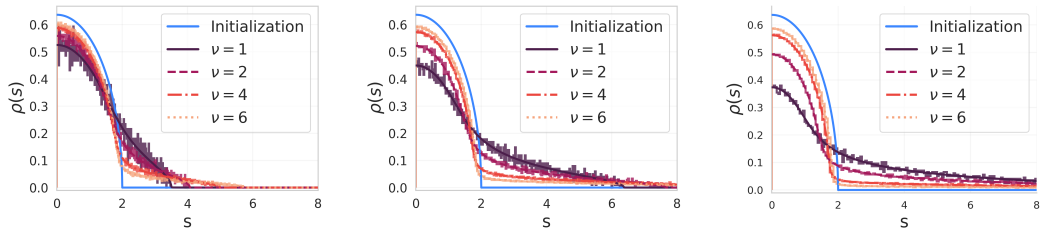


Figure 19: Weight-covariance spectrum of a deep ($L = 2$) linear neural network after one gradient step in the proportional limit at fixed $\nu = 5$, $\chi = 4$ and varying richness γ_0. As γ_0 increases, the bulk shifts and broadens substantially, indicating an extensive learning-induced deformation of the weight covariance.



(a) $\gamma_0 = 1$ (b) $\gamma_0 = 2$ (c) $\gamma_0 = 4$

Figure 20: Depth $L = 3$ linear networks with extensive $C = \mathcal{O}(N)$ classes. Singular value density plots after one gradient step at different ν_s and for different level of richness γ_0.

F.3 Multi-Layer Setting

It is easy to extend the calculation to a multi-layer setting. Let the forward pass be

$$\boldsymbol{H}^\ell = \left(\frac{1}{\sqrt{N}}\boldsymbol{W}^{\ell-1}(0)\right)\ldots\left(\frac{1}{\sqrt{D}}\boldsymbol{W}^0(0)\right) \in \mathbb{R}^{N\times D} \tag{394}$$

and backwards

$$\boldsymbol{G}^{\ell+1} = \left(\frac{1}{\sqrt{N}}\boldsymbol{W}^{\ell+1}(0)^\top\right)\ldots\left(\frac{1}{\sqrt{N}}\boldsymbol{W}^{L-1}(0)^\top\right)\boldsymbol{A} \in \mathbb{R}^{N\times C}. \tag{395}$$

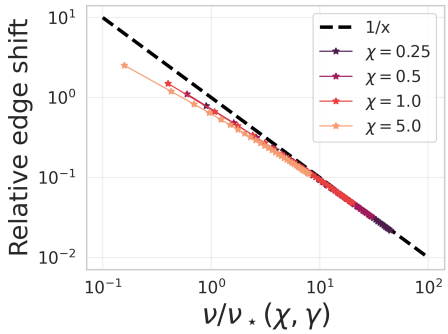


Figure 21: Scaling collapse of the finite-width correction to the upper bulk edge. The relative edge shift $(\lambda_+(\nu) - \lambda_+^\infty)/\lambda_+^\infty$ is plotted against the rescaled width $\nu/\nu_\star(\chi,\gamma)$. Curves for different χ collapse at large $\nu/\nu_\star$ and follow the predicted $\nu_\star/\nu$ scaling, shown as the dashed $1/x$ line.

After one gradient step, the weight update for a generic hidden layer ℓ is

$$\boldsymbol{W}^\ell(1) = \boldsymbol{W}^\ell(0) + \frac{\gamma_0}{\sqrt{CD}}\boldsymbol{G}^{\ell+1}\boldsymbol{B}^\star(\boldsymbol{H}^\ell)^\top. \tag{396}$$

We wish to study the spectrum of

$$\boldsymbol{M}^\ell = \frac{1}{N}\boldsymbol{W}^\ell(1)^\top\boldsymbol{W}^\ell(1). \tag{397}$$

In order to do that, we can define the following fields

$$\boldsymbol{v}^1(t) = \frac{1}{\sqrt{N}}\boldsymbol{H}^{\ell\top}\boldsymbol{v}^0(t) \in \mathbb{R}^D \tag{398}$$

$$\boldsymbol{v}^2(t) = \frac{1}{\sqrt{D}}\boldsymbol{B}^\star\boldsymbol{v}^1(t) \in \mathbb{R}^C \tag{399}$$

$$\boldsymbol{v}^3(t) = \frac{\gamma_0}{\sqrt{C}}\boldsymbol{G}^{\ell+1}\boldsymbol{v}^2(t) + \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\boldsymbol{v}^0(t) \in \mathbb{R}^N \tag{400}$$

$$\boldsymbol{v}^4(t) = \frac{\sqrt{C}}{N}\boldsymbol{G}^{\ell+1\,\top}\boldsymbol{v}_3(t) \in \mathbb{R}^C \tag{401}$$

$$\boldsymbol{v}^5(t) = \frac{\sqrt{D}}{C}\boldsymbol{B}^{\star\top}\boldsymbol{v}^4(t) \in \mathbb{R}^D \tag{402}$$

$$\boldsymbol{v}^6(t) = \frac{\gamma_0\sqrt{N}}{D}\boldsymbol{H}^\ell\boldsymbol{v}^5(t) + \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)^\top\boldsymbol{v}^3(t) \in \mathbb{R}^N. \tag{403}$$

Then

$$\boldsymbol{v}^6(t) = \boldsymbol{M}^\ell\boldsymbol{v}^0(t) \tag{404}$$

and the probe dynamics is simply

$$\partial_t\boldsymbol{v}^0(t) = -\boldsymbol{v}^6(t). \tag{405}$$

And the calculation proceeds by just computing averages over the random independent Gaussian weights. For instance, for three hidden layers linear network, if we to study the spectrum of the

hidden layer $\ell = 1$, the moment generating function can be expressed as

$$\mathcal{Z}[j] = \int \prod_{a=0}^{6} D\boldsymbol{v}^a(t) D\hat{\boldsymbol{v}}^a(t) \; \exp\left[i \int dt \hat{\boldsymbol{v}}^0(t) \cdot \Big(\partial_t \boldsymbol{v}^0(t) + \boldsymbol{v}^6(t) - \boldsymbol{j}(t) \Big) + i \sum_{a=0}^{6} \int dt \boldsymbol{v}^a(t) \cdot \hat{\boldsymbol{v}}^a \right] \tag{406}$$

$$\times \exp\left[- i \int dt \left(\frac{1}{\sqrt{N}} \hat{\boldsymbol{v}}^1(t) \cdot \boldsymbol{W}^{0\top} \boldsymbol{v}^0(t) + \frac{1}{\sqrt{D}} \hat{\boldsymbol{v}}^2(t) \cdot \boldsymbol{B} \boldsymbol{v}^1(t) \right) \right] \tag{407}$$

$$\times \exp\left[- i \int dt \left(\frac{\sqrt{C}}{N} \hat{\boldsymbol{v}}^4(t) \cdot \boldsymbol{A}^\top \boldsymbol{v}_3(t) + \frac{\sqrt{D}}{C} \hat{\boldsymbol{v}}^5(t) \cdot \boldsymbol{B}^\top \boldsymbol{v}^4(t) + \frac{\gamma_0 \sqrt{N}}{D} \hat{\boldsymbol{v}}^6(t) \cdot \boldsymbol{W}^0 \boldsymbol{v}^5(t) \right) \right] \tag{408}$$

$$\times \exp\left[- i \int dt \frac{1}{\sqrt{N}} \hat{\boldsymbol{v}}^6(t) \cdot \boldsymbol{W}^1(0)^\top \boldsymbol{v}^3(t) + \frac{\gamma_0}{\sqrt{C}} \hat{\boldsymbol{v}}^3(t) \cdot \boldsymbol{A} \boldsymbol{v}^2(t) + \frac{1}{\sqrt{N}} \hat{\boldsymbol{v}}^3(t) \cdot \boldsymbol{W}^1(0) \boldsymbol{v}^0(t) \Big) \right] \tag{409}$$

F.3.1 Gaussian averages over the weights

From $\boldsymbol{W}^0(0)$

$$\exp\left[- i \mathrm{Tr} \boldsymbol{W}^0(0)^\top \int dt \Big(\frac{1}{\sqrt{N}} \boldsymbol{v}^0(t) \hat{\boldsymbol{v}}^1(t)^\top + \frac{\gamma_0 \sqrt{N}}{D} \hat{\boldsymbol{v}}^6(t) \boldsymbol{v}^5(t)^\top \Big) \right] \tag{410}$$

$$= \exp\left(- \frac{1}{2} \int dt dt' C_v^0(t,t') \hat{\boldsymbol{v}}^1(t) \cdot \hat{\boldsymbol{v}}^1(t') - \frac{\gamma_0^2 \nu}{2} \int dt dt' C_v^5(t,t') \hat{\boldsymbol{v}}^6(t) \cdot \hat{\boldsymbol{v}}^6(t') \right) \tag{411}$$

$$\times \exp\left(\gamma_0 N \int dt dt' R_{06}(t,t') R_{51}(t',t) \right). \tag{412}$$

From $\boldsymbol{W}^1(0)$

$$\exp\left[- i \mathrm{Tr} \boldsymbol{W}^1(0)^\top \int dt \Big(\frac{1}{\sqrt{N}} \hat{\boldsymbol{v}}^3(t) \boldsymbol{v}^0(t)^\top + \frac{1}{\sqrt{N}} \boldsymbol{v}^3(t) \hat{\boldsymbol{v}}^6(t)^\top \Big) \right] \tag{413}$$

$$= \exp\left(- \frac{1}{2} \int dt dt' C_v^0(t,t') \hat{\boldsymbol{v}}^3(t) \cdot \hat{\boldsymbol{v}}^3(t') - \frac{1}{2} \int dt dt' C_v^3(t,t') \hat{\boldsymbol{v}}^6(t) \cdot \hat{\boldsymbol{v}}^6(t') \right) \tag{414}$$

$$\times \exp\left(N \int dt dt' R_{06}(t,t') R_{33}(t',t) \right). \tag{415}$$

From $\boldsymbol{A}(0)$

$$\exp\left[- i \mathrm{Tr} \boldsymbol{A}(0)^\top \int dt \Big(\frac{\gamma_0}{\sqrt{C}} \hat{\boldsymbol{v}}^3(t) \boldsymbol{v}^2(t)^\top + \frac{\sqrt{C}}{N} \boldsymbol{v}_3(t) \hat{\boldsymbol{v}}^4(t)^\top \Big) \right] \tag{416}$$

$$= \exp\left(- \frac{\gamma_0^2}{2} \int dt dt' C_v^2(t,t') \hat{\boldsymbol{v}}^3(t) \cdot \hat{\boldsymbol{v}}^3(t') - \frac{1}{2} \frac{\chi}{\nu} \int dt dt' C_v^3(t,t') \hat{\boldsymbol{v}}^4(t) \cdot \hat{\boldsymbol{v}}^4(t') \right) \tag{417}$$

$$\times \exp\left(\gamma_0 C \int dt dt' R_{33}(t,t') R_{24}(t',t) \right). \tag{418}$$

From $\boldsymbol{B}^\star = \boldsymbol{B}$

$$\exp\left[-i\mathrm{Tr}\boldsymbol{B}^\top \int dt\Big(\frac{1}{\sqrt{D}}\hat{\boldsymbol{v}}^2(t)\boldsymbol{v}^1(t)^\top + \frac{\sqrt{D}}{C}\boldsymbol{v}^4(t)\hat{\boldsymbol{v}}^5(t)^\top\Big)\right] \tag{419}$$

$$= \exp\left(-\frac{1}{2}\int dtdt' C_v^1(t,t')\hat{\boldsymbol{v}}^2(t)\cdot\hat{\boldsymbol{v}}^2(t') - \frac{1}{2\chi}\int dtdt' C_v^4(t,t')\hat{\boldsymbol{v}}^5(t)\cdot\hat{\boldsymbol{v}}^5(t')\right) \tag{420}$$

$$\times \exp\left(D\int dtdt' R_{42}(t,t')R_{15}(t',t)\right). \tag{421}$$

By enforcing these definitions

$$R_{06}(t,t') = -\frac{i}{N}\boldsymbol{v}^0(t)\cdot\hat{\boldsymbol{v}}^6(t') \tag{422}$$

$$R_{51}(t',t) = -\frac{i}{D}\boldsymbol{v}^5(t')\cdot\hat{\boldsymbol{v}}^1(t) \tag{423}$$

$$R_{33}(t',t) = -\frac{i}{N}\boldsymbol{v}^3(t')\cdot\hat{\boldsymbol{v}}^3(t) \tag{424}$$

$$R_{24}(t',t) = -\frac{i}{C}\boldsymbol{v}^2(t')\cdot\hat{\boldsymbol{v}}^4(t) \tag{425}$$

$$R_{42}(t,t') = -\frac{i}{C}\boldsymbol{v}^4(t)\cdot\hat{\boldsymbol{v}}^2(t') \tag{426}$$

$$R_{15}(t',t) = -\frac{i}{D}\boldsymbol{v}^1(t')\cdot\hat{\boldsymbol{v}}^5(t) \tag{427}$$

one gets at the saddle point

$$\hat{R}_{06} = \gamma_0 R_{51} + R_{33} \tag{428}$$

$$\hat{R}_{51} = \gamma_0\nu R_{06} \tag{429}$$

$$\hat{R}_{33} = R_{06} + \gamma_0\frac{\chi}{\nu}R_{24} \tag{430}$$

$$\hat{R}_{24} = \gamma_0 R_{33} \tag{431}$$

$$\hat{R}_{42} = \frac{1}{\chi}R_{15} \tag{432}$$

$$\hat{R}_{15} = R_{42} \tag{433}$$

giving the DMFT equations

$$v_1(t) = u_1(t) + \gamma_0\nu\int dsR_{06}(t,s)v_5(s), \quad u_1 \sim \mathcal{N}(0, C_v^0) \tag{434}$$

$$v_2(t) = u_2(t) + \frac{1}{\chi}\int dsR_{15}(t,s)v_4(s), \quad \mathcal{N}(0, C_v^1) \tag{435}$$

$$v_3(t) = u_3(t) + \int ds\Big(R_{06} + \gamma_0\frac{\chi}{\nu}R_{24}\Big)v_3(s), \quad u_3 \sim \mathcal{N}(0, C_v^0 + \gamma_0^2 C_v^2) \tag{436}$$

$$v_4(t) = u_4(t) + \gamma_0\int dsR_{33}(t,s)v_2(s), \quad u_4 \sim \mathcal{N}(0, \frac{\chi}{\nu}C_v^3) \tag{437}$$

$$v_5(t) = u_5(t) + \int dsR_{42}(t,s)v_1(s) \quad u_5 \sim \mathcal{N}(0, \frac{1}{\chi}C_v^4) \tag{438}$$

$$v_6(t) = u_6(t) + \int ds\Big(\gamma_0 R_{51}(t,s) + R_{33}(t,s)\Big)v_0(s), \quad u_6 \sim \mathcal{N}(0, C_v^3 + \gamma_0^2\nu C_v^5) \tag{439}$$

and finally

$$\partial_t v_0(t) = -u_6(t) - \int ds\Big(\gamma_0 R_{51}(t,s) + R_{33}(t,s)\Big)v_0(s) + j(t). \tag{440}$$

In Fourier space

$$R_{06}(\omega) = -\frac{1}{i\omega + R_{33}(\omega) + \gamma_0 R_{51}(\omega)} = -\mathcal{H}(\omega) \tag{441}$$

$$R_{15}(\omega) = -\frac{\gamma_0 \nu \mathcal{H}(\omega)}{1 + \gamma_0 \nu \mathcal{H}(\omega) R_{42}(\omega)} \tag{442}$$

$$R_{51}(\omega) = \frac{R_{42}(\omega)}{1 + \gamma_0 \nu R_{42}(\omega) \mathcal{H}(\omega)} \tag{443}$$

$$R_{24}(\omega) = \frac{\frac{1}{\chi} R_{15}(\omega)}{1 - \frac{\gamma_0}{\chi} R_{15}(\omega) R_{33}(\omega)} \tag{444}$$

$$R_{42}(\omega) = \frac{\gamma_0 R_{33}(\omega)}{1 - \frac{\gamma_0}{\chi} R_{33}(\omega) R_{15}(\omega)} \tag{445}$$

$$R_{33}(\omega) = \frac{1}{1 + \mathcal{H}(\omega) - \gamma_0 \frac{\chi}{\nu} R_{24}(\omega)}. \tag{446}$$

The equations above give the closed response-sector DMFT for the spectrum, which can then be obtained from

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\mathcal{H}(i\lambda - \epsilon). \tag{447}$$

G Experimental Details and Additional Figures

G.1 CIFAR

Models are trained on CIFAR-10 Krizhevsky (2009) with SGD in mean-field/μP scaling. The architecture is a residual CNN with 5 residual blocks and constant hidden width N. Training uses batch size 128 for $T = 1000$ gradient steps, without momentum or learning-rate scheduling. We sweep width N and output multiplier γ_0. For spectral plots, a convolutional tensor

$$W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k_1 \times k_2}$$

is square-unfolded as

$$W_{\text{sq}} \in \mathbb{R}^{k_1 C_{\text{out}} \times k_2 C_{\text{in}}} \tag{448}$$

and singular values are computed from

$$\frac{W_{\text{sq}}}{\sqrt{k_2 C_{\text{in}}}}$$

with aspect ratio

$$\phi = \frac{k_1 C_{\text{out}}}{k_2 C_{\text{in}}}. \tag{449}$$

G.2 ImageNet

We train a width-256 μP ResNet18 on ImageNet Russakovsky et al. (2015) using the public Flax/JAX implementation accompanying Vyas et al. Vyas et al. (2023). The model is trained with Adam in the μP parameterization, using minibatches of 1024 images split into microbatches of size 128. We track training time by the number of images seen and save model checkpoints at logarithmically spaced targets, including $10^3, 10^4, 10^5$, and 10^6 images seen. The reported loss is the validation cross-entropy evaluated on a held-out subset of ImageNet.

To probe the evolution of the weight spectrum, we extract convolutional kernels from saved checkpoints and compute their singular-value distributions. In the main plot we use the hidden 256-channel convolutional layer, which does not directly touch either the input pixels or the final readout. For a convolutional kernel $\boldsymbol{W} \in \mathbb{R}^{k_1 \times k_2 \times C\text{in} \times C_{\text{out}}}$, we form the square unfolding

$$\boldsymbol{W}_{\text{flat}} \in \mathbb{R}^{k_1 C_{\text{out}} \times k_2 C_{\text{in}}}$$

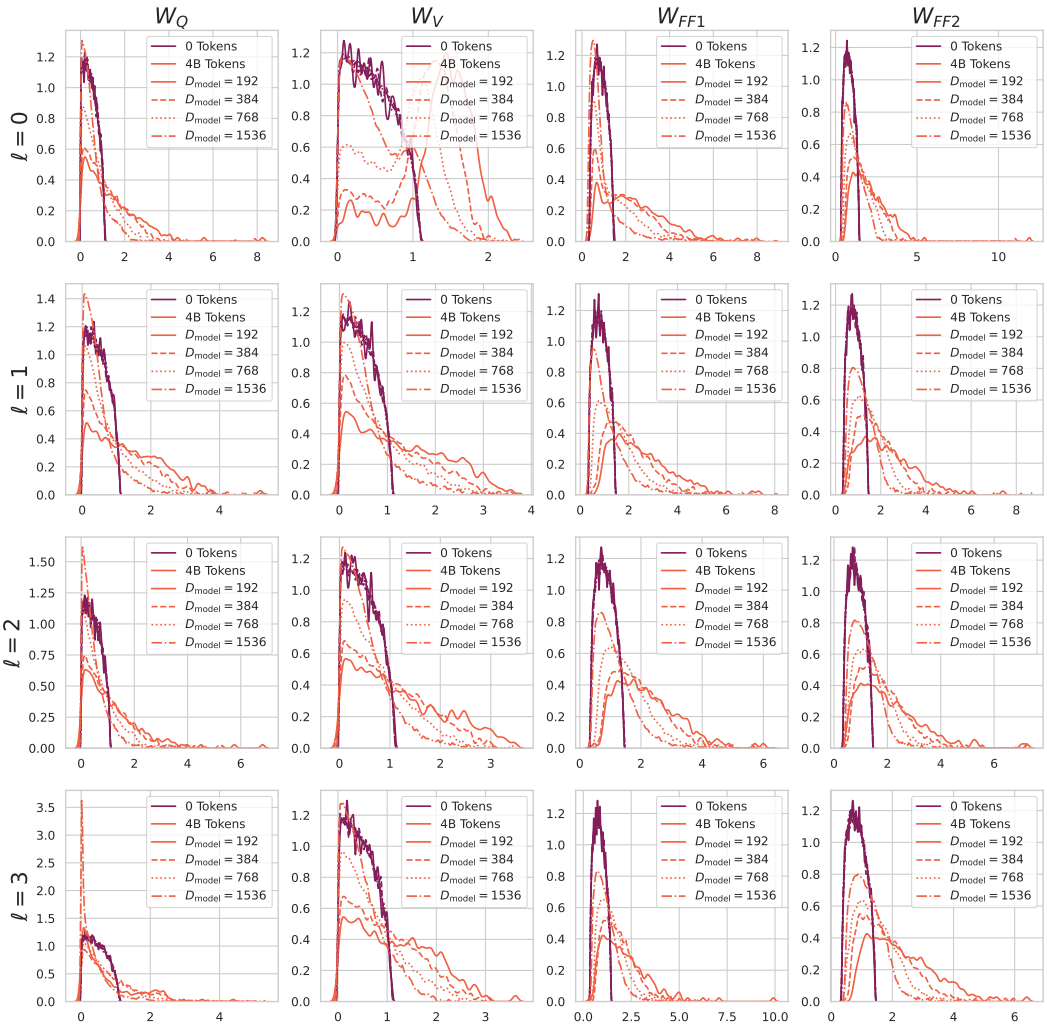


Figure 22: Weight spectra across the first 4 hidden layers of the transformer after pretraining on $T = 4$ B tokens.

obtained by grouping one spatial index with output channels and the other spatial index with input channels. We then normalize, so that an unstructured random matrix with the same aspect ratio has Marchenko–Pastur singular-value edge

$$s_{\text{edge}} = 1 + \sqrt{\frac{k_1 C_{\text{out}}}{k_2 C_{\text{in}}}}.$$

We plot the empirical singular-value density together with singular values exceeding this MP edge.

G.3 GPT Style Pretraining

Models are trained in μP with Adam Yang et al. (2021) on the C4 corpus Raffel et al. (2020). The models are based on the OLMo repository models and training scripts OLMo et al. (2024), but modified to be in μP width scaling and CompleteP depth scaling Yang et al. (2021); Dey et al. (2025). Scaling attention layers is achieved by scaling up the number of attention heads keeping $d_{\text{head}} = 64$ Bordelon et al. (2024b). The base model is $D_{\text{model}} = 768$ with $H = 12$ heads and $L = 6$ hidden layers. The C4 data are tokenized with the GPT-2 tokenizer whose vocabulary is ≈ 50k. Weight decay is 0.1 and a cosine annealing schedule for the $T = 4$B run is used across all the models.

G.4 Compute Resources and Estimate of Compute Time

Theory Plots The theory curves are computed with a single H100 in a jupyter notebook (attached). Most figures can be generated within a few minutes, less than $\sim$ 20 minutes. These were run sequentially and different settings were executed in separate notebooks.

CIFAR-10 Experiments For CIFAR-10, we also use a single H100 on jupyter notebook. Training takes around 10 minutes.

ImageNet We train on ImageNet for one epoch using a single H100 GPU. This takes about 90 minutes.

GPT experiments For the language model pretraining experiments, we use 3x Nvidia A100 per run for 4B tokens varying the widths D_{model}. The runtime ranges between 1-6 hours depending on model size.