Random Matrix Theory of Early-Stopped Gradient Flow: A Transient BBP Scenario

Florentin Coeurdoux[1]* **Grégoire Ferré**[1] **Jean-Philippe Bouchaud**[1,2,3]

[1]Capital Fund Management, Paris, France
[2]X-CFM Chair of EconophysiX, Ecole polytechnique, Palaiseau, France
[3]Académie des Sciences, Paris, France

Abstract

Empirical studies of trained models often report a transient regime in which signal is detectable in a finite gradient descent time window before overfitting dominates. We provide an analytically tractable random-matrix model that reproduces this phenomenon for gradient flow in a linear teacher–student setting. In this framework, learning occurs when an isolated eigenvalue separates from a noisy bulk, before eventually disappearing in the overfitting regime. The key ingredient is anisotropy in the input covariance, which induces fast and slow directions in the learning dynamics. In a two-block covariance model, we derive the full time-dependent bulk spectrum of the symmetrized weight matrix through a 2×2 Dyson equation, and we obtain an explicit outlier condition for a rank-one teacher via a rank-two determinant formula. This yields a transient Baik–Ben Arous–Péché (BBP) transition [1]: depending on signal strength and covariance anisotropy, the teacher spike may never emerge, emerge and persist, or emerge only during an intermediate time interval before being reabsorbed into the bulk. We map the corresponding phase diagrams and validate the theory against finite-size simulations. Our results provide a minimal solvable mechanism for early stopping as a transient spectral effect driven by anisotropy and noise.

1 Introduction

Understanding when gradient-based learning extracts signal from data, and when it instead amplifies noise and overfits, is a central problem in modern learning theory. A recurring empirical observation is that the spectrum of trained weight matrices often passes through a transient regime in which an isolated eigenvalue separates from a noisy bulk before the spectrum becomes more complicated at later stages of training [2, 3, 4, 5]. Although this phenomenon is frequently associated with the practical success of early stopping, its underlying mechanism remains poorly understood.

In this paper we isolate a minimal, analytically tractable mechanism for such a transient spectral separation. We study gradient flow in a linear teacher–student model and show that anisotropy in the input covariance can produce a *transient BBP transition*: an outlier may emerge from the bulk at intermediate times (with a non-zero overlap with the teacher direction) and later be reabsorbed as training continues, in which case the teacher direction cannot be reconstructed [1]. The effect is driven by the coexistence of fast and slow directions in the data covariance. Fast directions learn the teacher early, whereas slow directions continue to inject noise and can eventually hide the signal again.

*Corresponding author: florentin.coeurdoux@cfm.com. Code to reproduce all figures is available at https://github.com/FlorentinCDX/rmt-early-stopping

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Our model is deliberately simple. It is not intended to reproduce the full heavy-tailed phenomenology observed in deep nonlinear networks at convergence; rather, it is designed to capture and explain the *bulk-plus-spike* regime in a setting where the full dynamics can be analyzed explicitly. The key solvable case is a two-block covariance model in which the singular values of the input matrix take only two values. In that simplified yet realistic setting, the bulk spectrum of the symmetrized weight matrix is described by a 2×2 Dyson equation, while the rank-one teacher induces a finite-rank perturbation whose outlier condition can be reduced to a rank-two determinant.

Three regimes emerge from our model, depending on signal strength and input covariance anisotropy: (i) a weak-signal regime in which no outlier ever emerges, so the teacher is never spectrally detectable; (ii) a strong-signal regime in which an outlier emerges and remains separated from the bulk at all times; and (iii) an "early-stopping" regime in which an outlier exists only during a finite time window, providing a precise random-matrix formulation of early stopping.

Contributions In the solvable setting above, we (a) derive a 2×2 Dyson equation characterizing the full time-dependent bulk spectrum of the symmetrized weight matrix; (b) establish an explicit rank-two BBP-type outlier condition valid at every finite time t; (c) classify the possible learning regimes and give closed-form phase diagrams in the (θ, t) and (θ, λ_-) planes; (d) identify numerically an optimal early-stopping time, defined through the teacher–eigenvector overlap, that lies strictly inside the transient BBP window; and (e) map the regimes as a function of the sample ratio γ, and show numerically that the same transient persists for a continuous power-law spectrum and in nonlinear networks (Sections 4.3–4.5).

Related works The BBP phase transition, first discussed by Edwards and Jones [6] and established by Baik, Ben Arous and Péché [1], has since become a standard tool in spiked random matrix models and finite-rank perturbation theory [7]. Exact learning dynamics in linear and deep-linear models, including gradient-flow learning and early stopping, have been studied extensively [8, 9, 10, 11], also with anisotropic inputs [12]; spectral bias is well understood in kernel regression [13], and the exact high-dimensional dynamics of SGD under anisotropic data has been derived [14, 15]. Spectral properties of weight matrices have been linked to generalisation empirically [2, 3, 4, 5, 16, 17] and, in two-layer networks, theoretically: spikes that emerge in the weights after large gradient steps carry the learned features [18, 19, 20], and a teacher with directions of decreasing importance is learned through a cascade of spectral transitions [21, 22]. Dynamical BBP transitions appear in the Hessian of two-layer networks trained by online SGD on Gaussian mixtures [23, 24], an inverse BBP transition in the dynamical Hessian of gradient descent for phase retrieval [25], and a related transition in the Hessian spectrum during feature learning [26]. More broadly, random matrix methods have become increasingly useful in the study of high-dimensional learning dynamics and implicit regularisation [27, 28, 29, 30, 31, 32, 33]. Our contribution is to provide an arguably oversimplified but explicit random-matrix mechanism for a *transient* spectral spike during learning, i.e. a BBP transition followed by an inverse BBP transition, and to tie this transient spectral geometry to a computable optimal stopping time.

2 Model and Setup

2.1 Gradient flow on a linear model

We consider inputs $X \in \mathbb{R}^{N \times M}$ and targets $Y \in \mathbb{R}^{N \times M}$ in the high-dimensional regime

$$N, M \to \infty, \qquad \frac{M}{N} \to \gamma \geq 1.$$

We train a linear model with weight matrix $A \in \mathbb{R}^{N \times N}$ by minimizing

$$A_{\mathrm{LS}} = \arg\min_A J(A), \qquad J(A) = \frac{1}{2}\|AX - Y\|_F^2. \tag{1}$$

Although the minimizer is explicit and reads $A_{\mathrm{LS}} \equiv YX^\top (XX^\top)^{-1}$, our interest is not the optimizer itself but the spectral evolution of the gradient-flow trajectory

$$\frac{dA_t}{dt} = -\nabla J(A_t), \qquad A_0 = A_{\mathrm{init}}.$$

Since

$$\nabla J(A) = (AX - Y)X^\top = AXX^\top - YX^\top,$$

the flow solves the linear matrix ODE

$$\dot{A}_t = -A_t XX^\top + YX^\top.$$

Its explicit solution is

$$A_t = A_{\text{init}} e^{-tXX^\top} + YX^\top (XX^\top)^{-1}\big(I - e^{-tXX^\top}\big). \tag{2}$$

This is a standard semigroup evolution, and in the present least-squares setting it provides a natural continuous-time framework for early stopping [10]. At long times it converges to the least-squares estimator $A_\infty \equiv A_{\text{LS}}$. The key point for us is that the learning speed is anisotropic: the eigenvalues of $XX^\top$ determine how fast the corresponding directions are learned or damped. This observation motivates the two-block covariance model introduced below.

2.2 Random matrix assumptions

To model a teacher–student setting, we assume that the targets are generated by a rank-one teacher

$$Y = A^\theta_{\text{teach}} X + Z, \qquad A^\theta_{\text{teach}} = \theta\, vv^\top, \tag{3}$$

where $v \in \mathbb{R}^N$ is a fixed unit vector and $\theta \geq 0$ is the signal amplitude. The noise matrix $Z \in \mathbb{R}^{N\times M}$ has i.i.d. $\mathcal{N}(0, 1/M)$ entries, independent of everything else. We take A_{init} to be a GOE matrix at the usual $1/\sqrt{N}$ scale, independent of X and Z.

Substituting (3) into (2) yields the decomposition

$$A^\theta_t = \underbrace{A_{\text{init}} e^{-tXX^\top} + ZX^\top (XX^\top)^{-1}\big(I - e^{-tXX^\top}\big)}_{\text{noise part}} + \underbrace{\theta\, vv^\top \big(I - e^{-tXX^\top}\big)}_{\text{teacher part}}. \tag{4}$$

We consider the additive symmetrization

$$S^\theta_t := A^\theta_t + (A^\theta_t)^\top,$$

because it is the simplest construction that preserves the bulk-plus-spike structure of the problem while remaining fully solvable at finite time: the noise contribution becomes a Wigner-type matrix with a piecewise-constant variance profile, and the teacher contribution becomes a finite-rank symmetric perturbation. The more standard multiplicative choice $A_t A_t^\top$ would introduce a Wishart-type bulk whose time-dependent spectrum requires a substantially heavier analysis; we return to this in the conclusion. Using (4), we obtain

$$S^\theta_t = S^0_t + \theta\Big[vv^\top\big(I - e^{-tXX^\top}\big) + \big(I - e^{-tXX^\top}\big)vv^\top\Big], \tag{5}$$

where the pure-noise part is

$$S^0_t = A_{\text{init}} e^{-tXX^\top} + e^{-tXX^\top} A^\top_{\text{init}} + ZX^\top (XX^\top)^{-1}\big(I - e^{-tXX^\top}\big) + \big(I - e^{-tXX^\top}\big)(XX^\top)^{-1} XZ^\top. \tag{6}$$

Our goal is to describe the spectrum of S^θ_t for all times $t \geq 0$ and all $\theta \geq 0$. Note that we alternatively write S_t instead of S^0_t for notational simplicity.

2.3 Two-block structure of the input covariance

To obtain an analytically tractable model, we assume that the singular values of X take only two values: a fraction α are equal to $\lambda_+ = 1$ and a fraction $1 - \alpha$ are equal to $\lambda_- \in (0, 1]$. Equivalently, $XX^\top$ has two eigenvalues, 1 and λ_-^2, with proportions α and $1 - \alpha$. The smaller λ_- is, the more anisotropic the learning problem becomes.[2]

[2]This is where our assumption that $\gamma \geq 1$ is needed. Indeed, when $\gamma < 1$ the matrix $XX^\top$ has a null space of dimension $(1 - \gamma)N$, which leads to a 3×3 Dyson system. See Appendix B.

In order to simplify (6), let us write the compact singular value decomposition

$$X = U\Lambda V^\top,$$

where $U \in O(N)$, $V \in \mathbb{R}^{M\times N}$ has orthonormal columns, and

$$\Lambda = \mathrm{diag}(\underbrace{1,\dots,1}_{\alpha N}, \underbrace{\lambda_-,\dots,\lambda_-}_{(1-\alpha)N}) \in \mathbb{R}^{N\times N}.$$

Conjugating by U and using the rotational invariance of A_{init} and of the Gaussian noise, we obtain

$$\widetilde{A}_{\mathrm{init}} := U^\top A_{\mathrm{init}} U, \qquad \widetilde{Z} := U^\top Z V, \qquad \widetilde{v} := U^\top v,$$

with $\widetilde{A}_{\mathrm{init}}$ distributed as GOE (Gaussian Orthogonal Ensemble) and $\widetilde{Z}$ having i.i.d. Gaussian entries of variance $1/M$. In this basis,

$$U^\top A_t^\theta U = \widetilde{A}_{\mathrm{init}} e^{-t\Lambda^2} + \widetilde{Z}\Lambda^{-1}\big(I - e^{-t\Lambda^2}\big) + \theta\,\widetilde{v}\widetilde{v}^\top\big(I - e^{-t\Lambda^2}\big). \tag{7}$$

Therefore the symmetrized matrix has the same spectrum as

$$\widetilde{S}_t^\theta = \widetilde{S}_t^0 + \theta\big(\widetilde{v}w_t^\top + w_t\widetilde{v}^\top\big), \qquad w_t := \big(I - e^{-t\Lambda^2}\big)\widetilde{v}, \tag{8}$$

where

$$\widetilde{S}_t^0 = \widetilde{A}_{\mathrm{init}} e^{-t\Lambda^2} + e^{-t\Lambda^2}\widetilde{A}_{\mathrm{init}}^\top + \widetilde{Z}\Lambda^{-1}\big(I - e^{-t\Lambda^2}\big) + \big(I - e^{-t\Lambda^2}\big)\Lambda^{-1}\widetilde{Z}^\top. \tag{9}$$

This representation makes the structure transparent: the noise part is a two-block Wigner-type matrix, while the teacher contributes a finite-rank perturbation which is rank one only in the isotropic case $\lambda_- = 1$ and rank two otherwise.

3 Bulk spectrum via the 2×2 Dyson equation

We now characterize the bulk spectrum of the noise matrix in the abstract two-block setting.

3.1 Two-block Wigner-type ensemble

Define the block matrices

$$D_{1,t} = \mathrm{diag}(\underbrace{a_t,\dots,a_t}_{\alpha N}, \underbrace{b_t,\dots,b_t}_{(1-\alpha)N}), \qquad D_{2,t} = \mathrm{diag}(\underbrace{c_t,\dots,c_t}_{\alpha N}, \underbrace{d_t,\dots,d_t}_{(1-\alpha)N}),$$

with (for $\lambda_+ = 1$)

$$a_t = e^{-t}, \qquad b_t = e^{-\lambda_-^2 t}, \qquad c_t = 1 - e^{-t}, \qquad d_t = \frac{1 - e^{-\lambda_-^2 t}}{\lambda_-}. \tag{10}$$

Let $W \in \mathbb{R}^{N\times N}$ be GOE with the usual $1/N$ scaling, and let $\Gamma \in \mathbb{R}^{N\times N}$ have i.i.d. entries of variance $1/M$, independent of W. Then (9) can be rewritten as

$$S_t = W D_{1,t} + D_{1,t} W + \Gamma D_{2,t} + D_{2,t}\Gamma^\top. \tag{11}$$

In the limit $N, M \to \infty$ with $M/N \to \gamma \geq 1$, the off-diagonal variance profile is piecewise constant:

$$N\,\mathbb{E}[S_{ij}^2] \to \sigma_{\tau(i),\tau(j)},$$

where $\tau(i) \in \{A, B\}$ is the block label of i and

$$\sigma_{AA}(t) = 4a_t^2 + \frac{2c_t^2}{\gamma}, \qquad \sigma_{BB}(t) = 4b_t^2 + \frac{2d_t^2}{\gamma}, \qquad \sigma_{AB}(t) = (a_t + b_t)^2 + \frac{c_t^2 + d_t^2}{\gamma}. \tag{12}$$

Thus S is a Wigner-type matrix with a two-block variance profile [29, 28].

3.2 The 2×2 Dyson equation

Let

$$m_A(z,t) = \lim_{N\to\infty} \frac{1}{\alpha N} \sum_{i\in A} \big[(zI - S_t)^{-1}\big]_{ii}, \qquad m_B(z,t) = \lim_{N\to\infty} \frac{1}{(1-\alpha)N} \sum_{i\in B} \big[(zI - S_t)^{-1}\big]_{ii}.$$

In the large-N limit, these partial Stieltjes transforms concentrate to deterministic limits satisfying the following closed self-consistent system.

Proposition 3.1 (Dyson equation). *For each $z \in \mathbb{C}^+$ and $t \geq 0$, the pair $(m_A(z,t), m_B(z,t))$ is the unique solution of*

$$\begin{cases} \dfrac{1}{m_A(z,t)} = z - \alpha\, \sigma_{AA}(t)\, m_A(z,t) - (1-\alpha)\, \sigma_{AB}(t)\, m_B(z,t), \\ \dfrac{1}{m_B(z,t)} = z - \alpha\, \sigma_{AB}(t)\, m_A(z,t) - (1-\alpha)\, \sigma_{BB}(t)\, m_B(z,t), \end{cases} \tag{13}$$

with the Stieltjes-branch conditions $\Im m_A(z,t), \Im m_B(z,t) < 0$ for $\Im z > 0$ and $m_A(z,t), m_B(z,t) \sim 1/z$ as $|z| \to \infty$. The global Stieltjes transform and limiting spectral density are

$$m(z,t) = \alpha\, m_A(z,t) + (1-\alpha)\, m_B(z,t), \qquad \rho_{S_t}(\lambda) = -\frac{1}{\pi} \lim_{\eta\downarrow 0} \Im\, m(\lambda + i\eta, t). \tag{14}$$

A short proof of this result is given in Appendix A. Note that when $\alpha = 1$ or $\alpha = 0$, (13) reduces to the scalar Dyson equation of a semicircle law. When $a = b$ and $c = d$, one has $\sigma_{AA} = \sigma_{AB} = \sigma_{BB}$ and the two-block system again collapses to a single semicircle. In the general case the complex behaviour of the spectral dynamics is illustrated in Figure 1 below. To conclude this part we mention that, should we have postulated an n-block structure, we would have obtained n coupled equations for the corresponding set $\{m_{A_1}(z), m_{A_2}(z), \ldots, m_{A_n}(z)\}$. The $\gamma < 1$ case, which introduces a frozen null-space block, leads to a 3×3 Dyson system and is treated in Appendix B; the resulting permanent noise narrows but does not eliminate the transient regime discussed below.

3.3 Empirical validation and dynamical phases

We validate the Dyson prediction by simulating the full gradient flow (2) at logarithmically spaced times $t \in [0.1, 2000]$, averaging over 20 independent realizations with $N = 500$, $\gamma = 1$, $\alpha = 0.5$, $\lambda_+ = 1$, $\lambda_- = 0.1$. At each time we form $S_t = A_t + A_t^\top$ and compare the empirical spectral density with the theoretical density ρ_{S_t} obtained from the 2×2 Dyson equation via the Stieltjes inversion formula (14).

Three dynamical phases are visible in Figure 1. At early times the spectrum is close to a single semicircle, since both blocks still behave similarly and the matrix is dominated by the GOE initialization. At intermediate times the two learning scales separate, the variance profile becomes genuinely two-block, and the bulk deforms accordingly. At very late times the GOE contribution has essentially decayed, while the factor $(XX^\top)^{-1}$ amplifies the slow block and broadens the support. Across the full dynamical range, the theoretical curve obtained from the Dyson equation agrees well with the empirical histogram. When the rank-one teacher is switched on ($\theta = 6$, red vertical lines), one clearly sees how an outlier first detaches from the bulk at intermediate times and is then reabsorbed at late times; the full analysis of this transient BBP transition is the subject of the next section. Numerical experiments with a full-rank power-law input spectrum, where the covariance eigenvalues are drawn from a power-law distribution, confirm that the same three dynamical phases persist well beyond the two-block setting (Figure 5 and Appendix F).

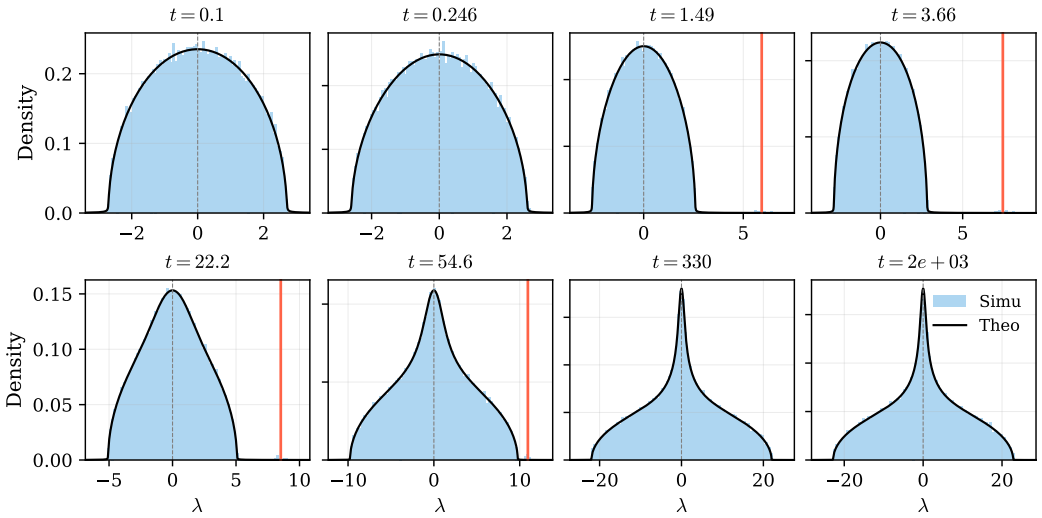

Figure 1: Empirical spectral density of S_t (blue histogram, 20 realizations, $N = 500$) versus the theoretical density from the 2×2 Dyson equation (13)–(14) (black curve) at eight training times. Parameters: $\lambda_+ = 1$, $\lambda_- = 0.1$, $\gamma = 1$, and $\alpha = 0.5$. In the signal case $\theta = 6$, the red vertical line marks the outlier location as predicted in Section 4 when an outlier is present. One clearly sees how the outlier first emerges from the bulk before being reabsorbed.

4 Rank-two transient BBP transition

4.1 Outlier condition for a rank-one teacher

We now analyze the effect of the rank-one teacher on S_t by considering

$$S_t^\theta = S_t + \theta\big(\widetilde{v} w_t^\top + w_t \widetilde{v}^\top\big), \qquad w_t = (I - e^{-t\Lambda^2})\widetilde{v}.$$

Because w_t is proportional to $\widetilde{v}$ only in the isotropic case $\lambda_- = 1$, the perturbation is genuinely rank two at finite time, and the standard rank-one BBP formula [7] does not apply directly. We can however reduce the problem to a 2×2 determinantal condition.

Proposition 4.1 (Outlier condition)**.** *Let $L_-(t)$ and $L_+(t)$ denote the lower and upper edges of the limiting bulk spectrum of S_t. For $\theta \geq 0$, the perturbed matrix S_t^θ has an outlier eigenvalue $\xi_{\theta,t}$ if and only if it solves*

$$\big(1 - \theta\, \psi_t(\xi_{\theta,t})\big)^2 = \theta^2\, \phi_t(\xi_{\theta,t})\chi_t(\xi_{\theta,t}), \qquad \xi_{\theta,t} \in \mathbb{R} \setminus [L_-(t), L_+(t)], \tag{15}$$

where the deterministic functions ϕ_t, ψ_t, χ_t are defined in Appendix C, Eq. (21).

Proposition 4.1 is proved in Appendix C.

Remark 4.2 (Isotropic limit)**.** *In the isotropic case $\lambda_- = 1$, one has $w_t = f_t \widetilde{v}$ with $f_t = 1 - e^{-t}$. Hence $\psi_t(z) = f_t \phi_t(z)$ and $\chi_t(z) = f_t^2 \phi_t(z)$, so (15) reduces to*

$$1 - 2\theta f_t\, \phi_t(\xi_{\theta,t}) = 0,$$

which is exactly the standard rank-one BBP equation with effective spike strength $2\theta f_t$.

Remark 4.3 (Boundary times)**.** *At $t = 0$, one has $w_t = 0$, hence $\psi_0 \equiv \chi_0 \equiv 0$, and (15) has no solution outside the bulk. At the other extreme, as $t \to \infty$, one has $w_t \to \widetilde{v}$, so*

$\phi_t(z), \psi_t(z), \chi_t(z) \to m_\infty(z)$ *and* (15) *reduces to*

$$1 - 2\theta\, m_\infty(\xi_{\theta,\infty}) = 0,$$

which is the usual steady-state spike equation for the limiting matrix S_∞^θ.

4.2 BBP thresholds and signal regimes

The upper outlier first appears when (15) acquires a solution at the upper edge $L_+(t)$. The corresponding critical signal level is

$$\theta_c(t) := \frac{1}{\psi_t(L_+(t)^+) + \sqrt{\phi_t(L_+(t)^+)\chi_t(L_+(t)^+)}}. \tag{16}$$

For fixed signal strength θ, define the fitting set

$$\mathcal{T}_\theta := \{t > 0 :\ \theta > \theta_c(t)\}.$$

From our observations, we can state the following conjecture.

Conjecture 4.1 (Three signal regimes)**.** *In the limit* $N, M \to \infty$ *with* $M/N \to \gamma > 0$*, exactly one of the following scenarios occurs:*

1. ***Weak-signal regime:*** $\mathcal{T}_\theta = \varnothing$*. No outlier ever separates from the bulk, and the teacher direction is never spectrally detectable.*
2. ***Strong-signal regime:*** $\mathcal{T}_\theta = (t_1, \infty)$ *for some* $t_1 > 0$*. An outlier emerges at time* t_1 *and remains separated from the bulk thereafter.*
3. ***Early-stopping regime:*** $\mathcal{T}_\theta = (t_1, t_2)$ *for some* $0 < t_1 < t_2 < \infty$*. An outlier exists only on a finite time window. This is the transient BBP regime associated with early stopping.*

In the early-stopping regime, t_1 and t_2 are interpreted as follows: for $t < t_1$ the model is underfitted and the teacher direction cannot be determined; for $t_1 < t < t_2$ the teacher is spectrally detectable in the sense that the eigenvector $u^\star$ corresponding to the spike has a computable, non-zero overlap $(u^\star \cdot v)^2$ with the teacher direction (see e.g. [34], section 14.2.2); and for $t > t_2$ the outlier is reabsorbed into the bulk, corresponding to a late-time overfitting phase. The optimal stopping time, defined as the $t_{\text{opt}} \in (t_1, t_2)$ that maximizes $(u^\star \cdot v)^2$, is discussed in Section 4.4 below.

We illustrate this behaviour by plotting the evolution in time of $\theta_c(t)$ in Figure 2. When $\lambda_+ = \lambda_- = 1$, the function is monotone, so that only regimes 1 and 2 are possible depending on the value of θ. On the contrary, in the anisotropic regime $\lambda_- = 0.1$, the function is no longer monotone: θ_c first decreases (the fast block reveals the teacher) and then increases again (the slow block, amplified by Λ^{-1}, injects enough noise to hide the signal). This means that, for an intermediate range of values of θ, a spike is visible only during a finite time interval – this is the early-stopping regime, and the minimum of θ_c is the best detection threshold achievable along the whole training trajectory.

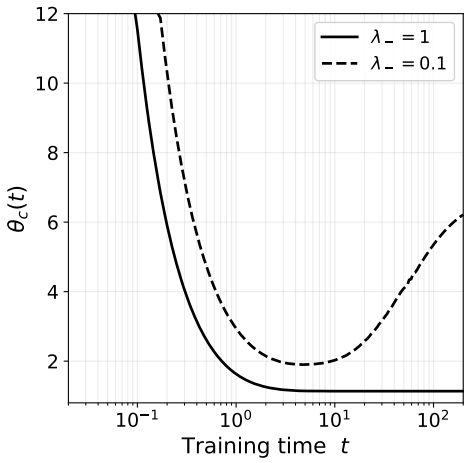

Figure 2: Evolution of θ_c defined in (16) for $\alpha = 0.5$ and two values of λ_-.

Remark 4.4. *Conjecture 4.1 is hard to prove in full generality due to the rather complex functional form of θ_c. Indeed, a proof would require showing that either the function is monotonically decreasing, or that it is quasi-convex (for instance if it has a unique global minimizer which is also the unique critical point), for whatever range of parameters.*

We can however state some partial results. For instance, we show in Appendix E that when $\lambda_- \to 1$ (i.e. in the isotropic limit), the function θ_c becomes monotonically decreasing. This entails that the isotropic regime leads to either weak or strong signal regime depending on the value of θ, as confirmed by Figure 2.

4.3 Phase diagrams and the role of anisotropy

The curve $\theta = \theta_c(t)$ in the (θ, t) plane directly separates the three regimes of Conjecture 4.1. Horizontal cuts at fixed θ give the fitting set $\mathcal{T}_\theta$, and the minimum of the boundary curve corresponds to the minimum of $\theta_c(t)$ visible in Figure 3. As the covariance becomes more anisotropic (smaller λ_-), the transient regime broadens; as the signal becomes stronger, the fitting window opens earlier and may become persistent.

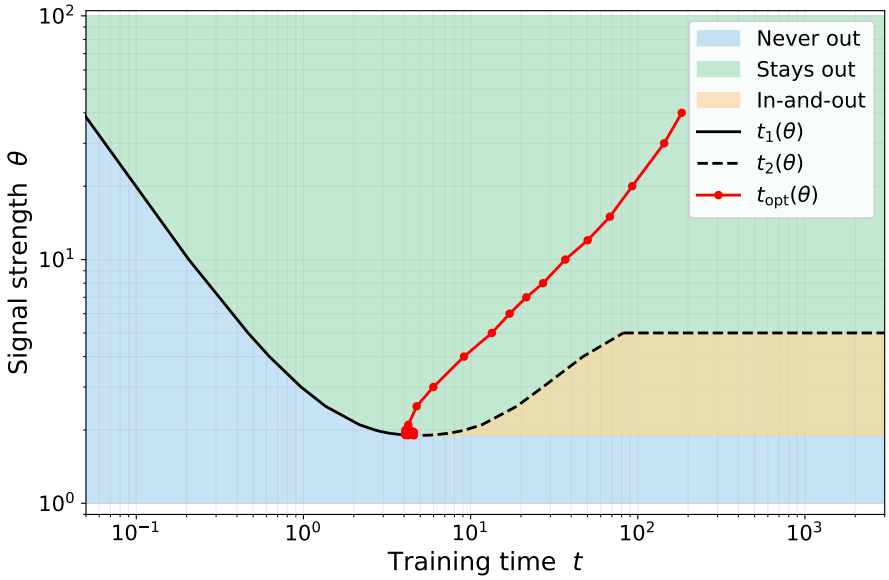

Figure 3: Phase diagram in the (θ, t) plane for $\lambda_- = 0.1$, $\gamma = 1$ and $\alpha = 0.5$. The critical curve $\theta = \theta_c(t)$ separates three regimes: no outlier (underfitting), transient outlier (early-stopping window), and persistent outlier. For a fixed intermediate value of θ, the outlier emerges at t_1 and is reabsorbed at t_2. The overlaid red curve shows the optimal stopping time $t_{\text{opt}}(\theta)$ defined in Section 4.4: in our experiments it lies strictly inside the transient window (t_1, t_2) whenever the latter exists.

To examine how anisotropy controls the possible learning regimes, we scan the (θ, λ_-) plane and, for each pair, count the number of sign changes of

$$F(\theta, t) := \left(1 - \theta\psi_t(L_+(t)^+)\right)^2 - \theta^2\phi_t(L_+(t)^+)\chi_t(L_+(t)^+)$$

over $t \in [0.05, 3000]$. One can derive this function through a simple manipulation of the outlier condition (15): a sign change of F shows by continuity that zero was crossed, hence (15) had a solution.

Figure 4(a) shows the resulting (θ, λ_-) diagram. The orange "in-and-out" wedge, the regime in which early stopping is the only way to recover the teacher, opens only below a critical anisotropy $\lambda_-^\star$: for nearly isotropic inputs only the weak- and strong-signal regimes survive. Even for strongly anisotropic inputs, the wedge is bounded below in θ: below a minimum signal strength, no stopping time recovers the teacher. Panel (b) repeats the construction in the (θ, γ) plane, using the 3×3 Dyson system of Appendix B for $\gamma < 1$. Fewer data samples widen the bulk, through the frozen null-space noise and through noisier least-squares directions, so all boundaries move to larger θ:

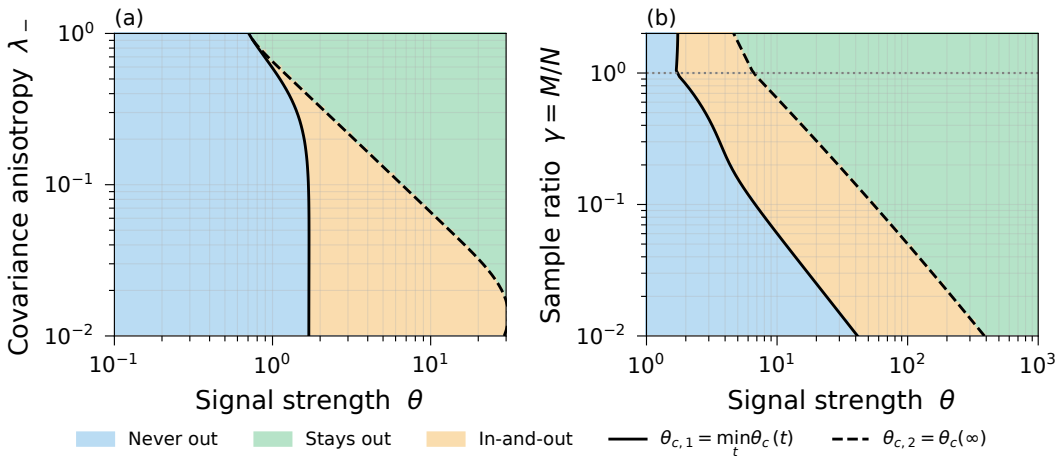

Figure 4: Outlier-regime classification obtained from $\theta_c(t)$, $t \in [0.05, 3000]$ ($\alpha = 0.5$, $\lambda_+ = 1$). (a) (θ, λ_-) plane at $\gamma = 1$. (b) (θ, γ) plane at $\lambda_- = 0.1$; dotted line: $\gamma = 1$, below which a frozen null-space block appears (Appendix B). Blue: no outlier; green: persistent outlier; orange: transient outlier (early-stopping regime).

the best detection threshold grows from 1.69 at $\gamma = 1$ to about 41 at $\gamma = 0.01$, and at fixed θ the early-stopping window narrows and eventually closes. Note that, in the in-and-out case, we could also estimate the time interval over which the signal is detected.

4.4 Teacher recovery and optimal stopping

The regime classification of Conjecture 4.1 tells us *when* the teacher is spectrally detectable, but not *how well* it is reconstructed at any given time. The natural scalar summary of teacher recovery is the squared overlap

$$q_t := (u_t^\star \cdot v)^2$$

between the outlier eigenvector $u_t^\star$ and the true teacher direction v, with the convention $q_t = 0$ when no outlier exists. This is the quantity that controls the rank-one content of the trained weight matrix along the teacher direction.

Because the outlier is a simple real zero of the 2×2 Schur complement $M_t(z)$ introduced in Appendix C, the overlap q_t can be obtained, via the Woodbury identity, as a residue of $v^\top (zI - S_t^\theta)^{-1} v$ at $z = \xi_{\theta,t}$. This residue is a rational function of the deterministic quantities ϕ_t, ψ_t, χ_t and their derivatives at $\xi_{\theta,t}$; we sketch the derivation in Appendix C and do not expand the expression here, since it is not used in what follows. In the isotropic limit $\lambda_- = 1$ it collapses to the classical Benaych-Georges–Nadakuditi formula for rank-one BBP. What matters for the discussion below is a qualitative feature of q_t: it is continuous on the fitting set $\mathcal{T}_\theta$, vanishes on its boundary (as the outlier merges into the bulk edge and the eigenvector delocalizes over the bulk), and is strictly positive in its interior. In the early-stopping regime $\mathcal{T}_\theta = (t_1, t_2)$, q_t therefore attains at least one interior maximum, which defines an *optimal early-stopping time*

$$t_{\mathrm{opt}}(\theta) \in \operatorname*{arg\,max}_{t \in (t_1, t_2)} q_t.$$

Numerical evidence We computed $t_{\mathrm{opt}}(\theta)$ by one-dimensional maximization of q_t over the window $(t_1(\theta), t_2(\theta))$ already located by the phase-diagram computation, for the full range of θ covered by Figure 3. The result is overlaid as the red curve in Figure 3. In our experiments q_t is unimodal on (t_1, t_2), so $t_{\mathrm{opt}}(\theta)$ is a single point, and it lies strictly inside the transient window for every θ in the early-stopping regime. The curve interpolates smoothly into the strong-signal regime, tracing the locus of best teacher recovery along the entire (θ, t) plane. An optimal stopping time also exists for a full-rank power-law covariance (Section 4.5, Figure 5). A rigorous decomposition of the test error in terms of q_t is left to future work.

4.5 Beyond the solvable setting

Nonlinear architectures To test whether the transient BBP survives nonlinearity and depth, we trained two models on the same data and teacher (Appendix G) and tracked the spectrum of each symmetrized weight matrix. In a two-layer network $\widehat{Y} = W_2 \tanh(W_1 X)$, the scenario of Figure 1 is reproduced in both layers, and synchronously: an outlier aligned with the teacher leaves the bulk of W_1 and of W_2 at the same time, and both are later reabsorbed as the bulk broadens (Figure 6). In a one-layer self-attention model $\widehat{Y} = (W_V X)\,\mathrm{softmax}\big((W_K X)^\top (W_Q X)\big)$, only the value matrix W_V, which enters the output linearly, shows the transient BBP (Figure 7). The query and key matrices, which act through the softmax, do not: W_Q carries only a weak signal that decays slowly, and the outliers of W_K are orthogonal to the teacher, a reminder that a spectral outlier alone does not certify that the teacher has been learned.

Beyond two blocks Our theory relies on a two-block covariance, which gives the dynamics only two time scales. To test whether the transient BBP survives beyond this simplification, we simulated gradient flow with a full-rank covariance whose eigenvalues follow a power law $p(\lambda) \propto \lambda^{-1.5}$ on $[0.1, 5]$ (Appendix F), so that directions are learned over a continuum of time scales. No closed-form prediction is available here, yet the three regimes of Conjecture 4.1 persist (Figure 5). A strong signal ($\theta = 10$) is recovered and stays detectable, while a weak one ($\theta \lesssim 1$, not shown) is never detected. For intermediate signals ($\theta = 2, 3$), the teacher is recovered only during a finite window: as in Section 4.4, the overlap q_t reaches an interior maximum, then returns to zero as the slow directions fit the noise. The transient BBP is therefore set by the spread of the covariance spectrum, i.e. its condition number, not by its two-block form.

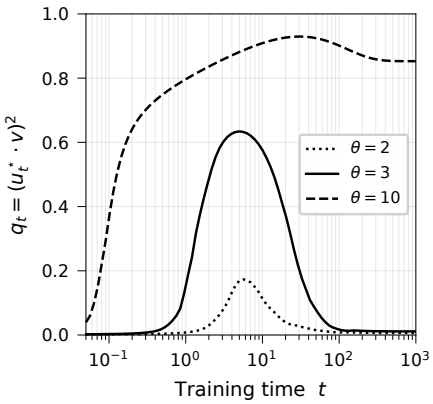

Figure 5: Teacher overlap q_t for a power-law spectrum ($\lambda \in [0.1, 5]$, $N = 400$).

5 Conclusion and Extensions

We introduced a solvable random-matrix model in which the spectrum of a symmetrized weight matrix undergoes a *transient BBP transition* along gradient flow: depending on signal strength and covariance anisotropy, the teacher spike never emerges, emerges and persists, or exists only during a finite time window, in which the teacher overlap peaks at an optimal stopping time. The driving ingredient is anisotropy of the input covariance. Fast directions reveal the teacher early, whereas slow directions keep fitting noise and eventually hide the signal again, so overfitting is not an abstract late-time pathology here but a concrete consequence of spectral heterogeneity in the data.

Connection to empirical observations Our model qualitatively reproduces features reported in empirical spectral studies of neural-network training [2, 3, 4, 5, 16]: isolated eigenvalues at intermediate training times, a broadening bulk at late times, and a best generalization time tied to the transient rather than the asymptotic spectrum. It does *not* capture the heavy-tailed bulk and the multi-layer nonlinear coupling of deep networks at convergence.

Finally, for infinitely wide networks the Neural Tangent Kernel [35, 36] plays the role of $XX^\top$, so the transient BBP may be a useful toy model for early stopping in more complex architectures.

References

[1] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[2] 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.

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

[4] Max Staats, Matthias Thamm, and Bernd Rosenow. Boundary between noise and information applied to filtering neural network weight matrices. *Physical Review E*, 108(2):L022302, 2023.

[5] David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R Walter. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*, 2024.

[6] Samuel F Edwards and Raymund C Jones. The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595–1603, 1976.

[7] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[8] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

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

[10] Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1370–1378. PMLR, 2019.

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

[12] Taishi Watanabe, Ryo Karakida, and Jun-nosuke Teramae. The impact of anisotropic covariance structure on the training dynamics and generalization error of linear networks. *arXiv preprint arXiv:2601.06961*, 2026.

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

[14] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214(1):1–90, 2025.

[15] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024.

[16] Brian Richard Olsen, Sam Fatehmanesh, Frank Xiao, Adarsh Kumarappan, and Anirudh Gajula. From sgd to spectra: A theory of neural network weight dynamics. *arXiv preprint arXiv:2507.12709*, 2025.

[17] Gert Aarts, Biagio Lucini, and Chanju Park. Stochastic weight matrix dynamics during learning and Dyson Brownian motion. *Physical Review E*, 111:015303, 2025.

[18] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[19] 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. In *International Conference on Machine Learning*, pages 36106–36159. PMLR, 2024.

[20] 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. In *International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR, 2025.

[21] 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. In *International Conference on Learning Representations*, 2026.

[22] Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. Optimal scaling laws in learning hierarchical multi-index models. In *Advances in Neural Information Processing Systems*, 2026.

[23] Gérard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. Spectral alignment of stochastic gradient descent for high-dimensional classification tasks. *The Annals of Applied Probability*, 35(4):2767–2822, 2025.

[24] Gérard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. Local geometry of high-dimensional mixture models: Effective spectral theory and dynamical transitions. *The Annals of Statistics*, 54(4):1682–1711, 2026.

[25] Tony Bonnaire, Giulio Biroli, and Chiara Cammarota. The role of the time-dependent hessian in high-dimensional optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):083401, 2025.

[26] Andrea Montanari and Zihao Wang. Phase transitions for feature learning in neural networks. *arXiv preprint arXiv:2602.01434*, 2026.

[27] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[28] Oskari H Ajanki, László Erdős, and Torben Krüger. Universality for general Wigner-type matrices. *Probability Theory and Related Fields*, 169(3):667–727, 2017.

[29] Oskari H Ajanki, László Erdős, and Torben Krüger. Stability of the matrix Dyson equation and random matrices with correlations. *Probability Theory and Related Fields*, 173(1):293–373, 2019.

[30] Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

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

[32] Blake Bordelon and Cengiz Pehlevan. Disordered dynamics in high dimensions: Connections to random matrices and machine learning. *arXiv preprint arXiv:2601.01010*, 2026.

[33] Zhenyu Liao and Michael W. Mahoney. Random matrix theory for deep learning: Beyond eigenvalues of linear models. *IEEE Signal Processing Magazine*, 2025. To appear.

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

[35] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, 2018.

[36] Lechao Xiao, Jeffrey Pennington, and Samuel S Schoenholz. Disentangling trainability and generalization in deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10462–10472. PMLR, 2020.

A Proof of Proposition 3.1

Proof. We fix a time $t \geq 0$ and drop all time indexes in the proof for notational simplicity. Write

$$r_i := (D_1)_{ii}, \qquad q_i := (D_2)_{ii},$$

so that $r_i \in \{a, b\}$ and $q_i \in \{c, d\}$ depending on whether i belongs to block A or block B. From (11), for $i < j$ we have

$$S_{ij} = (r_i + r_j)W_{ij} + q_j\Gamma_{ij} + q_i\Gamma_{ji}.$$

Since the families $\{W_{ij}\}_{i\leq j}$ and $\{\Gamma_{ij}\}_{i,j}$ are independent and centered, the upper-triangular entries of S are independent and centered as well. Moreover,

$$N\,\mathbb{E}S_{ij}^2 = (r_i + r_j)^2 + \frac{N}{M}\,(q_i^2 + q_j^2) \;\longrightarrow\; (r_i + r_j)^2 + \frac{q_i^2 + q_j^2}{\gamma}.$$

Hence the variance profile is piecewise constant, and for off-diagonal entries it is exactly

$$\sigma_{AA} = 4a^2 + \frac{2c^2}{\gamma}, \qquad \sigma_{BB} = 4b^2 + \frac{2d^2}{\gamma}, \qquad \sigma_{AB} = (a + b)^2 + \frac{c^2 + d^2}{\gamma},$$

as claimed in (12). The diagonal entries also have variance of order N^{-1} and fit into the same Wigner-type framework.

Now consider the standard resolvent

$$G(z) := (S - zI)^{-1}, \qquad z \in \mathbb{C}^+.$$

By the general theory of Wigner-type matrices with variance profile $\mathfrak{s}_{ij} := \mathbb{E}|S_{ij}|^2$, the diagonal resolvent entries converge to deterministic limits $\widehat{m}_i(z)$ solving the quadratic vector equation

$$-\frac{1}{\widehat{m}_i(z)} = z + \sum_{j=1}^{N} \mathfrak{s}_{ij}\,\widehat{m}_j(z),$$

with $\Im\widehat{m}_i(z) > 0$ for $\Im z > 0$ [29, 28].

Since the variance profile has only two block values, the solution is constant on each block:

$$\widehat{m}_i(z) = \widehat{m}_A(z) \quad (i \in A), \qquad \widehat{m}_i(z) = \widehat{m}_B(z) \quad (i \in B).$$

Therefore the quadratic vector equation reduces to

$$-\frac{1}{\widehat{m}_A(z)} = z + \alpha\,\sigma_{AA}\,\widehat{m}_A(z) + (1 - \alpha)\,\sigma_{AB}\,\widehat{m}_B(z),$$
$$-\frac{1}{\widehat{m}_B(z)} = z + \alpha\,\sigma_{AB}\,\widehat{m}_A(z) + (1 - \alpha)\,\sigma_{BB}\,\widehat{m}_B(z).$$

Finally, Proposition 3.1 is written in terms of the resolvent

$$(zI - S)^{-1} = -\,(S - zI)^{-1} = -G(z).$$

Thus, setting

$$m_A(z) := -\widehat{m}_A(z), \qquad m_B(z) := -\widehat{m}_B(z),$$

we obtain exactly

$$\frac{1}{m_A(z)} = z - \alpha\,\sigma_{AA}\,m_A(z) - (1 - \alpha)\,\sigma_{AB}\,m_B(z),$$
$$\frac{1}{m_B(z)} = z - \alpha\,\sigma_{AB}\,m_A(z) - (1 - \alpha)\,\sigma_{BB}\,m_B(z).$$

Moreover $\Im m_A(z), \Im m_B(z) < 0$ for $\Im z > 0$, and $m_A(z), m_B(z) \sim 1/z$ as $|z| \to \infty$. The averaged limit

$$m(z) = \alpha\,m_A(z) + (1 - \alpha)\,m_B(z)$$

is therefore the Stieltjes transform of the limiting empirical spectral measure in the convention of (14), and the density is recovered by the inversion formula

$$\rho_S(\lambda) = -\frac{1}{\pi}\lim_{\eta\downarrow 0}\Im\, m(\lambda + i\eta).$$

This proves the proposition. □

B The case $\gamma < 1$: A 3×3 Dyson system

When $M < N$ (i.e. $\gamma = M/N < 1$), the input covariance $XX^\top$ has a null space of dimension $(1-\gamma)N$. In the null space, the gradient flow never moves: the exponential $e^{-tXX^\top}$ acts as the identity, and the noise term $ZX^\top(XX^\top)^{-1}(I - e^{-tXX^\top})$ vanishes (since $X^\top$ annihilates the null space). Consequently, the weight matrix retains its random initialisation in those directions for all time.

This introduces a third block in the variance profile of the symmetrised noise matrix, leading to a 3×3 Dyson system that generalises the 2×2 system of Section 3. We also drop the time index for simplicity.

B.1 Three-block structure

We split the N indices into three classes:

Block	Fraction	a_p	c_p
A (fast)	$\alpha_A = \gamma/2$	e^{-t}	$1 - e^{-t}$
B (slow)	$\alpha_B = \gamma/2$	$e^{-\lambda_-^2 t}$	$(1 - e^{-\lambda_-^2 t})/\lambda_-$
C (null)	$\alpha_C = 1 - \gamma$	1	0

The null-space block C has $a_C = 1$ (the initialisation is never damped) and $c_C = 0$ (no noise is injected through X).

Using the general variance formula $\sigma_{pq} = (a_p + a_q)^2 + (c_p^2 + c_q^2)/\gamma$, we obtain six variance parameters:

$$\begin{aligned}
\sigma_{AA} &= 4a_A^2 + \frac{2c_A^2}{\gamma}, \\
\sigma_{BB} &= 4a_B^2 + \frac{2c_B^2}{\gamma}, \\
\sigma_{CC} &= 4, \\
\sigma_{AB} &= (a_A + a_B)^2 + \frac{c_A^2 + c_B^2}{\gamma}, \\
\sigma_{AC} &= (1 + a_A)^2 + \frac{c_A^2}{\gamma}, \\
\sigma_{BC} &= (1 + a_B)^2 + \frac{c_B^2}{\gamma}.
\end{aligned} \tag{17}$$

Note that $\sigma_{CC} = 4$ is time-independent: the null-space block is a frozen GOE contribution at all times.

B.2 The 3×3 Dyson equation

The partial Stieltjes transforms $m_A(z)$, $m_B(z)$, $m_C(z)$ satisfy:

$$\begin{cases}
\dfrac{1}{m_A(z)} = z - \dfrac{\gamma}{2}\,\sigma_{AA}\,m_A - \dfrac{\gamma}{2}\,\sigma_{AB}\,m_B - (1-\gamma)\,\sigma_{AC}\,m_C, \\
\dfrac{1}{m_B(z)} = z - \dfrac{\gamma}{2}\,\sigma_{AB}\,m_A - \dfrac{\gamma}{2}\,\sigma_{BB}\,m_B - (1-\gamma)\,\sigma_{BC}\,m_C, \\
\dfrac{1}{m_C(z)} = z - \dfrac{\gamma}{2}\,\sigma_{AC}\,m_A - \dfrac{\gamma}{2}\,\sigma_{BC}\,m_B - (1-\gamma)\,\sigma_{CC}\,m_C,
\end{cases} \tag{18}$$

with the Stieltjes-branch conditions $\Im\, m_p(z) < 0$ for $\Im\, z > 0$ and $m_p(z) \sim 1/z$ as $|z| \to \infty$.

The global Stieltjes transform and spectral density are

$$m(z) = \frac{\gamma}{2}\, m_A(z) + \frac{\gamma}{2}\, m_B(z) + (1-\gamma)\, m_C(z), \qquad \rho_S(\lambda) = -\frac{1}{\pi} \lim_{\eta \downarrow 0} \Im\, m(\lambda + i\eta). \tag{19}$$

B.3 Effect of the null space on the transient BBP window

Compared to the oversampled regime $\gamma > 1$, the null-space block C introduces a frozen GOE component of variance $\sigma_{CC} = 4$ that competes with the signal at all times. This has two consequences:

1. The bulk spectrum is generically wider (since the frozen initialisation noise never decays), making it harder for an outlier to emerge.
2. The null-space directions do not contribute to ψ or χ (since they carry no learned signal), but they do contribute to ϕ (through m_C), which enters the right-hand side of the outlier equation.

As a result, the transient BBP window $[t_1, t_2]$ is generically *narrower* for $\gamma < 1$ than for $\gamma > 1$: more directions carry permanent noise without contributing to learning, raising the effective detection threshold $\theta_c(t)$. Figure 4(b) in Section 4.3 quantifies this effect over $\gamma \in [0.01, 2]$.

C Outlier equation and BBP transition

C.1 Weighted finite-rank teacher perturbation

In the covariance eigenbasis, the signal contribution to the symmetrized matrix is

$$\theta\big(\widetilde{v} w_t^\top + w_t \widetilde{v}^\top\big), \qquad w_t = (I - e^{-t\Lambda^2})\widetilde{v}.$$

This perturbation is rank two unless w_t is proportional to $\widetilde{v}$, which happens only in the isotropic case. Thus the standard rank-one BBP formula does not directly apply at finite time [7].

C.2 Rank-two determinant and outlier condition

To simplify notation, we drop the tildes and the time subscript and write the perturbed matrix as

$$S^\theta = S + \theta(wv^\top + vw^\top).$$

Introduce

$$U = [w, v] \in \mathbb{R}^{N\times 2}, \qquad V = [v, w] \in \mathbb{R}^{N\times 2},$$

and let

$$R(z) = (zI - S)^{-1}.$$

By the matrix determinant lemma,

$$\det(zI - S^\theta) = \det(zI - S)\, \det\big(I_2 - \theta V^\top R(z) U\big).$$

The 2×2 Schur complement is

$$M(z) = I_2 - \theta \begin{pmatrix} v^\top R(z) w & v^\top R(z) v \\ w^\top R(z) w & w^\top R(z) v \end{pmatrix} = \begin{pmatrix} 1 - \theta\psi(z) & -\theta\phi_t(z) \\ -\theta\chi(z) & 1 - \theta\psi(z) \end{pmatrix},$$

where

$$\phi_t(z) = v^\top R(z) v, \qquad \psi_t(z) = v^\top R(z) w, \qquad \chi_t(z) = w^\top R(z) w. \tag{20}$$

Concentration of quadratic forms Since v is asymptotically isotropic in the eigenbasis of the noise matrix and the two blocks have deterministic weights, the quadratic forms concentrate according to

$$\begin{aligned} \phi_t(z) &\xrightarrow[N\to\infty]{} \alpha\, m_{A,t}(z) + (1-\alpha)\, m_{B,t}(z), \\ \psi_t(z) &\xrightarrow[N\to\infty]{} \alpha\, f_{A,t}\, m_{A,t}(z) + (1-\alpha)\, f_{B,t}\, m_{B,t}(z), \\ \chi_t(z) &\xrightarrow[N\to\infty]{} \alpha\, f_{A,t}^2\, m_{A,t}(z) + (1-\alpha)\, f_{B,t}^2\, m_{B,t}(z), \end{aligned} \tag{21}$$

with

$$f_{A,t} = 1 - e^{-\lambda_+^2 t}, \qquad f_{B,t} = 1 - e^{-\lambda_-^2 t}.$$

Outlier equation Eigenvalues of S^θ outside the bulk correspond to real zeros of $\det M(z)$. Computing the determinant gives

$$\det M(z) = \big(1 - \theta\psi_t(z)\big)^2 - \theta^2\phi_t(z)\chi_t(z) = 0, \tag{22}$$

which is the outlier equation used in Proposition 4.1.

Eigenvector overlap with the teacher Given a simple outlier root ξ_θ of (22), the projections of the associated eigenvector $u^\star$ of S^θ onto v and w concentrate to deterministic limits, while the component orthogonal to $\mathrm{span}(v, w)$ is delocalized over the bulk. The squared overlap $q_t = (u^\star \cdot v)^2$ with the teacher direction is obtained, via the Woodbury identity, as the residue of the resolvent $v^\top(zI - S^\theta)^{-1}v$ at $z = \xi_\theta$; this reduces to differentiating $\det M(z)$ at ξ_θ and yields a rational function of ϕ_t, ψ_t, χ_t and their derivatives at ξ_θ. We omit the explicit expression, which is not used in the main text beyond the qualitative properties of q_t discussed in Section 4.4.

D Fixed-point algorithm for the Dyson equation

Proposition 3.1 characterizes $(m_A(z), m_B(z))$ as the unique physical solution of (13) with $\Im m_A(z), \Im m_B(z) < 0$ for $\Im z > 0$. Numerically, we solve this system by fixed-point iteration, initializing $m_A^{(0)} = m_B^{(0)} = 1/z$ and iterating

$$m_A^{(n+1)} = \frac{1}{z - \alpha\sigma_{AA}m_A^{(n)} - (1-\alpha)\sigma_{AB}m_B^{(n)}}, \qquad m_B^{(n+1)} = \frac{1}{z - \alpha\sigma_{AB}m_A^{(n)} - (1-\alpha)\sigma_{BB}m_B^{(n)}}. \tag{23}$$

Numerical implementation For $z = x + i\varepsilon$ with $\varepsilon > 0$, the iteration preserves the physical branch and converges robustly in our experiments. In practice we use $\varepsilon = 10^{-2}$ and stop when the ℓ^∞ change between successive iterates is below 10^{-10}. The Dyson-equation solutions, the linear simulations ($N \le 500$, up to 30 realizations) and the phase-diagram scans run on a standard CPU; the nonlinear experiments of Appendix G use a single GPU.

Bulk edge and phase boundary The upper edge L_+ is estimated from the rightmost point on a fine real grid where the numerical density exceeds a small threshold. For each fixed θ, we then evaluate

$$F(\theta) = \big(1 - \theta\psi(L_+^+)\big)^2 - \theta^2\phi(L_+^+)\chi(L_+^+)$$

Doing this on a coarse grid in t and refining each sign change by bisection yields the transition times and the phase diagrams shown in the main text.

E Isotropic expansion of θ_c

When $\lambda_- \to 1$, we can show via standard Taylor expansion that

$$\theta_c(t) = \theta_{\mathrm{iso}}(t)\Big[1 + (1-\alpha)(1-\lambda_-)\,G(t)\Big] + O\big((1-\lambda_-)^2\big),$$

where

$$\theta_{\mathrm{iso}}(t) = \frac{\sqrt{(e^t-1)^2 + 2\gamma}}{\sqrt{2\gamma}\,(e^t-1)}, \qquad G(t) = \frac{(e^t-1)^3 + 4\gamma\, t\, e^t}{(e^t-1)\big[(e^t-1)^2 + 2\gamma\big]}.$$

Simple differentiation shows that θ_{iso} is decreasing monotonically. On the other hand, a standard asymptotic calculus leads to

$$\theta_c(t) \underset{t\to 0}{\sim} \frac{1 + 2(1-\alpha)(1-\lambda_-)}{t}, \qquad \theta_c(t) \xrightarrow[t\to\infty]{} \frac{1 + (1-\alpha)(1-\lambda_-)}{\sqrt{2\gamma}}, \tag{24}$$

entailing that G grows at a similar scale as θ_{iso} around zero. This shows that θ_c is indeed monotonically decreasing from infinity to the limiting value on the right hand side of (24), in accordance with our expectations and Figure 2.

F Full-rank power-law spectrum: numerical validation

The two-block covariance model of Section 2.3 is analytically tractable because it reduces the Dyson system to two equations. Here we check numerically that the transient BBP phenomenology is not an artefact of that simplification, by replacing the two-block spectrum with a continuous power-law distribution.

Setup We draw the N singular values of the input matrix X independently from the truncated power-law density

$$p(\lambda) \propto \lambda^{-\beta}, \qquad \lambda \in [\lambda_{\min}, \lambda_{\max}],$$

with $\beta = 1.5$, $\lambda_{\min} = 0.1$, $\lambda_{\max} = 5$, and $N = 400$. The rest of the setup follows Section 2 exactly: the teacher is rank-one with signal amplitude θ, the initialisation is GOE at scale $1/\sqrt{N}$, and the noise matrix has i.i.d. $\mathcal{N}(0, 1/N)$ entries. Working in the eigenbasis of $XX^\top$ (with eigenvalues $\mu_i = \lambda_i^2$), the closed-form gradient-flow solution (2) becomes

$$A_t = Q\left(e^{-t\mu} \cdot G \; + \; \frac{1 - e^{-t\mu}}{\mu} \cdot W_Z\right),$$

where Q is the eigenvector matrix of $XX^\top$, $\mu = \mathrm{diag}(\mu_1, \ldots, \mu_N)$, $G = Q^\top A_{\text{init}}$, and $W_Z = Q^\top (XY_0^\top)$ encodes the teacher-plus-noise contribution. Simulations are run over 30 independent realisations at each of 85 logarithmically spaced times $t \in [0.05, 1000]$ and averaged.

Results Figure 5 (Section 4.5) shows the squared overlap $q_t = (u_t^\star \cdot v)^2$ between the top eigenvector of $S_t = A_t + A_t^\top$ and the teacher direction v, for $\theta = 2, 3$ (transient) and $\theta = 10$ (persistent).

The three qualitative regimes of Conjecture 4.1 are all visible: for $\theta \lesssim 1$ the overlap never exceeds the random baseline $1/N$ (weak-signal regime); for intermediate θ the overlap rises, attains an interior maximum, and decays back to baseline (early-stopping regime); and for large θ the overlap rises and remains elevated throughout training (strong-signal regime). The peak overlap increases with θ and the transient window widens, consistent with the phase-diagram picture of Section 4.3. These observations confirm that the transient BBP mechanism is robust to the specific form of the covariance spectrum and is not a peculiarity of the two-block simplification.

G Nonlinear architectures

Setup We use the data and teacher of the main text: $N = M = 500$, $\alpha = 0.5$, $\lambda_- = 0.1$, $\gamma = 1$, and $Y = \theta v v^\top X + Z$. Models are trained on $\frac{1}{2}\|\widehat{Y} - Y\|_F^2$ by full-batch gradient descent, with time $t = k\eta$ after k steps of size η. All square weights start from independent GOE matrices, and results are averaged over 30 realizations at $\theta = 15$. For every weight W we compute the spectrum of $W + W^\top$; an eigenvalue is declared an outlier when its mean over realizations exceeds the corresponding extreme eigenvalue obtained without signal ($\theta = 0$) by more than $\max(4\,\mathrm{std}, 0.15)$.

- **Two-layer network:** $\widehat{Y} = W_2 \tanh(W_1 X)$, $\eta = 0.015$, $t \le 200$.
- **One-layer self-attention:** $\widehat{Y} = (W_V X)\,\mathrm{softmax}\big(s\,(W_K X)^\top (W_Q X)\big)$, softmax over keys, with the samples as the attention set, $\eta = 0.01$, $t \le 300$. The temperature s is fixed at initialization so that the scores have standard deviation 0.5; with the usual $1/\sqrt{N}$ scaling the softmax stays essentially uniform and nothing is learned on this time scale.

Two-layer network Figure 6 is the analogue of Figure 1. In both layers a pair of outliers leaves the bulk at $t \approx 0.5$, is still present at $t = 7$, and is reabsorbed by $t = 20$ as the bulk broadens, while the bulk itself deforms differently in the two layers. At $t = 7$ the top eigenvector has overlap $(u^\star \cdot v)^2 \approx 0.34$ with the teacher for W_1 and 0.38 for W_2.

Self-attention The three weights behave differently (Figure 7). The value matrix W_V reproduces the transient behaviour: its outliers are present from $t = 2$ to $t = 50$ and reabsorbed by $t = 300$, with overlap ≈ 0.47 at $t = 10$. The query matrix W_Q carries a weaker signal (overlap ≈ 0.23) that decays slowly. The key matrix W_K also develops outliers, but their overlap with v stays at the random level $1/N$: a spectral outlier alone does not certify that the teacher has been learned.

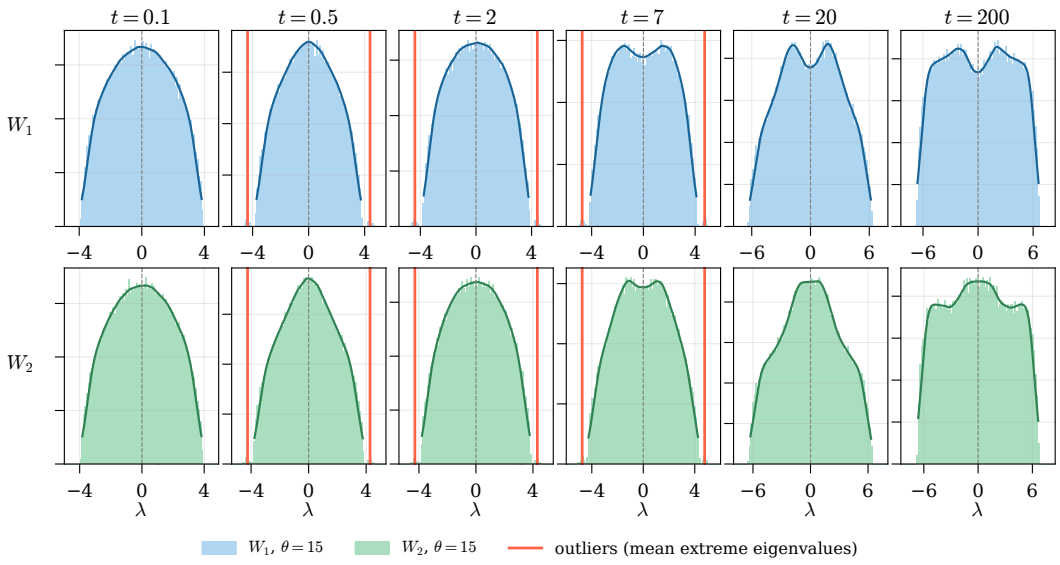

Figure 6: Two-layer $\tanh$ network, $\theta = 15$: spectral density of $W_\ell + W_\ell^\top$ for each layer (rows) at six training times (30 realizations). Red vertical lines: mean largest and smallest eigenvalues, shown when they separate from those obtained without signal ($\theta = 0$). As in Figure 1, the outliers emerge from the bulk and are then reabsorbed.

Why the outliers come in pairs For unit vectors a, r with $c = a \cdot r$, the symmetrization of a rank-one matrix $\sigma a r^\top$ satisfies $(ar^\top + ra^\top)(a \pm r) = (c \pm 1)(a \pm r)$: it has the two non-zero eigenvalues $\sigma(c+1)$ and $\sigma(c-1)$, with eigenvectors $(a \pm r)/\|a \pm r\|$. In the linear model the learned signal is $\theta\, v\,[(1 - e^{-t\Sigma})v]^\top$, whose two directions are close; the negative eigenvalue is small and stays inside the bulk, so a single outlier is visible. A layer inside a product instead learns a component whose hidden side is set by the random initialization of the other layers. Near initialization, gradient flow gives $\dot{W}_1 \supset \theta\,(W_2(0)^\top v)(\Sigma v)^\top$ and $\dot{W}_2 \supset \theta\, v\,(W_1(0)\Sigma v)^\top$. The hidden directions $W_2(0)^\top v$ and $W_1(0)\Sigma v$ are nearly orthogonal to v, so $c \approx 0$ and $W + W^\top$ shows a pair $\pm\sigma$, each eigenvector carrying about half of the teacher overlap; this is why the overlaps level off near $1/2$. The same holds for self-attention. Writing $\widehat{Y} = W_V X P$ with P the softmax matrix, gradient flow gives $\dot{W}_V \supset \theta\, v\,(XPX^\top v)^\top$: W_V writes along v but reads along $XPX^\top v$, which the softmax mixing of the samples makes nearly orthogonal to v (measured overlap 0.03). W_Q and W_K receive their gradient through the softmax, and their learned components are again non-symmetric: W_Q reads close to v (overlap 0.71) and writes a random direction, while neither side of the component learned by W_K is aligned with v. This explains the pairs in all three rows of Figure 7, and why the pair of W_K carries no teacher overlap. We checked this mechanism directly: initializing $W_2(0) = I$ makes the learned component of W_1 symmetric, and the negative outlier of W_1 disappears.

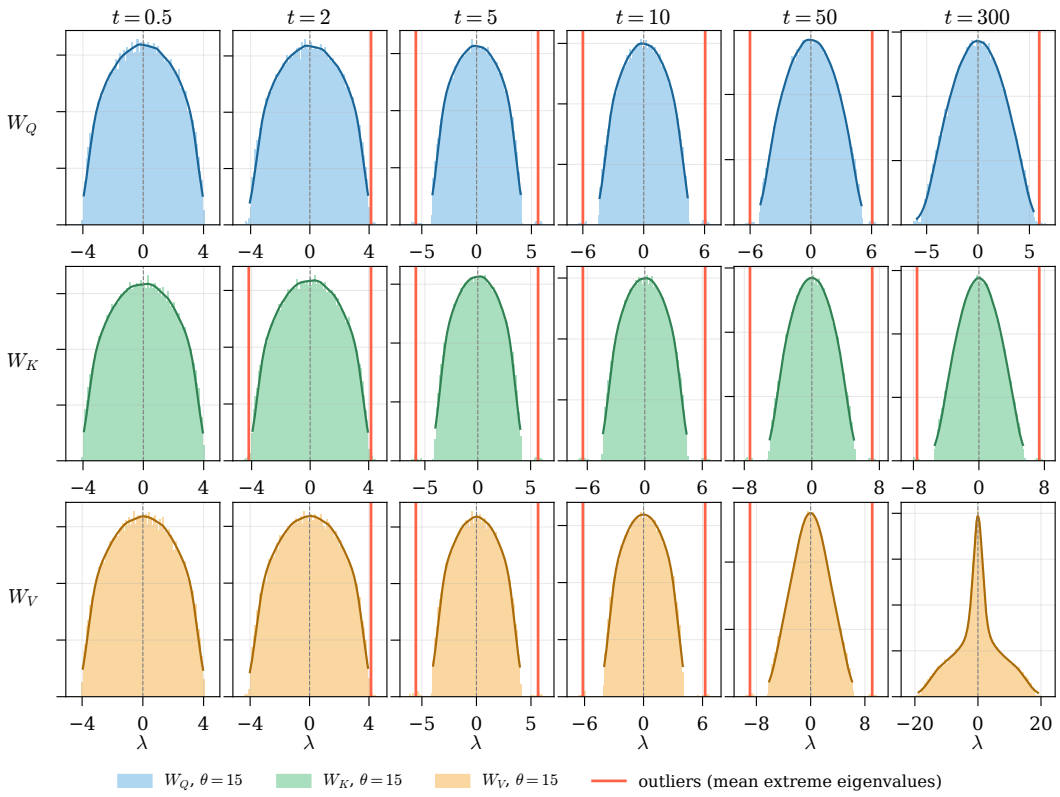

Figure 7: One-layer self-attention, $\theta = 15$: spectra of W_Q, W_K, W_V (rows) at six training times; conventions as in Figure 6. The outliers of W_V are transient. Those of W_K are orthogonal to the teacher.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe all main contributions: the 2×2 Dyson equation for the bulk spectrum, the rank-two outlier condition, the three-phase classification (weak-signal, strong-signal, early-stopping), and the phase diagrams. The scope is explicitly restricted to a linear teacher–student setting with a two-block covariance.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

Justification: The conclusion explicitly acknowledges the key limitations: the restriction to a linear teacher–student model, the use of additive rather than multiplicative symmetrization, and the inability to capture the heavy-tailed bulk and nonlinear coupling observed in deep networks at convergence.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All results are numbered and cross-referenced. The proof of Proposition 1 (Dyson equation) is given in Appendix A; the derivation of the rank-two outlier condition (Proposition 2) is in Appendix C. All assumptions (high-dimensional limit $M/N \to \gamma \geq 1$, two-block covariance, GOE initialization, rank-one teacher) are stated explicitly in Section 2.

   Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: All simulation parameters are reported in the main text (system size $N = 500$, aspect ratio $\gamma = 1$, block proportion $\alpha = 0.5$, singular values $\lambda_+ = 1$, $\lambda_- = 0.1$, 20 realizations, logarithmically spaced time grids). The fixed-point algorithm for the Dyson equation is fully described in Appendix D.

Guidelines:

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

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

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

Answer: [Yes]

Justification: The code is publicly available at `https://github.com/FlorentinCDX/rmt-early-stopping`. All data are synthetic and are generated by the code itself; no external dataset is used. The repository provides one script per figure of the paper (main text and appendices), a script that regenerates all figures, the list of dependencies, a quickstart notebook, and unit tests for the theoretical predictions, and its README explains how to run them. All simulation parameters and numerical algorithms are also described in the paper and appendices.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All experimental details are given in Sections 3.3 and 4.3 and Appendix D, including the system size, block fractions, covariance parameters, signal amplitude, number of independent realizations, and the time grids used for each figure.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The paper reports full empirical spectral distributions (histograms over 20 realizations in the main experiments, 8 in the power-law appendix), which provide richer statistical information than scalar error bars. Agreement with the theoretical prediction is assessed distribution-by-distribution, which is the standard validation approach in the random matrix literature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D states that all experiments (matrix simulations with $N \leq 500$, up to 20 realizations, and phase-diagram scans) complete in under 10 minutes on a standard CPU laptop, requiring no specialized hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research is purely theoretical and computational, involving no human subjects, personal data, or ethically sensitive applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents foundational theoretical results in random matrix theory applied to learning dynamics. There is no direct path to negative societal applications; the work has no foreseeable harmful uses beyond its purely scientific contribution to understanding generalization and early stopping.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper introduces no pre-trained models, datasets, or other assets that could pose risks of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, code packages, or models. All mathematical tools and results referenced are properly cited in the bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new code, datasets, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects research and therefore requires no IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as any important or original component of the core research methodology. All theoretical derivations and numerical experiments were performed without LLM assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.