Spectral Decoupling in Attention Heads: Tracking the Onset of Memorization in Vision Transformers

Dibyajyoti Jena*, Sumantra Dutta Roy*, and Vinayak Abrol†
*Dept. of Electrical Communication Engg., Indian Institute of Technology Delhi
†Dept. of Computer Science and Engg., Indraprastha Institute of Information Technology Delhi
Emails: {eez248388, sumantra}@ee.iitd.ac.in, abrol@iiitd.ac.in

***Abstract*—Despite the empirical dominance of Vision Transformers (ViTs), a rigorous theoretical characterization of their learning dynamics remains elusive. Lacking the explicit inductive biases of convolutional architectures, ViTs are particularly susceptible to severe memorization in data-constrained regimes. We address this generalization-memorization gap through the lens of Random Matrix Theory (RMT), systematically investigating the internal training dynamics of ViT attention projection (key, query, value) matrices. By analyzing their Empirical Singular Value Distributions (ESVD), we uncover a critical spectral phase transition: while early training rapidly extracts task-specific features, manifesting as spectral 'spikes' emerging from the RMT 'bulk', the stabilization of these outliers coincides precisely with the stagnation of validation accuracy. Crucially, we demonstrate that even after feature stabilization, the distribution's 'bulk' continues to evolve, providing empirical evidence that subsequent gradient updates drive structural data memorization rather than generalizable learning. We show that the spectral drift of these top singular values acts as an empirical indicator of the onset of overfitting. Our RMT-grounded perspective offers principled heuristics for monitoring network stability & provides a theoretical foundation for more resilient Transformer designs.**

***Index Terms*—Vision Transformers, Random Matrix Theory, Signal Propagation, Overfitting, Generalization.**

I. Introduction

The meteoric rise of Vision Transformers (ViTs) [1] has redefined computer vision, yet their empirical success remains theoretically opaque. Unlike Convolutional Neural Networks (CNNs), which rely on explicit inductive biases like translation invariance & locality, ViTs operate largely as structural blank slates. While this flexibility enables massive scalability, it renders the architecture notoriously prone to spectral instability & rote memorization when trained in data-constrained regimes. Consequently, mitigating overfitting in ViTs often remains a heuristic driven endeavor [2] rather than a principled science.

To demystify these internal learning dynamics, Random Matrix Theory (RMT) has emerged as a uniquely suited framework for high-dimensional weight analysis [3, 4]. Foundational works [3, 5], have demonstrated that implicit self-regularization in deep neural networks can be characterized by analyzing the heavy-tailed spectral distributions of fully trained models. These approaches have successfully established that examining post-training weight matrices against classical RMT 'null models' can predict generalization trends without access to testing data.

While static post-hoc spectral analysis provides valuable global quality metrics, it naturally obscures the temporal, epoch-by-epoch mechanisms of how a network transitions from learning to memorizing. Further, analyzing deep networks via generic layer operators can overlook the distinct architectural signatures of the Transformer's core engine: the attention mechanism. In ViTs, where explicit inductive biases are absent, understanding the precise, dynamic interplay of the attention matrices (key, query, value) during the training is critical to addressing the generalization-memorization gap [6].

In this paper, we bridge this gap by conducting a longitudinal RMT analysis of ViT attention heads across data-constrained datasets (e.g., TinyImageNet, CIFAR). At initialization, the empirical singular value distribution (ESVD) of these square projection matrices is strictly governed by the Quarter-Circle [7]. By tracking the spectral drift continuously throughout training, we uncover a critical phase transition: early feature extraction manifests as dominant singular values ('spikes') escaping the RMT 'bulk', while the subsequent, continued evolution of the bulk provides empirical evidence of structural data memorization. Ultimately, we demonstrate that monitoring the divergence between these stabilized spikes and the shifting bulk offers a data-independent heuristic to pinpoint the onset of overfitting.

II. Related work

Originally developed to model the energy levels of heavy nuclei, Random Matrix Theory (RMT) has emerged as a powerful statistical framework for analyzing the high-dimensional landscapes of deep neural networks. Foundational works by Pennington et al. [8, 9, 10] utilized RMT to characterize the geometry of deep learning loss surfaces and the stability of gradient descent. This line of inquiry introduced critical engineering concepts such as dynamical isometry and optimal initialization strategies. Building on this foundation, work in [11] demonstrated that universal outlier behaviors within these empirical spectra act as key indicators of network stability and learning capacity. The primary advantage of RMT in this context is its ability to extract macroscopic properties of the learning algorithm directly from the microscopic structure of the random weight matrices.

The application of RMT to monitor active training dynamics was significantly advanced by Martinet et al. [3]. Their framework provided extensive evidence linking the spectral evolution of weight matrices to distinct phases of neural network

training. Crucially, their work established that large spectral outliers, 'spikes' that escape the random Marchenko-Pastur 'bulk', serve as definitive signatures of well-trained layers that have successfully absorbed latent data structures. This demonstrated that RMT can effectively diagnose model quality and implicit self-regularization without requiring access to held-out testing data.

While a growing body of literature investigates signal propagation and scaling limits in Transformers at extreme depths [12, 13, 14], the specific application of RMT to understand signal routing, particularly the generalization-memorization trade-off remains underexplored [15]. Although the fundamental principles of RMT apply universally, the architectural nuances of the self-attention mechanism and the absence of spatial inductive biases introduce complex spectral dynamics. Our work bridges this gap by shifting from the static, post-hoc RMT analysis of prior literature to a longitudinal tracking of Transformer weights, directly correlating spectral phase transitions with the onset of overfitting.

III. Spectral Evolution of Weight Matrices

We employ tools from RMT to characterize the training dynamics of ViTs by analyzing the spectral evolution of the attention mechanism's projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{N\times N}$, treating them as empirical realizations of large random matrix ensembles.

A. Theoretical Distribution of Singular Values

At initialization, weight matrices $W \in \mathbb{R}^{N\times N}$ are typically drawn from i.i.d. entries. In the limit $N \to \infty$, the eigenvalue density $\rho_\Lambda(\lambda)$ of the covariance matrix $C = WW^T$ converges to the Marchenko-Pastur (MP) distribution. For square matrices ($Q = 1$) with variance σ^2, the density simplifies to:

$$\rho_\Lambda(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{\frac{4\sigma^2-\lambda}{\lambda}}, \quad \lambda \in (0, 4\sigma^2]$$

Standard deep learning frameworks often utilize a uniform initialization $\mathcal{U}(-\alpha, \alpha)$ with $\alpha = 1/\sqrt{N}$, yielding a variance $\sigma^2 = 1/3$ (normalized for the $1/N$ factor). Substituting this into the MP density provides the baseline 'null' distribution for the network's untrained state:

$$\rho_\Lambda(\lambda) = \frac{3}{2\pi}\sqrt{\frac{4/3-\lambda}{\lambda}}$$

B. The Quarter-Circle Law

To analyze weight matrices directly via Singular Value Decomposition (SVD), we transform the eigenvalue density using the relation $\lambda = s^2$, where s denotes the singular values of the matrix. Applying the change of variables $\rho_S(s) = \rho_\Lambda(s^2)|\frac{d\lambda}{ds}|$ with $\frac{d\lambda}{ds} = 2s$, the ESVD for a matrix with $\sigma^2 = 1/3$ becomes:

$$\rho_S(s) = \frac{\sqrt{3}}{\pi}\sqrt{4-3s^2}, \quad s \in \left(0, \frac{2}{\sqrt{3}}\right]$$

By the Universality Principle, Quarter-Circle [7] remains invariant of the underlying entry-wise distribution (e.g., Gaussian or Uniform) for large N, providing a robust 'null model' for an untrained network. Figure 1 illustrates the tight alignment between this theoretical density & the empirical spectrum of a 1000×1000 uniformly initialized matrix.

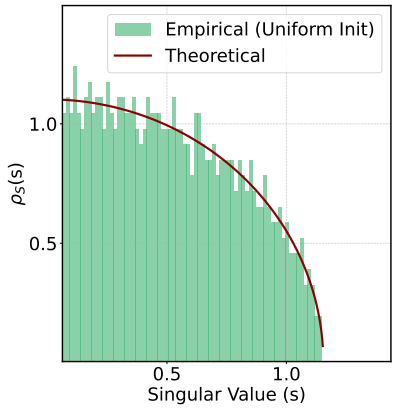

Figure 1. Empirical singular value distribution of a 1000 x 1000 matrix, along with a theoretical overlay matching it perfectly.

Probing Training Dynamics: Spikes vs. Bulk As training progresses, the weight matrices W deviate from their i.i.d. initialization. We track this evolution by sampling W at discrete epoch intervals and computing its SVD. We decompose the spectrum into two distinct components:

- **The Bulk (Noise):** The dense region of singular values that closely follows the RMT prediction. In deep learning, the bulk is often hypothesized to represent "thermal noise" or unlearned parameters [3].
- **The Spikes (Information):** Outlying singular values $s_i > s_+$ that escape the bulk. The BBP (Baik-Ben Arous-Péché) transition [16] suggests that these spikes emerge when the signal-to-noise ratio exceeds a critical threshold, effectively encoding the learned task features.

Monitoring the Memorization Gap We hypothesize that the 'spectral drift' - the movement of these top singular values, serves as a proxy for feature extraction. Crucially, we investigate the regime where the spikes stabilize while the bulk continues to shift. We propose that this divergence marks the transition from generalization (learning low-rank task features) to memorization (fitting high-rank noise in the data-constrained regime).

C. Theoretical Framework: The Spike-Bulk Phase Transition

While the Quarter-Circle Law characterizes the static, untrained state of the Vision Transformer, we must formalize the temporal dynamics that drive the network from generalization to memorization. To formally ground the temporal dynamics of the ESVD, we analyze the implicit bias of gradient descent on Cross-Entropy (CE) loss. Let $W(t) \in \mathbb{R}^{N\times N}$ denote an attention projection matrix at training time t. In data-constrained regimes, ViTs rapidly enter the interpolation regime (perfect training accuracy).

Theorem 3.1 (Asymptotic Spectral Decoupling): Under gradient flow on exponentially-tailed losses (e.g., CE), the tem-

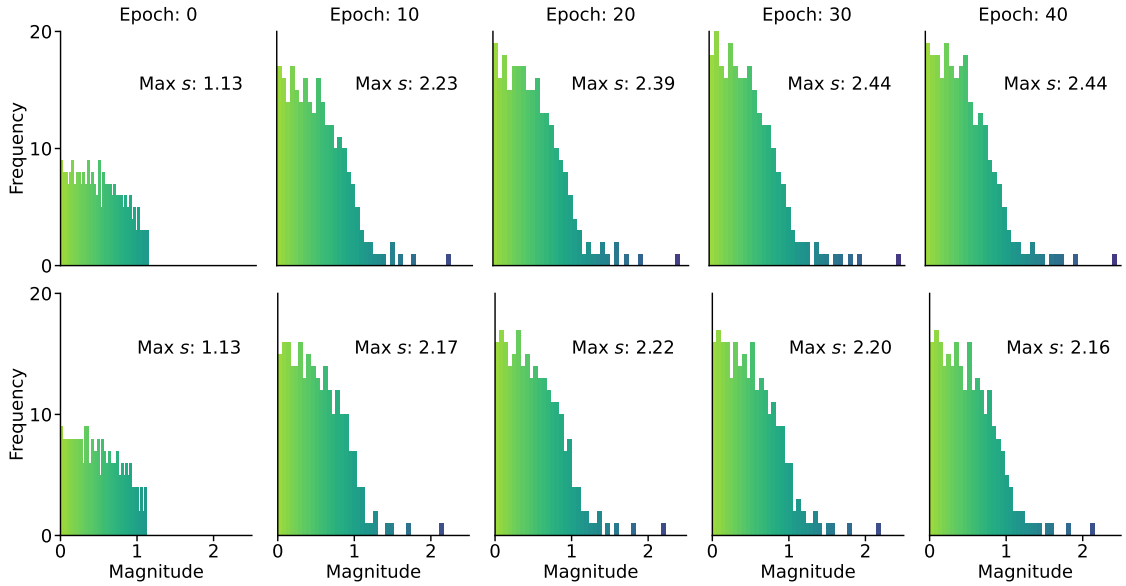

Figure 2. ESVD of the Layer 1 W_Q matrix for TinyImageNet (top) and CIFAR-10 (bottom). This early layer demonstrates that dataset-specific spectral alignment happens from the very onset of training.

poral evolution of the ESVD of $W(t)$ decouples into two dynamic phases separated by an interpolation threshold, t_{crit}:

1) **Signal Extraction:** For $t \leq t_{crit}$, the top k singular values ('spikes') rapidly extract primary task features. For $t > t_{crit}$, the network enters the margin-maximization regime; the signal singular vectors converge directionally to the max-margin solution, and their magnitude growth exponentially decelerates to a logarithmic rate ($\mathcal{O}(\log t)$), manifesting empirically as a plateau in spectral drift.
2) **Noise Absorption:** For $t > t_{crit}$, gradients orthogonal to the max-margin subspace remain bounded away from zero. Lacking inductive bias, the network minimizes residual CE loss by fitting high-dimensional dataset noise, continuously driving structural deformations predominantly within the RMT bulk.

Proof Sketch: Extending the implicit bias of gradient descent to deep homogeneous networks [17, 18], the weight matrix in the late-training interpolation regime ($t > t_{crit}$) asymptotically decomposes as:

$$W(t) = W_{margin} \log(t) + \eta(t)$$

where W_{margin} is a strict low-rank max-margin classifier for the separable features, and $\eta(t)$ is a bounded perturbation matrix fitting sample-specific noise. For the dominant spectral components governed by $W_{margin} \log(t)$, of these top k singular values scales as $\frac{d}{dt} s_{1..k}(t) \propto \frac{1}{t}$. As training surpasses t_{crit}, this derivative becomes vanishingly small, halting directional updates and stabilizing the Wasserstein distance. Conversely, the bounded noise perturbation $\eta(t)$ must utilize the matrix's full capacity to interpolate residuals. This continuous injection of gradients continuously shuffles the lower-rank eigenvalues, structurally shifting the ESVD bulk while the spikes remain anchored.

Corollary 3.1.1: [Depth-Invariant Timing] While the absolute signal magnitude $\|W_{margin}\|$ varies across network depth, the critical transition time t_{crit} is dictated by the global loss landscape reaching the interpolation threshold. Consequently, the $\mathcal{O}(1/t)$ logarithmic slowdown occurs simultaneously across all attention layers, providing a globally synchronized, depth-invariant indicator of overfitting (as empirically observed in Figure 3).

IV. Experimental Setup

To empirically validate our theoretical framework, we train a custom Vision Transformer (ViT) comprising eight attention blocks (eight heads per block, embedding dimension $d = 256$) followed by a standard MLP classification head. To track the internal spectral dynamics, we probe the network every five epochs, extracting the attention projection matrices (W_Q, W_K, W_V). The model is trained across three benchmark datasets: TinyImageNet (200 classes, 500 images/class, 64×64 pixels), CIFAR-10 (10 classes, 5000 images/class), and CIFAR-100 (100 classes, 500 images/class). All hyperparameters are held constant across experiments to isolate the effect of dataset scale and complexity.

Because the self-attention mechanism lacks explicit spatial inductive biases, ViTs require massive datasets to naturally learn generalizable features. Our data-constrained setup deliberately forces the over-parameterized model into the interpolation regime. We observe this stark generalization-memorization gap most prominently on TinyImageNet, where training accuracy reaches 100% by epoch 50, yet validation accuracy severely stagnates at 35%. CIFAR-100 and CIFAR-10 exhibit similar overfitting plateaus, achieving validation

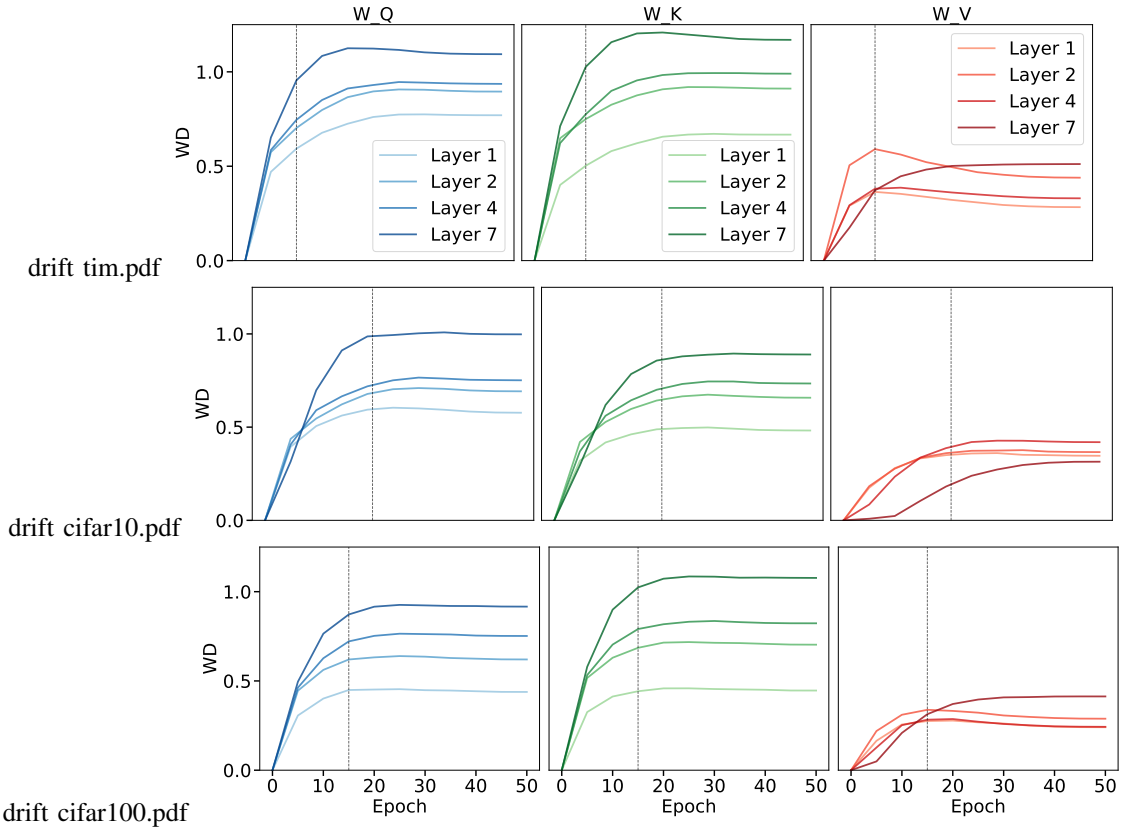

Figure 3. Evolution of spectral drift across network depths. The Wasserstein distance (WD) of the dominant five singular values plateaus exactly at the epoch of minimum validation error (dashed lines) across TinyImageNet (Row 1), CIFAR-10 (Row 2), and CIFAR-100 (Row 3), empirically marking the transition to structural memorization.

accuracies of only 50% and 77%, respectively. This setting, where a model memorizes perfectly but fails to generalize, provides the ideal empirical testbed to visualize the spectral evolution of the weight matrices and validate the precise onset of structural memorization.

V. Experimental Results

A. Visualizing the Spike-Bulk Phase Transition

We first visualize the structural evolution of the weights during training. Figure 2 plots the ESVD of the query projection matrix (W_Q) from the first attention layer during training. We specifically probe the first layer to demonstrate that dataset-specific spectral alignment begins at the earliest stages of the network. At initialization (epoch 0), the ESVD adheres to the Quarter-Circle Law, with a maximum singular value (s_{max}) of 1.13 across both datasets. As training commences (Phase I: Signal Extraction), the network learns dominant task-specific features. Visually, this manifests as top singular values ("spikes") escaping the RMT bulk. By epoch 20, s_{max} significantly increases to 2.39 for TinyImageNet and 2.22 for CIFAR-10. As training progresses further, we observe the critical phase transition predicted by Theorem 1. The dominant singular values stabilize and remain structurally anchored from epoch 20 to 40. While these spikes stabilize, the network continues to receive gradient updates, and training loss continues to approach zero. Visual inspection confirms that these subsequent structural changes occur predominantly within the shifting dense bulk of the distribution. This provides direct empirical evidence that once primary features are extracted, the network utilizes its over-parameterized capacity to memorize high-rank dataset noise.

B. Quantifying Spectral Drift and Depth-Invariance

To quantify this phase transition and establish its direct relationship with overfitting, Figure 3 tracks the "Spectral Drift" of the attention projection matrices (W_Q, W_K, W_V). We define Spectral Drift mathematically as the Wasserstein distance of the top 5 singular values from their initial random state at epoch 0. This metric is recorded across four network depths (layers 1, 2, 4, and 7). The vertical dashed lines in Figure 3 represent the exact epoch where the validation error achieves its minimum (epoch 10 for TinyImageNet, 21 for CIFAR-10, and 15 for CIFAR-100). A universal pattern emerges across all datasets and projection matrices: the spectral drift exhibits a steep, monotonic increase during early epochs, but abruptly plateaus almost exactly at the point of minimum validation loss. This confirms the spectral decoupling hypothesis. Once the Wasserstein distance flattens, the top singular values become directionally constant ($\frac{d}{dt} \approx 0$). Consequently, any further reduction in training loss is driven entirely by the bulk

absorbing structural noise. Furthermore, Figure 3 provides empirical validation for Corollary 3.1.1; while deeper layers (e.g., Layer 7) exhibit a higher absolute magnitude of drift, the critical plateau time (t_{crit}) is remarkably uniform across all network depths.

We explicitly acknowledge the boundaries of this initial study, including the small to mid scale experiments and a generic ViT architecture instead of standard models like ViT-B/16. There are many directions to extend this line of work and further validate the claims, including (and not limited to) ablations on the number of top singular values, generalizing to image classification architectures, and checking if the theorem holds tight under heavy regularization in vision models.

VI. Conclusion

We systematically investigated the internal training dynamics of ViTs through the lens of RMT by tracking the empirical singular value distributions of attention projection matrices and uncovered a critical spectral phase transition that formally explains the ViT generalization-memorization gap. Our findings demonstrate a strict decoupling in the learning process: the extraction of generalizable features manifests as spectral 'spikes' that plateau precisely at the point of minimum validation error, while the continued absorption of dataset noise persistently deforms the RMT 'bulk'. By quantifying this divergence via Wasserstein spectral drift, we established a depth-invariant, and data-independent signature for the onset of overfitting. Ultimately, this RMT framework moves deep learning analysis beyond trial-and-error, providing practitioners with a principled early-stopping criteria and a foundation for designing spectrally stable architectures.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021.

[3] C. H. Martin and M. W. Mahoney, "Implicit-regularization in deep neural networks: Evidence from random matrix theory and implications for learning," *Journal of Machine Learning Research*, vol. 22, no. 165, 2021.

[4] M. Thamm, M. Staats, and B. Rosenow, "Random matrix analysis of deep neural network weight matrices," *Phys. Rev. E*, vol. 106, 2022.

[5] C. H. Martin, T. Peng, and M. W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications*, vol. 12, no. 1, 2021.

[6] J. Jiang, W. Huang, M. Zhang, T. Suzuki, and L. Nie, "Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[7] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, 1967.

[8] J. Pennington and Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017.

[9] J. Pennington and P. bahri, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] J. Pennington, S. Schoenholz, and S. Ganguli, "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] N. P. Baskerville, J. P. Keating, F. Mezzadri, J. Najnudel, and D. Granziol, "Universal characteristics of deep neural network loss surfaces from random matrix theory," *Journal of Physics A: Mathematical and Theoretical*, vol. 55, no. 49, 2022.

[12] A. Kedia, M. A. Zaidi, S. Khyalia, J. Jung, H. Goka, and H. Lee, "Transformers get stable: An end-to-end signal propagation theory for language models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024.

[13] H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei, "Deepnet: Scaling transformers to 1,000 layers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, 2024.

[14] L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi, "Signal propagation in transformers: Theoretical perspectives and the role of rank collapse," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[15] T. N. Saada, A. Naderi, and J. Tanner, "Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers," in *Forty-second International Conference on Machine Learning*, 2025.

[16] J. Baik, G. B. Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, vol. 33, no. 5, 2005.

[17] K. Lyu and J. Li, "Gradient descent maximizes the margin of homogeneous neural networks," in *International Conference on Learning Representations (ICLR)*, 2020.

[18] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *Journal of Machine Learning Research*,

vol. 19, no. 1, 2018.