Spectral Dynamics in Neural Network Training: Mathematical Foundations for Understanding Representational Development

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding the mathematical foundations underlying neural network training dynamics is essential for mechanistic interpretability research. We develop a continuous-time, matrix-valued stochastic differential equation (SDE) framework that rigorously connects SGD optimization to the evolution of spectral structure in weight matrices. We derive exact SDEs showing that singular values follow Dyson Brownian motion with eigenvalue repulsion, and characterize stationary distributions as gamma-type densities with power-law tails that explain the empirically observed 'bulk+tail' spectral structure in trained networks. Through controlled experiments on transformer and MLP architectures, we validate our theoretical predictions and demonstrate quantitative agreement between SDE-based forecasts and observed spectral evolution, providing a mathematical framework for mechanistic interpretability researchers to predict when interpretable structure emerges during training and monitor the development of internal representations.

1 Introduction

Deep neural networks have fundamentally transformed machine learning, achieving unprecedented performance across diverse domains [13, 32, 9]. Yet despite their empirical success, our theoretical understanding of how neural networks learn remains remarkably incomplete [36, 20]. Central to this understanding is the evolution of weight matrices, whose spectral properties—the distribution and dynamics of singular values—provide deep insights into optimization dynamics, generalization behavior, and implicit regularization [24, 18]. This gap is particularly relevant for mechanistic interpretability, which seeks to understand neural networks through analysis of their internal representations [22]. While most interpretability research focuses on analyzing trained networks, understanding the mathematical foundations of how representations develop during training could provide valuable insights into when and why interpretable structure emerges during the learning process.

At initialization, weight matrices exhibit well-characterized random matrix statistics described by the Marchenko-Pastur law, which characterizes the eigenvalue distribution of large random matrices, and related results from random matrix theory (RMT). However, training fundamentally alters these spectral properties, producing empirically observed 'bulk+tail' structured distributions that correlate strongly with generalization performance [23, 19]. Existing theoretical frameworks fail to explain this transformation: while RMT describes initial conditions and stochastic differential equations (SDEs) can model SGD, current analyses focus on scalar parameters or low-rank models, failing to capture the full matrix-valued dynamics [16]. Most critically, no unified framework connects the microscopic stochastic dynamics of SGD to the macroscopic spectral evolution observed empirically.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this paper, we bridge this gap by developing a continuous-time, matrix-valued SDE framework with carefully designed small-scale experiments that captures the full dynamics of singular value evolution under SGD. Our key contributions are:

1. We derive exact SDEs for individual singular values under isotropic SGD noise, connecting the microscopic parameter updates to macroscopic spectral dynamics. Under the assumption of negligible gradients, we show that squared singular values follow a Dyson Brownian motion with $\beta = 1$, explaining the eigenvalue repulsion and spectral spreading.
2. We characterize the stationary spectral distribution in the non-negligible gradient regime using mean-field theory. We prove that the limiting distribution follows a gamma-type density with power-law tails, recovering the empirically observed 'bulk+tail' structure in trained networks and providing the first theoretical explanation for this empirical phenomenon.
3. Through experiments on transformer [32], vision transformer [4], and MLP [28] architectures, we demonstrate quantitative agreement between our SDE-based predictions and spectral evolution and propose an algorithm to forecast singular value dynamics from minimal gradient information. The code for all of our experiments is available here: `https://anonymous.4open.science/r/featureevolution-0E1C/`

We show that SGD's stochastic noise acts like a "spectral sculptor"—initially spreading eigenvalues apart via repulsion, then concentrating them into beneficial empirically observed 'bulk+tail' structured patterns that enable generalization, connecting microscopic mini-batch randomness to macroscopic spectral evolution. Our work demonstrates how small-scale experimentation can unlock fundamental insights with implications for initialization strategies, optimization algorithm design, and understanding why deep learning works.

2 Related Works

Spectral Analysis of Neural Network Weights. RMT establishes that at initialization, weight matrices follow Wigner's semicircle law [34] and the Marčenko–Pastur distribution [17], with edge statistics governed by Tracy-Widom distributions [30]. Training induces pronounced deviations: singular-value spectra become highly anisotropic [29], evolving 'bulk+tail' structured distributions linked to class structure [23] and implicit regularization [19]. Martin and Mahoney [19] identified 5+1 phases of spectral evolution and showed that batch size affects spectral properties, with smaller batches leading to stronger implicit self-regularization. Extensions to empirically observed 'bulk+tail' structured matrix ensembles [2] provide theoretical foundations past Gaussian universality classes.

SGD as Stochastic Dynamics. SGD can be approximated by stochastic differential equations (SDEs), with constant-rate SGD behaving like an Ornstein–Uhlenbeck process [15] and anisotropic noise structures enabling escape from sharp minima [37]. Weight updates have been mapped to Dyson Brownian motion [1], explaining eigenvalue repulsion as a Coulomb-gas phenomenon. The mathematical foundation relies on Itô calculus for matrix functions [6] and Fokker-Planck equations for interacting particle systems [26]. The implicit regularization effects of gradient descent have been explored [21].

Training Dynamics and Interpretability. Recent mechanistic interpretability research has revealed that neural networks undergo distinct phases during training, such as grokking transitions [25] and sudden attention head specialization [33]. Understanding the mathematical foundations underlying these phenomena could inform interpretability research by providing principled frameworks for analyzing training dynamics.

3 Methodology

Our approach transitions from discrete microscopic SGD dynamics to continuous macroscopic spectral evolution. We produce notation for the training of neural networks via a spatiotemporal interpretation of the evolution of the weight matrices with stochasticity: $dW(x,t) = -\eta \frac{\partial \mathcal{L}}{\partial W(x,t)} dt + \sqrt{2\eta D_W(x,t)}\, d\mathcal{W}_W(x,t)$, where $\frac{\partial \mathcal{L}}{\partial W(x,t)}$ is the gradient of the loss, η is the learning rate, D is

an effective diffusion constant described by $d\mathcal{W}_W$, and $d\mathcal{W}_b$ are independent matrix/vector-valued Wiener processes, capturing the stochastic dynamics of SGD. We split our analysis into two cases: negligible gradient in loss and non-negligible gradient in loss. For the first limit (negligible gradient in loss, or $\frac{\partial \mathcal{L}(W)}{\partial W} \approx 0$), we perform SVD/eigenvalue decomposition and use Ito Calculus (see Appendix 6.1) to arrive at the result of **Theorem 3.1**:

Theorem 3.1 (Stochastic Dynamics of Singular Values)**.** *Let $W \in \mathbb{R}^{m \times n}$ evolve via stochastic gradient descent with noise. Then, the singular values $\sigma_k(W)$ follow the SDE:*

$$d\sigma_k(t) = \left[-\eta u_k^T (\nabla_W \mathcal{L}) v_k + \eta D \left(\frac{m-n+1}{2\sigma_k} + \sum_{j \neq k} \frac{\sigma_k}{\sigma_k^2 - \sigma_j^2} \right) \right] dt + \sqrt{2\eta D} d\beta_k(t)$$

where u_k, v_k are the singular vectors and D is the effective diffusion strength.

This theorem shows that under SGD noise, singular values behave like interacting particles that repel each other (the $\sum_{j \neq k}$ terms), explaining why the spectrum spreads out during training rather than collapsing. This SDE can then be mapped to Dyson-Brownian Motion processes (see Appendix 6.6 whose statistics are described by the Marčenko–Pastur (**Lemma 6.9**) and Tracy-Widom distributions respectively (**Lemma 6.11**).

After deriving these microscopic underpinnings, we may consider the second case (non-negligible gradient in loss, or $\frac{\partial \mathcal{L}(W)}{\partial W} \neq 0$) and transition into a macroscopic limit by considering the empirical spectral density distribution $\rho(\lambda, t)$ such that: $\rho(\lambda, t) = \frac{1}{r} \sum_{k=1}^{r} \delta(\lambda - \lambda_k(t))$ As $r \to \infty$, we assume that $\rho(\lambda, t)$ converges to a deterministic density function, normalized such that $\int \rho(\lambda, t) d\lambda = 1$.

To study the dynamics of the squared singular values $\{\lambda_j\}$ then, we adopt a mean-field perspective by assuming the effective influence of the complex loss function $\mathcal{L}(W)$ is captured by a potential $\mathcal{L}_{MF}$ that depends only on this set. This postulates the form: $\mathcal{L}(W) \approx \mathcal{L}_{MF}(\{\lambda_j\}) = \frac{c}{2} \sum_{j=1}^{r} (\lambda_j - \lambda^*)^2$, where these eigenvalues λ_j are "driven" to λ^* as in **Lemma 6.10**.

Although deep neural network training objectives are generally nonconvex, it was shown in prior work [5] that every stationary point of the original nonconvex problem coincides with the global optimum of a suitably defined subsampled convex program. As such, while the parameter-space landscape may admit many critical points, their spectral signatures at stationarity are governed by a convex variational principle. Consequently, modeling the large-r limit of the empirical spectral density $\rho(\lambda, t)$ via a deterministic mean-field potential $\mathcal{L}_{MF}$ is fully justified as in **Corollary 6.15**. Evaluating this spectral density distribution, neglecting its highest order terms, and we derive **Theorem 3.2**:

Theorem 3.2 (Stationary Distribution of Singular Values)**.** *Under the stationary mean-field approximation, the probability density function of the singular values follows a Gamma-type distribution:*

$$p_\sigma(\sigma) = 2 \frac{\left(\frac{\beta_1}{4\eta D} \right)^{\frac{m-n+3}{4}}}{\Gamma\left(\frac{m-n+3}{4} \right)} \sigma^{\frac{m-n+1}{2}} e^{-\left(\frac{\beta_1}{4\eta D} \right) \sigma^2}$$

where β_1 is an effective noise constant representing the mean-field restoring force of the gradient, D is the diffusion constant, and the weight matrix W is $m \times n$.

This gamma-type distribution with power-law tails captures the "bulk+tail" structure observed empirically—most singular values cluster in a bulk region, while a few large values form the heavy tail that correlates with good generalization.

We present the proofs for both theorems in Appendix 6.1. Note that we treat **Theorem 3.2** phenomenologically by fitting to within bounds beyond the Tracy-Widom distribution predictions (see Corollary 6.1). Thus, we take a "mean-field" approach to deconstructing this system. See **Lemma 6.17** and **Lemma 6.18** for the estimation of the effective diffusion and noise constants governing the stochastic dynamics.

4 Experiments

4.1 Experimental Setup

We validate our spectral-SDE framework on three canonical architectures, namely a GPT-2 model; a Vision Transformer (ViT) [4]; and a MLP—all trained with SGD. This selection is motivated by extensive prior work showing 'bulk+tail' spectra arise across MLPs, CNNs, and transformers [19], and that different architectures exhibit distinct spectral biases and mode-learning rates [35]. We initialize all weights from architecture-specific Gaussian priors, then train GPT-2 on Shakespeare text [10], and ViT/MLP on MNIST [14] and CIFAR-100 [12]. More experimental details are in Appendix 8.

4.2 Singular Value Evolution Simulation and Analysis

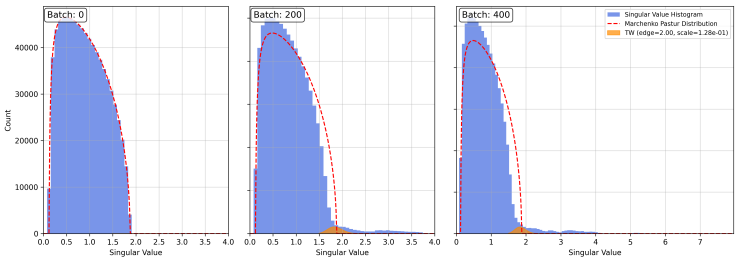

Figure 1: Singular-value histograms at batches 0, 200, and 400, overlaid with the Marčenko–Pastur (MP) bulk law (red dashed) and the Tracy–Widom (TW) edge curve (green).

In Figure 1, at initialization the empirical spectrum adheres almost exactly to the MP prediction with no outliers beyond the TW edge, verifying random initialization assumptions. Our analysis begins by modeling the stochasticity of SGD with an isotropic noise term, resulting in a Langevin-type SDE. We acknowledge that this is an idealization; the true noise covariance of SGD is known to be anisotropic and parameter-dependent. However, this assumption provides a tractable starting point that allows us to establish a clear, analytical connection to the classical frameworks of Random Matrix Theory. This approach enables us to isolate and understand the fundamental repulsive dynamics that serve as a baseline for spectral evolution. We explicitly address the extension to the more realistic anisotropic case in our appendix (see **Proposition 6.16**), which we identify as a crucial direction for future work. At batch 200, the MP/TW fits begin to underpredict mass near the spectrum's edge. This underprediction implies that growing correlations within the weight matrix diminish the effective size of the random matrix, consequently amplifying the dominance of edge statistics and fluctuations. Physically, this has the interpretation that we are still traversing either local minima or plateaus within the loss function landscape (producing weakly correlated learned features). At batch 400, this tail becomes increasingly pronounced: the bulk shifts rightward and a persistent shoulder of large singular values forms outside the MP support, showing that our matrix is becoming increasingly correlated. This is predicted to result from mostly gradient-derived information. Thus we understand the singular values beyond the TW threshold as highly correlated learned features. We empirically test our theory via Algorithm 1. The algorithm uses gradient loss and singular value information to use the dynamic equation previously described to predict singular values over time.

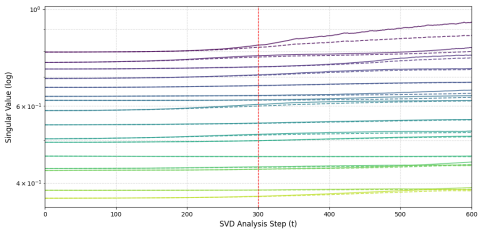

Figure 2: Predicted singular values (dashed) versus true.

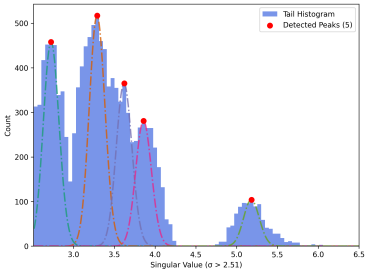

Figure 3: Predicted heavy tails via Theorem 3.2.

In Figure 2, we track the top 8 singular values of a representative linear layer in a MLP over 800 training batches with CIFAR-10, plotting empirical trajectories against our bootstrap-drift predictions. The leading modes rise faster and the gradual increases of lower modes are reproduced by the prediction algorithm, with deviations starting around batch 300 as the spectrum begins to develop empirically observed 'bulk+tail' structure. This close alignment across all 8 modes up to the heavy-tail regime demonstrates that our continuous-time, matrix-valued SDE framework accurately forecasts the full singular-value dynamics from minimal gradient information. However, we hypothesize the anisotropic noise causes a bulk of the observed deviation for the singular values. This hypothesis is as follows: the larger singular values might experience greater effects from anisotropic noise due to preferential alignment of noise with their dominant singular vectors and potentially larger Hessian components (see **Proposition 6.16**). In Figure 3, the fits predict the qualitative shape well, but underpredict counts significantly, which we attribute to the aforementioned anisotropic noise hypothesis.

4.3 Practical Applications

Our framework provides mechanistic interpretability researchers with quantitative tools for understanding representation development. By monitoring the transition from Marchenko-Pastur to bulk+tail spectral structure, researchers can: (1) identify critical training phases when interpretable features emerge - the deviation from random matrix statistics signals the onset of structured learning; (2) predict which layers develop interpretable structure first by tracking layer-wise spectral evolution rates; (3) design interventions that encourage interpretable representations by tuning η and D to elucidate feature consolidation. The spectral repulsion mechanism we identify suggests that SGD naturally separates features into distinct modes, potentially explaining why neural networks often learn disentangled representations amenable to interpretation.

5 Conclusion

We develop a continuous-time, matrix-valued SDE framework connecting SGD's microscopic dynamics to macroscopic spectral evolution, revealing that squared singular values follow Dyson Brownian motion and produce gamma-type distributions with power-law tails that explain the empirically observed 'bulk+tail' structure in trained networks. Through controlled experiments, we demonstrate quantitative agreement between our predictions and observed spectral evolution, with our forecasting algorithm accurately predicting singular value trajectories until empirically observed 'bulk+tail' structure emerges. While our current analysis assumes isotropic noise, future extensions to anisotropic SGD fluctuations could bridge the gap to real optimization dynamics and enable new preconditioning schemes. For mechanistic interpretability specifically, our framework offers a principled method to predict when random initializations give way to structured, potentially interpretable representations, providing researchers with quantitative markers for when to deploy interpretability tools during training. By providing a complete theoretical characterization of how spectral structure emerges during training, we offer a foundation for future work exploring the relationship between optimization dynamics and the development of interpretable representations.

References

[1] Tim Aarts and Jeroen de Wit. Dyson brownian motion of neural network weights, 2024.

[2] Antonio Auffinger, Gerard Ben Arous, and Sandrine Peche. Poisson convergence for the largest eigenvalues of Heavy Tailed Random Matrices, May 2008.

[3] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Tolga Ergen and Mert Pilanci. The convex landscape of neural networks: Characterizing global optima and stationary points via lasso models. *IEEE Trans. Inf. Theor.*, 71(5):3854–3870, February 2025.

[6] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. June 1990.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015.

[8] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Naveen Goel, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. In *International Conference on Learning Representations (ICLR)*, 2018.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[10] Andrej Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

[11] Andrej Karpathy. Karpathy/nanoGPT, May 2025.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[14] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[15] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. In *JMLR*, volume 18, pages 1–35, 2017.

[16] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference, January 2018.

[17] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik*, 1(4):457–483, 1967.

[18] Charles H. Martin and Michael W. Mahoney. Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning, October 2018.

[19] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *JMLR*, 22(172):1–73, 2021.

[20] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning, July 2017.

[21] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning, April 2015.

[22] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

[23] Vardan Papyan. A mean field theory of batch normalization. *ICLR*, 2020.

[24] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice, November 2017.

[25] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting in small transformers. *arXiv preprint arXiv:2201.02177*, 2022.

[26] Hannes Risken and Till Frank. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer Science & Business Media, September 1996.

[27] Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the convergence rate. *Communications in Mathematical Sciences*, 16(6):2309–2331, 2018.

[28] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[29] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014.

[30] Craig A. Tracy and Harold Widom. On Orthogonal and Symplectic Matrix Ensembles. *Communications in Mathematical Physics*, 177(3):727–754, April 1996.

[31] Craig A. Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023.

[33] Johannes von Oswald, Nikita Krasheninnikov, Luigi Gresele, Bledi Berisha, and Tomaso Poggio. Transformers as support vector machines. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35227–35243. PMLR, 2023.

[34] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.*, 62(3):548–564, 1955.

[35] Ziqi Yao, Alfred O. Hero, and Jose C. Principe. Spectral bias and learning dynamics in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[36] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, February 2017.

[37] Jun Zhu, Tafadzwa Li, and Suvrit Sra. On the variance of the adaptive learning rate and beyond. In *ICLR*, 2019.

A Appendix

A.1 Main Theorem Proofs

Theorem 3.1

Proof. We regard the SGD update with isotropic noise as the Itô SDE on the weight matrix

$$dW \;=\; A\,dt \;+\; \sqrt{2\eta D}\,d\mathcal{W}, \qquad A = -\,\eta\,\nabla_W \mathcal{L},$$

where $d\mathcal{W}$ is a matrix-valued Wiener increment with independent entries. Writing the SVD $W = U\Sigma V^T$ and denoting the kth singular value by σ_k, we apply Itô's lemma to the scalar function $f(W) = \sigma_k(W)$. First, by standard matrix-perturbation theory,

$$\nabla_W \sigma_k \;=\; u_k\, v_k^T, \qquad \Delta_W \sigma_k = \frac{m-n+1}{2\sigma_k} \;+\; \sum_{j\neq k} \frac{\sigma_k}{\sigma_k^2 - \sigma_j^2}.$$

Hence the general Itô formula

$$df(W) = \sum_{i,j} \frac{\partial f}{\partial W_{ij}}\, dW_{ij} \;+\; \tfrac{1}{2} \sum_{i,j,p,q} \frac{\partial^2 f}{\partial W_{ij}\partial W_{pq}}\, dW_{ij}\, dW_{pq},$$

together with

$$dW_{ij} = A_{ij}\,dt + \sqrt{2\eta D}\,d\mathcal{W}_{ij}, \qquad dW_{ij}\,dW_{pq} = 2\eta D\,\delta_{ip}\,\delta_{jq}\,dt,$$

yields

$$d\sigma_k = \Big\langle \nabla_W \sigma_k,\ A \Big\rangle\,dt \ +\ \eta D\,\Delta_W \sigma_k\,dt \ +\ \sqrt{2\eta D}\,\langle \nabla_W \sigma_k,\ d\mathcal{W}\rangle.$$

Substituting $\nabla_W \sigma_k = u_k v_k^T$ gives

$$d\sigma_k = \Big[-\eta\,u_k^T(\nabla_W \mathcal{L})v_k \ +\ \eta D\Big(\tfrac{m-n+1}{2\sigma_k} + \sum_{j\neq k} \tfrac{\sigma_k}{\sigma_k^2-\sigma_j^2}\Big)\Big]dt \ +\ \sqrt{2\eta D}\,d\beta_k(t),$$

where $d\beta_k = u_k^T\,d\mathcal{W}\,v_k$ is a scalar Wiener increment.

Finally, set $\lambda_k = \sigma_k^2$ and apply Itô again:

$$d\lambda_k = 2\sigma_k\,d\sigma_k + (d\sigma_k)^2 = 2\sigma_k\,d\sigma_k + 2\eta D\,dt.$$

Substituting the above expression for $d\sigma_k$ and simplifying yields

$$d\lambda_k = \Big[-2\sqrt{\lambda_k}\,\eta\,u_k^T(\nabla_W \mathcal{L})v_k + \eta D\,(m{-}n{+}3) + 2\eta D \sum_{j\neq k} \frac{\lambda_k}{\lambda_k-\lambda_j}\Big]dt + 2\sqrt{\lambda_k}\,\sqrt{2\eta D}\,d\beta_k(t),$$

which is the desired result. □

At initialization ($t = 0$), the singular value spectrum of a random weight matrix is indeed described by the Marchenko-Pastur (MP) law. The contribution of our theorem is not to re-derive this initial state, but rather to characterize the initial dynamics that drive the spectrum away from this random configuration. The theorem formally describes the repulsive force ($\sum_{j\neq k}$ ...) induced by SGD's stochastic updates, which is the fundamental mechanism that introduces structure into the spectrum. We note that the dynamics of the squared singular values, $\lambda_k = \sigma_k^2$, are more precisely termed a Wishart process, a matrix-valued generalization related to Dyson Brownian motion.

Theorem 3.2

Proof. Under the stationary mean-field approximation with vanishing gradient, each squared singular value λ_t evolves according to the one-dimensional SDE

$$d\lambda_t \;=\; (\alpha_0 - \beta_1\,\lambda_t)\,dt \;+\; \sqrt{8\eta D\,\lambda_t}\,dW_t,$$

where $\alpha_0 = \eta D(m-n+3)$ and $\beta_1 > 0$ is a constant. The corresponding stationary Fokker–Planck equation for the density $p(\lambda)$ is (by setting $\frac{\partial p(\lambda,t)}{\partial t} = 0$

$$\frac{\partial p(\lambda,t)}{\partial t} = -\frac{\partial}{\partial \lambda}[(\alpha_0 - \beta_1\lambda)p(\lambda,t)] + \frac{1}{2}\frac{\partial^2}{\partial \lambda^2}[8\eta D\lambda p(\lambda,t)]$$

$$0 \;=\; -\frac{d}{d\lambda}\big[(\alpha_0 - \beta_1\lambda)\,p(\lambda)\big] + \frac{1}{2}\,\frac{d^2}{d\lambda^2}\big[8\eta D\,\lambda\,p(\lambda)\big].$$

Integrating once under the zero-flux boundary condition gives

$$(\alpha_0 - \beta_1\lambda)\,p(\lambda) \;=\; 4\eta D\,\frac{d}{d\lambda}\big[\lambda\,p(\lambda)\big].$$

Rearranging and separating variables, we have

$$\frac{p'(\lambda)}{p(\lambda)} = \Big(\frac{\alpha_0}{4\eta D} - 1\Big)\frac{1}{\lambda} \;-\; \frac{\beta_1}{4\eta D}.$$

Integrating gives us

$$\ln p(\lambda) = \Big(\frac{\alpha_0}{4\eta D} - 1\Big)\ln\lambda \;-\; \frac{\beta_1}{4\eta D}\,\lambda \;+\; \ln C,$$

so that

$$p(\lambda) = C\,\lambda^{\frac{\alpha_0}{4\eta D}-1}\exp\Big(-\tfrac{\beta_1}{4\eta D}\,\lambda\Big),$$

with C fixed by normalization below

$$C = \frac{\big(\frac{\beta_1}{4\eta D}\big)^{\frac{\alpha_0}{4\eta D}}}{\Gamma\big(\frac{\alpha_0}{4\eta D}\big)}.$$

Noting $\alpha_0 = \eta D(m-n+3)$ gives us the claimed form.

Finally, since $\sigma = \sqrt{\lambda}$, we find the push-forward density

$$p_\sigma(\sigma) = 2\sigma\,p(\sigma^2) = 2\frac{\big(\frac{\beta_1}{4\eta D}\big)^{\frac{m-n+3}{4}}}{\Gamma\big(\frac{m-n+3}{4}\big)}\,\sigma^{\frac{m-n+3}{2}-1}\exp\Big(-\tfrac{\beta_1}{4\eta D}\,\sigma^2\Big),$$

as desired. □

In deriving the stationary distribution in **Theorem 3.2**, we adopt a mean-field approximation. This approach decouples the interacting system of singular values, allowing us to analyze the dynamics of a single value within an effective potential. We recognize that this is a significant simplification, as the Coulomb-type repulsion term is fundamental to the transient dynamics. However, our goal here is to model the effective stationary state that emerges after prolonged training. In this limit, it is reasonable to approximate the complex, N-body interaction by an average restoring force, captured by the β_1 term. While this model neglects higher-order correlations, it yields a tractable Fokker-Planck equation whose solution successfully recovers the characteristic shape of the "bulk and tail" structure observed empirically.

A.2 Backpropagation as a Discrete Spatial-Temporal System

In this section, we recast layer-wise backpropagation as a recursion in discrete space x (layer index) and time t (training iteration), laying the foundation for the continuous limit.

Theorem A.1 (Error Signal Recursion)**.** *In the discrete spatial–temporal interpretation, the error signal $\delta(x,t)$ satisfies*

$$\delta(X_{\max},t) = \frac{\partial \mathcal{L}}{\partial a(X_{\max},t)} \odot f'\big(z(X_{\max},t)\big), \quad \delta(x,t) = \big(W(x{+}1,t)^T\,\delta(x{+}1,t)\big) \odot f'\big(z(x,t)\big),$$

for $x = X_{\max}-1,\dots,1$.

Proof. By definition $\delta(x,t) = \partial\mathcal{L}/\partial z(x,t)$. At the boundary $x = X_{\max}$,

$$\delta(X_{\max},t) = \frac{\partial \mathcal{L}}{\partial a(X_{\max},t)} \cdot \frac{\partial a}{\partial z} = \frac{\partial \mathcal{L}}{\partial a} \odot f'(z)\,.$$

For $x < X_{\max}$ we apply the chain-rule and get

$$\delta(x,t) = \frac{\partial \mathcal{L}}{\partial z(x,t)} = \big(W(x{+}1,t)^T\,\partial\mathcal{L}/\partial z(x{+}1,t)\big)\odot f'(z(x,t)) = \big(W(x{+}1,t)^T\,\delta(x{+}1,t)\big)\odot f'(z(x,t)).$$

□

Corollary A.2 (Gradient Formulas)**.** *The parameter gradients satisfy*

$$\frac{\partial \mathcal{L}}{\partial W(x,t)} = \delta(x,t)\,a(x-1,t)^T, \qquad \frac{\partial \mathcal{L}}{\partial b(x,t)} = \delta(x,t).$$

Proof. We see that this immediately by $\partial z = a\,\partial W + \partial b$ and the definition of δ. □

These theorems serve as the discrete foundation for Section 3, enabling the PDE and SDE derivations.

A.3 PDE Representation in Continuous Limit

In this section, by letting the layer and time increments vanish, we derive PDEs describing the deterministic flow of parameters.

Theorem A.3 (Continuum PDE for Weight Evolution)**.** *As $\Delta t, \Delta x \to 0$, the discrete update $W(x,t+\Delta t) - W(x,t) = -\eta\,\delta(x,t)\,a(x-1,t)^T$ converges formally to the PDE*

$$\partial_t W(x,t) = -\eta\,\delta(x,t)\,f\big(z(x-1,t)\big)^T, \quad \delta(x,t) = \big(\partial_x^\dagger \delta\big)(x,t) \odot f'\big(z(x,t)\big),$$

where $\partial_x^\dagger$ denotes the backward difference operator.

Proof. We first express

$$\frac{W(x,t+\Delta t) - W(x,t)}{\Delta t} = -\eta\,\delta(x,t)\,a(x-1,t)^T.$$

Sending $\Delta t \to 0$ yields the time-derivative. Meanwhile replacing the backward recursion for δ by the adjoint of the forward difference gives the continuous spatial dependence. □

This PDE describes the mean-drift component of training and underlies our stochastic perturbations

A.4 SGD as a Matrix-Valued Itô SDE

In this section, we show that random mini-batch gradients introduce Brownian-like noise into the weight dynamics.

Theorem A.4 (SGD as Itô SDE)**.** *Under mini-batch noise, with variance parameter D, the weight update $W(x,t+\Delta t) - W(x,t) = -\eta\nabla_W\mathcal{L}(x,t) + \sqrt{2\eta D}\,\xi$ converges to the Itô SDE*

$$dW(x,t) = -\eta\,\nabla_W \mathcal{L}\,dt + \sqrt{2\eta D}\,d\mathcal{W}(x,t),$$

where $d\mathcal{W}$ is matrix-valued Brownian motion.

Proof. By central-limit scaling of the mini-batch noise we see that $\frac{1}{\sqrt{\Delta t}} \sum_i (\nabla \mathcal{L}_i - \nabla \mathcal{L}) \xrightarrow{d} \mathcal{N}(0, D)$, hence in the limit $\Delta t \to 0$ it becomes the Wiener increment $\sqrt{2\eta D}\, d\mathcal{W}$. □

This result justifies the isotropic noise term in the SDE.

A.5 Itô's Lemma for Singular Values

In this section, we compute the drift and diffusion contributions to each singular value under the matrix Itô SDE.

Lemma A.5 (Gradient and Laplacian of σ_k)**.** *Let $W = U\Sigma V^T$ be the SVD of $W \in \mathbb{R}^{m\times n}$ with $\Sigma_{kk} = \sigma_k > 0$. Then*

$$\nabla_W \sigma_k = u_k v_k^T, \qquad \Delta_W \sigma_k = \frac{m-n+1}{2\sigma_k} + \sum_{j\neq k} \frac{\sigma_k}{\sigma_k^2 - \sigma_j^2}.$$

Proof. The gradient is standard from matrix perturbation theory. The Laplacian follows by differentiating twice and using orthonormality of singular vectors. □

Theorem A.6 (Itô SDE for σ_k)**.** *Under $dW = A\, dt + \sqrt{2\epsilon}\, d\mathcal{W}$, the kth singular value obeys*

$$d\sigma_k = \Big(Tr((u_k v_k^T)^T A) + \epsilon\, \Delta_W \sigma_k \Big)\, dt + \sqrt{2\epsilon}\, d\beta_k,$$

where $d\beta_k = u_k^T\, d\mathcal{W}\, v_k$ is scalar Brownian motion.

Proof. We apply the general Itô formula

$$df(W) = \sum_{i,j} f_{ij}\, dW_{ij} + \tfrac{1}{2} \sum_{i,j,k,l} f_{ij,kl}\, dW_{ij} dW_{kl},$$

with $f(W) = \sigma_k(W)$, and use $dW_{ij} dW_{kl} = 2\epsilon\, \delta_{ik}\delta_{jl}\, dt$, together with the lemma above. □

These theorems form the basis for the interacting SDEs of singular values.

A.6 Mapping to Dyson Brownian Motion

In this section, we show that in the zero-gradient regime, squared singular values follow a Dyson-type interacting particle SDE.

Theorem A.7 (Dyson–SDE Identification)**.** *Let $\lambda_k = \sigma_k^2$. Then in the gradient-flat regime $\nabla \mathcal{L} \approx 0$,*

$$d\lambda_k = \Big(\eta D(m-n+3) + 2\eta D \sum_{j\neq k} \frac{\lambda_k}{\lambda_k - \lambda_j} \Big)\, dt + 2\sqrt{2\eta D\, \lambda_k}\, d\beta_k,$$

which after time-rescaling becomes the $\beta = 1$ Dyson Brownian motion $dY_k = \Big(\frac{m-n+3}{2} + \sum_{j\neq k} \frac{Y_k}{Y_k - Y_j} \Big)\, ds + 2\sqrt{Y_k}\, dW_k$.

Proof. Compute $d\lambda_k$ via Itô on $f(\sigma) = \sigma^2$, use the previous SDE, drop the gradient term, and choose $s = t/(2\eta D)$ so that the prefactors match exactly the canonical form. □

This theorem helps to explain the eigenvalue repulsion and spectral spreading.

A.7 Stationary Fokker–Planck and Gamma Law

We show that solving the steady-state Fokker–Planck PDE for one SDE yields a Gamma-family density.

Proposition A.8 (Stationary Density is Gamma-type)**.** *For the one-particle SDE* $d\lambda_t = (\alpha_0 - \beta_1\lambda_t)\,dt + \sqrt{8\eta D\,\lambda_t}\,dW_t$*, the stationary solution of the Fokker–Planck equation is*

$$p(\lambda) \propto \lambda^{\frac{\alpha_0}{4\eta D}-1} \exp\Big(-\tfrac{\beta_1}{4\eta D}\,\lambda\Big), \quad \lambda > 0.$$

Proof. Setting the time-derivative to zero, we get

$$0 = -\partial_\lambda\big[(\alpha_0 - \beta_1\lambda)p\big] + \tfrac{1}{2}\partial_\lambda^2\big[8\eta D\,\lambda p\big].$$

Integrating once under zero-flux boundary conditions and separating variables, we get $\frac{p'}{p} = (\frac{\alpha_0}{4\eta D} - 1)\frac{1}{\lambda} - \frac{\beta_1}{4\eta D}$, and exponentiate to obtain the Gamma-form. □

Obtaining a Gamma form, we see that this justifies the heavy-tail exponents observed in our experimental results.

A.8 Further Mathematical Analysis

In this section, we assemble classical random–matrix and integral–transform results that underpin our spectral SDE framework. First, we recall the Marchenko–Pastur and Tracy–Widom edge laws, which describe the untrained and boundary fluctuations of large random weight matrices. Then we turn to Hilbert transforms and stationary mean-field equations, which provide the macroscopic density needed in our Fokker–Planck analysis of singular-value dynamics.

A.8.1 Derivation of the Stochastic Term β

Proposition A.9 (Solving for the Stochastic Term)**.** *Given the stochastic differential equation for the temporal evolution of an eigenvalue* λ_k*, the stochastic term* β *can be isolated.*

Proof. We begin with the SDE describing the evolution of the eigenvalue λ_k, which includes terms for the gradient of the loss, eigenvalue repulsion, and a stochastic component driven by β:

$$\frac{\partial \lambda_k}{\partial t} = -\sqrt{\eta}\lambda_k \frac{\partial L}{\partial W} + \sum_{j\neq k} \frac{\lambda_k}{\lambda_k - \lambda_j} + \sqrt{\eta\lambda_k}\beta \tag{1}$$

$$\frac{\partial \lambda_k}{\partial t} + \sqrt{\eta}\lambda_k \frac{\partial L}{\partial W} - \sum_{j\neq k} \frac{\lambda_k}{\lambda_k - \lambda_j} = \sqrt{\eta\lambda_k}\beta \tag{2}$$

Finally, we divide by the coefficient of β, which is $\sqrt{\eta\lambda_k}$, to obtain the expression for the stochastic term:

$$\beta = \frac{1}{\sqrt{\eta\lambda_k}} \left(\frac{\partial \lambda_k}{\partial t} + \sqrt{\eta}\lambda_k \frac{\partial L}{\partial W} - \sum_{j\neq k} \frac{\lambda_k}{\lambda_k - \lambda_j} \right) \tag{3}$$

This completes the derivation. □

A.8.2 Scaling Limits and Tracy-Widom Limit

At initialization, our weight matrices follow Wishart (or MP) statistics. Understanding the bulk and edge of this spectrum is essential both to verify our isotropic SGD noise reproduces classical limits and to identify regimes where empirically observed 'bulk+tail' structured deviations occur.

Lemma A.10. *Bulk and Edge of the Marchenko-Pastur Law We first let* $M = \frac{1}{n}XX^T$ *be an* $m \times m$ *Wishart matrix with* $X \in \mathbb{R}^{m\times n}$ *having i.i.d. entries of variance* 1*. As* $m, n \to \infty$ *with*

$m/n \to \gamma \in (0, 1]$, *the empirical spectral distribution of* M *converges to the Marchenko–Pastur density*

$$\rho_{MP}(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\gamma x}, \quad x \in [\lambda_-, \lambda_+],$$

where

$$\lambda_\pm = (1 \pm \sqrt{\gamma})^2.$$

This proof is presented in [17], where the Stieltjes transform of M is utilized. In all, this bulk law justifies our use of MP fits at $t = 0$, and sets the stage for tracking departures under SGD noise.

Lemma A.11 (Edge Scaling Constants)**.** *Under the same regime, we let* $\lambda_{(1)}$ *be the largest eigenvalue of* M*. We define*

$$\mu_{m,n} = \lambda_+, \qquad \sigma_{m,n} = (\lambda_+)^{1/2} \, \frac{(1 + \gamma^{-1/2})^{1/3}}{n^{2/3}}.$$

Then the centered and scaled variable

$$\chi_{m,n} = \frac{\lambda_{(1)} - \mu_{m,n}}{\sigma_{m,n}}$$

has fluctuations on order one as $m, n \to \infty$*.*

Corollary A.12 (Tracy–Widom F_1 Limit)**.** *It was shown in [31] and this was proven by expressing the gap probability as a Fredholm determinant of the Airy kernel, thereby yielding the limit.*

$$\lim_{m,n\to\infty} \mathbb{P}\big(\chi_{m,n} \le s\big) = F_1(s),$$

where F_1 *is the Tracy–Widom distribution for* $\beta = 1$ *(real symmetric ensembles).*

Definition A.13 (Airy Kernel and Process)**.** *The* Airy kernel *is*

$$K_{\mathrm{Ai}}(x, y) = \frac{\mathrm{Ai}(x)\,\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\,\mathrm{Ai}(y)}{x - y},$$

and the Airy process $\{\mathcal{A}(t)\}$ *is the determinantal process with kernel*

$$K_{\mathrm{Ai}}(t_1, \xi_1;\, t_2, \xi_2) = \int_0^\infty e^{-u(t_2 - t_1)} \mathrm{Ai}(\xi_1 + u)\,\mathrm{Ai}(\xi_2 + u)\, du.$$

Remark A.14. *The largest eigenvalue fluctuations of Dyson's Brownian motion (with* $\beta = 1$*) also converge to the Airy process, giving a dynamical Tracy–Widom law for* $\lambda_{(1)}(t)$ *under appropriate time scaling.*

A.8.3 Hilbert Transform and Stationary Density

To derive the macroscopic spectral density under our isotropic SDE, we solve a stationary Fokker–Planck equation via Hilbert transforms. The lemma below gives a closed-form for power-law inputs, enabling the Gamma-like stationary density.

Lemma A.15 (Hilbert Transform of Power Law Densities)**.** *If* $\rho(x) = C\,x^\alpha$ *on* $[0, R]$*, then its finite-interval Hilbert transform is*

$$H[\rho](\lambda) = \frac{1}{\pi} PV \left(\int_0^R \frac{C\,x^\alpha}{x - \lambda}\, dx \right) = C\,\lambda^\alpha \cot(\pi\alpha) + O(1),$$

for $\lambda \in (0, R)$ *and* $\alpha \notin \mathbb{Z}$*.*

Corollary A.16 (Stationary Density at Large r)**.** *Under the quadratic mean-field potential and isotropic noise, the large-*r *stationary* $\rho_{st}(\lambda)$ *solving*

$$\eta D(m - n + 3) - 2\pi\eta D \, \frac{d}{d\lambda}\big(\lambda\, H[\rho_{st}](\lambda)\big) = 0$$

behaves to leading order like

$$\rho_{st}(\lambda) \propto \lambda^{\frac{1}{4}(m-n+3)-1},$$

recovering the Gamma-type density in the effective single-particle Fokker–Planck.

This provides the explicit stationary spectrum that emerges from our isotropic SDE.

A.8.4 Anisotropic Analysis

Thus far, our analysis has assumed that the random fluctuations in SGD are isotropic meaning every direction in parameter space experiences the same noise strength. In practice, however, noise can be highly direction-dependent—layers, singular modes, or even individual parameters often see very different variance due to batch structure, learning-rate schedules, or architecture specifics. Accounting for this anisotropy is crucial if we hope to predict not only the locations of singular values, but also the relative spreading and alignment of singular vectors over training. The following proposition shows how the general Itô-lemma approach naturally incorporates a full covariance structure $\Sigma(W,t)$, yielding additional second–derivative corrections to the drift and a directionally weighted diffusion term. This richer SDE then serves as the foundation for a non-homogeneous Dyson-type PDE in the mean-field limit, capable of capturing empirically observed anisotropic spectral evolution. We carry out the derivations below, and we leave experimentation for anisotropic analysis for future work.

We have for anisotropic analysis,

$$dW = -\,\eta\,\nabla_W \mathcal{L}\,dt \;+\; B(W,t)\,d\mathcal{B}_t,$$

where $B(W,t)B(W,t)^T = 2\eta\,\Sigma(W,t)$ and $d\mathcal{B}_t$ is our standard matrix Wiener process.

Proposition A.17 (Anisotropic Noise - Changes to SDE for Singular Values)**.** *We let $W(t) = U(t)\,\Sigma(t)\,V(t)^T$ be the SVD of W, and we denote $\sigma_k(t)$ the k^{th} singular value. Then under the dynamics above, we have*

$$d\sigma_k = \Big\langle \nabla_W \sigma_k\,,\, -\eta\,\nabla_W \mathcal{L} \Big\rangle\,dt + \frac{1}{2}\sum_{i,j,p,q} \frac{\partial^2 \sigma_k}{\partial W_{ij}\partial W_{pq}} \big[2\eta\,\Sigma(W,t)\big]_{ip,jq}\,dt + \big\langle \nabla_W \sigma_k\,,\, B(W,t)\,d\mathcal{B}_t \big\rangle.$$

Since $\nabla_W \sigma_k = u_k v_k^T$ and $\dfrac{\partial^2 \sigma_k}{\partial W_{ij}\partial W_{pq}}$ is known from matrix-perturbation theory, the drift becomes

$$u_k^T\big(-\eta \nabla_W \mathcal{L}\big) v_k \;+\; \eta\,\mathrm{Tr}\Big[\Sigma(W,t)\,\nabla_W^2 \sigma_k\Big],$$

and the diffusion term is $\sqrt{2\eta}\,\langle u_k v_k^T,\ \sqrt{\Sigma(W,t)}\,d\mathcal{B}_t\rangle$.

Proof. For a scalar $f(W)$, we know by Ito's Lemma that

$$df \;=\; \sum_{i,j} \frac{\partial f}{\partial W_{ij}}\,dW_{ij} \;+\; \frac{1}{2}\sum_{i,j,p,q} \frac{\partial^2 f}{\partial W_{ij}\,\partial W_{pq}}\;dW_{ij}\,dW_{pq}.$$

Now, we proceed to substitute dW_{ij}. We get that

$$dW_{ij} = -\eta\,(\nabla_W \mathcal{L})_{ij}\,dt \;+\; \sum_{\alpha,\beta} B_{ij,\alpha\beta}\,d\mathcal{B}_{\alpha\beta}.$$

Hence

$$\sum_{i,j} \frac{\partial f}{\partial W_{ij}}\,dW_{ij} = \underbrace{\sum_{i,j} \frac{\partial f}{\partial W_{ij}}\big[-\eta(\nabla_W \mathcal{L})_{ij}\big]}_{=\langle \nabla_W f,\, -\eta\,\nabla_W \mathcal{L}\rangle}\;dt \;+\; \sum_{i,j,\alpha,\beta} \frac{\partial f}{\partial W_{ij}}\,B_{ij,\alpha\beta}\,d\mathcal{B}_{\alpha\beta}.$$

We see that only the noise part contributes second-order terms, hence we have

$$dW_{ij}\,dW_{pq} = \Big(\sum_{\alpha,\beta} B_{ij,\alpha\beta}\,d\mathcal{B}_{\alpha\beta}\Big)\Big(\sum_{\gamma,\delta} B_{pq,\gamma\delta}\,d\mathcal{B}_{\gamma\delta}\Big) = \sum_{\alpha,\beta} B_{ij,\alpha\beta}\,B_{pq,\alpha\beta}\;dt = 2\eta\,\Sigma_{ip,jq}(W,t)\,dt.$$

Thus, we have

$$\frac{1}{2}\sum_{i,j,p,q} \frac{\partial^2 f}{\partial W_{ij}\,\partial W_{pq}}\,dW_{ij}\,dW_{pq} = \underbrace{\eta \sum_{i,j,p,q} \frac{\partial^2 f}{\partial W_{ij}\,\partial W_{pq}}\,\Sigma_{ip,jq}(W,t)}_{=\frac{1}{2}\sum(\partial^2 f)\,[2\eta\,\Sigma]_{ip,jq}}\;dt.$$

Now, we proceed to group all the dt terms, giving us the drift term below

$$\langle \nabla_W f,\, -\eta\, \nabla_W \mathcal{L} \rangle \;+\; \frac{1}{2} \sum_{i,j,p,q} \frac{\partial^2 f}{\partial W_{ij} \partial W_{pq}} \big[2\eta\, \Sigma(W,t)\big]_{ip,jq},$$

and the remaining stochastic term is the martingale term given by

$$\sum_{i,j,\alpha,\beta} \frac{\partial f}{\partial W_{ij}}\, B_{ij,\alpha\beta}\, d\mathcal{B}_{\alpha\beta} = \big\langle \nabla_W f,\; B(W,t)\, d\mathcal{B}_t \big\rangle.$$

Finally, from matrix perturbation theory, we know that

$$\nabla_W \sigma_k = u_k\, v_k^T, \quad \frac{\partial^2 \sigma_k}{\partial W_{ij}\, \partial W_{pq}} \;=\; \big[\nabla_W^2 \sigma_k\big]_{ij,pq},$$

so the final SDE is

$$d\sigma_k = \underbrace{\langle u_k v_k^T,\, -\eta\, \nabla_W \mathcal{L} \rangle}_{\text{drift from loss}}\, dt + \frac{1}{2} \sum_{i,j,p,q} \big[\nabla_W^2 \sigma_k\big]_{ij,pq} \big[2\eta\, \Sigma(W,t)\big]_{ip,jq}\, dt + \big\langle u_k v_k^T,\; B(W,t)\, d\mathcal{B}_t \big\rangle,$$

as proposed. □

Lemma A.18. *(Estimating the Diffusion Constant for Stationary Distribution Fitting)*
In order to connect our theoretical diffusion coefficient D to observable quantities during training, we employ a simple dimensional-analysis argument. The diffusion term in our singular-value SDE has units of (singular-value)^2 per unit time, so D must scale like the variance of singular-value changes divided by the time step. Empirically, the broadening of the spectrum is characterized by the gap between the largest mode and a representative central mode—here taken as $\sigma_{\max} - \sigma_{\text{med}}$. Over t_b batch updates, this gap typically increases by an amount on the order of its own magnitude. Matching units then gives

$$[D] = \frac{L^2}{s} \mapsto \frac{(\sigma_{max} - \sigma_{med})^2}{t_b}$$

where σ_{max} corresponds the maximum singular value, σ_{med} corresponds to the median singular value, and t_b is the time (which is represented as the batch update number in our spatiotemporal interpretation).

Lemma A.19. *(Estimating the Noise Constant β_1 for Stationary Distribution Fitting)*
*Letting $L(w) = \frac{1}{N} \sum_{i=1}^N L_i(W)$ be the loss function, and defining the **batch gradient** (true gradient) as:*

$$\nabla L(w) = \frac{1}{N} \sum_{i=1}^{N} \nabla L_i(W)$$

*and **minibatch gradient** for a randomly sampled minibatch S_t of size B as:*

$$\nabla L_{S_t}(w) = \frac{1}{B} \sum_{j \in S_t} \nabla L_j(w)$$

*we model the **SGD noise** for a minibatch S_t as the difference:*

$$\beta_1(W, S_t) = \nabla L_{S_t}(W) - \nabla L(W)$$

Implying

$$\|\beta_1(W, S_t)\|^2 = \|\nabla L_{S_t}(W) - \nabla L(w)\|^2$$

*The minibatch gradient is an unbiased estimator: $\mathbb{E}_{S_t}[\nabla L_{S_t}(w)] = \nabla L(w)$. The **variance of the minibatch gradient**, which is the formal measure of SGD noise, is given by the expected squared norm of the noise term:*

$$Var(\nabla L_{S_t}(w)) \equiv \mathbb{E}_{S_t}\left[\|\nabla L_{S_t}(w) - \nabla L(w)\|^2\right] \equiv \mathbb{E}_{S_t}[\|\beta_1(W, S_t)\|^2$$

Thus we use this empirically determined value of β_1 for our fits for the stationary distributions.

B Algorithm Details

Algorithm 1 Predicting Singular–Value Dynamics via Bootstrapped Drift

Input: $W^{(0)} \in \mathbb{R}^{m\times n}$, η, T, k
Output: $\{\sigma^{(t)}\}_{t=0}^{T}$

$[U, \Sigma, V] \leftarrow \mathrm{svd}(W^{(0)})$; $\quad U_k \leftarrow U_{:,1:k}$, $\sigma \leftarrow \mathrm{diag}(\Sigma)_{1:k}$, $V_k \leftarrow V_{:,1:k}$
for $t = 0, \ldots, T-1$ **do**
  $G \leftarrow -\eta \nabla_W \ell^{(t)}$
  $M \leftarrow U_k^\top G V_k$
  for $i = 1, \ldots, k$ **do**
    $\Delta\sigma_i \leftarrow M_{ii}$
    $du_i,\ dv_i \ = \ \sum_{j\neq i} \Big(\frac{M_{ji}}{\sigma_i - \sigma_j + \varepsilon} + \frac{M_{ij}}{\sigma_i + \sigma_j + \varepsilon} \Big) \, (U_k[:, j],\ V_k[:, j])$
    $\sigma_i \leftarrow \max(\sigma_i + \Delta\sigma_i,\ \delta)$
    $\widetilde{U}_i \leftarrow U_k[:, i] + du_i, \quad \widetilde{V}_i \leftarrow V_k[:, i] + dv_i$
  end for
  $U_k, V_k \leftarrow \mathrm{orth}\big([\widetilde{U}_1, \ldots, \widetilde{U}_k]\big),\ \mathrm{orth}\big([\widetilde{V}_1, \ldots, \widetilde{V}_k]\big)$
  Align signs of U_k, V_k columns with previous
end for
return $\{\sigma^{(t)}\}$

B.1 Computational Complexity Analysis

Algorithm 1 offers significant computational advantages over naive approaches that recompute the full SVD at each time step. We analyze the complexity for an $m \times n$ weight matrix over T time steps, tracking the top k singular values.

Initial Setup. The initial SVD computation (Line 3) requires $O(\min(m^2 n, mn^2))$ operations, which is performed only once.

Per-Timestep Complexity. For each of the T time steps, the algorithm performs:

- **Gradient computation**: $O(G)$ operations, where G depends on the specific loss function and network architecture.
- **Projection**: Computing $M = U_k^T G V_k$ requires $O(kmn)$ operations.
- **Singular value updates**: For each of the k singular values, the drift computation involves $O(k)$ operations and the singular vector updates require $O(k \max(m, n))$ operations, yielding $O(k^2 \max(m, n))$ total.
- **Orthogonalization**: The Gram-Schmidt orthogonalization step costs $O(k^2 \max(m, n))$ operations.

Total Complexity. The overall computational complexity is:

$$O\big(\min(m^2 n, mn^2) + T(G + kmn + k^2 \max(m, n))\big)$$

Efficiency Gains. When $k \ll \min(m, n)$ (typically $k \leq 10$ for the leading modes), our algorithm achieves substantial speedups compared to naive full SVD recomputation at each step, which would require $O(T \min(m^2 n, mn^2))$ operations. For large matrices where $m, n \gg k$, the per-timestep cost reduces from $O(\min(m^2 n, mn^2))$ to $O(kmn + k^2 \max(m, n))$, representing a factor of $\Theta(\min(m, n)/k)$ improvement in the SVD-related computations.

C Additional Experimental Details

SGD. Our use of SGD follows the classic Ornstein–Uhlenbeck approximation for constant-rate noise [15], while observed anisotropies in batch-size and learning-rate interactions [8] directly inform

our extension to non-isotropic noise. Choosing a quadratic mean-field potential for large-width spectral dynamics is supported by recent convergence results in overparameterized models [3, 27].

GPT2. We use the nanoGPT implementation [11] which follows the transformer decoder-only architecture with four transformer layers, four attention heads per layer, and 256-dimensional embeddings. The learning rate starts at $5 * 10^{-4}$ with cosine decay to $5 * 10^{-5}$. We use a batch size of 12 sequences of 256 tokens each.

Vision Transformer (ViT). ViT is configured with two encoder layers, four attention heads, and a 256-dimensional embedding. Inputs ($H \times H$) are segmented into patches ($P \times P$), transformed by standard Transformer blocks (FFN expansion ratio $\alpha = 2$), and classified via a linear head initialized as $w \sim \mathcal{N}(0, 1/\sqrt{H_{dim}})$. We set $(H, P) = (28, 7)$ for MNIST and $(32, 8)$ for CIFAR-100.

Multilayer Perceptron (MLP). Our MLP comprises three hidden layer of 1024 dimensions. Weight matrices are initialized from $\mathcal{N}(0, 1/\text{fan}_{\text{in}})$, with biases initialized to zero.

Other Considered Models. ResNet architecture [7] is not used as it consists mostly of convolutional layers with structured weight sharing patterns making some spectral properties less interpretable for understanding loss landscapes.

D Additional Experimental Results

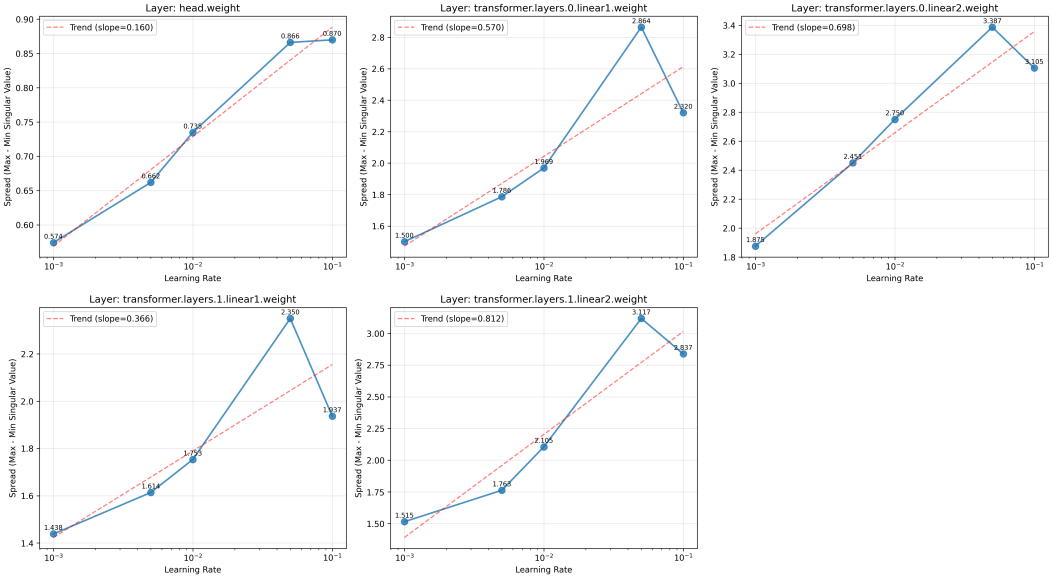

Figure 4: Spread of singular values (max–median) versus learning rate for different vision transformer weight matrices, with red dashed least-squares trends and slopes indicating sensitivity.

Across all layers, increasing the learning rate from 10^{-3} to 10^{-1} amplifies the spectral spread, indicating that higher noise levels drive greater anisotropy in the weight matrix. Moreover, the fitted trend-line slopes reveal that the second feed-forward projection in each layer is most sensitive to learning-rate scaling. In particular, in layer 1 (slope ≈ 0.81)—whereas the output head's weights remain comparatively stable (slope ≈ 0.16). These results show that isotropic SGD induces layer-dependent spectral broadening, with deeper feed-forward blocks experiencing the strongest effect. When a certain critical learning rate is hit, we see that the spread decreases, indicating more uniformity in singular values, potentially implying that fewer features are being learnt by the model, thus meriting further investigation to understand this phenomenon.