The Underlying Scaling Laws and Universal Statistical Structure of Complex Datasets

Noam Levi[†‡] **and Yaron Oz**[†]
[†]Raymond and Beverly Sackler School of Physics and Astronomy
Tel-Aviv University
Tel-Aviv 69978, Israel
[‡]École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
noam@mail.tau.ac.il

Abstract

We study universal traits which emerge both in real-world complex datasets, as well as in artificially generated ones. Our approach is to analogize data to a physical system and employ tools from statistical physics and Random Matrix Theory (RMT) to reveal their underlying structure. We focus on the feature-feature covariance matrix, analyzing both its local and global eigenvalue statistics. Our main observations are: *(i)* The power-law scalings that the bulk of its eigenvalues exhibit are vastly different for uncorrelated normally distributed data compared to real-world data, *(ii)* this scaling behavior can be completely modeled by generating Gaussian data with long range correlations, *(iii)* both generated and real-world datasets lie in the same universality class from the RMT perspective, as chaotic rather than integrable systems, *(iv)* the expected RMT statistical behavior already manifests for empirical covariance matrices at dataset sizes significantly smaller than those conventionally used for real-world training, and can be related to the number of samples required to approximate the population power-law scaling behavior, *(v)* the Shannon entropy is correlated with local RMT structure and eigenvalues scaling, is substantially smaller in strongly correlated datasets compared to uncorrelated ones, and requires fewer samples to reach the distribution entropy. These findings show that with sufficient sample size, the Gram matrix of natural image datasets can be well approximated by a Wishart random matrix with a simple covariance structure, opening the door to rigorous studies of neural network dynamics and generalization which rely on the data Gram matrix.

1 Introduction

Natural, or real-world, images are expected to follow some underlying distribution, which can be arbitrarily complex, and to which we have no direct access to. This distribution could have infinitely many nonzero moments, with varying relative importance compared to one another. In practice, we only have access to a very small subset of samples from the underlying distribution, which can be parameterized as $X \in \mathbb{R}^{d\times M}$, where d is the dimension of each image vector and M is the number of samples. The first moment of the data can always be set to 0, since we can remove the mean from each sample, without losing information regarding the distribution. The second moment, however, cannot be set to 0, and holds valuable information. This observation motivates the study of the empirical covariance (Gram) matrix, $\Sigma_M = \frac{1}{M} X X^T$.

The properties of Σ_M in real world data are entirely unknown a priori, as we do not know how to parameterize the process which generated natural images. Nevertheless, interesting observations

Preprint. Under review.

have been made. Empirical evidence shows that the spectrum of Σ_M for various datasets can be separated into a set of large eigenvalues ($\mathcal{O}(10)$), a bulk of eigenvalues which decay as a power law $\lambda_i \sim i^{-1-\alpha}$ [1, 2] and a large tail of small eigenvalues which terminates at some finite index n. Since the top eigenvalues represent the largest overlapping properties across different samples, these are not simply interpreted without more information on the underlying distribution. The bulk of the eigenvalues, however, can be understood as representing the correlation structure of different features amongst themselves, and has been key to understanding the emergence of neural scaling laws [3, 4].

In this work, we study both the scaling laws present in natural datasets, and their spectral statistics, with the goal of obtaining a universal, analytically tractable model for real world Gram matrices, regardless of their origins. While this may not be feasible for any Σ_M, fortunately, the standard datasets used today are high dimensional and contain many samples, a ubiquitous regime found in complex systems, and typically studied using tools from Random Matrix Theory (RMT).

RMT is a powerful tool for describing the spectral statistics of complex systems. It is particularly useful for systems that are chaotic but also have certain coherent structures. The theory predicts universal statistical properties, provided that the underlying matrix ensemble is large enough to sufficiently fill the space of all matrices with a given symmetry, a property known as ergodicity [5]. Ergodicity has been observed in a variety of systems, including chaotic quantum systems [6–8], financial markets, nuclear physics and many others [9–11]. To demonstrate that a similar universal structure is also observed for correlation matrices resulting from datasets, we will employ several diagnostic tools widely used in the field of quantum chaos. We will analyze the global and local spectral statistics of empirical covariance matrices generated from three classes of datasets: (i) Data generated by sampling from a normal distribution with a specific correlation structure for its features, (ii) Uncorrelated Gaussian Data (UGD), (iii) Real-world datasets composed of images, at varying levels of complexity and resolution. Our research aims to answer the following questions:

- Is power-law scaling a universal property across real-world datasets?; what determines the scaling exponent and what properties should an analytic model of the dataset have, in order to follow the same scaling?
- What are the universal properties of datasets that can be gleaned from the empirical covariance matrix and how are they related to local and global statistical properties of RMT?
- How to quantify the extent to which complex data is well characterized by its Gram matrix?
- What, if any, are the relations between datasets scaling, entropy and statistical chaos diagnostics?

Our primary contributions are:

1. We find that power-law scaling appears across various datasets. It is governed by a single scaling exponent α, and its origin is the strength of correlations in the underlying population matrix [1]. We accurately recover the behavior of the eigenvalue bulk of real-world datasets using Wishart matrices with the singular values of a Toeplitz matrix [13] as its covariance. We dub these *Correlated Gaussian Datasets* (CGDs).
2. We show that generically, the bulk of eigenvalues' distribution and spacings are well described by RMT predictions, verified by diagnostic tools typically used for quantum chaotic systems. This means that the CGD model is a correct proxy for real-world data Gram matrices.
3. We find that the effective convergence of the empirical covariance matrix as a function of the number of samples correlates with the corresponding RMT description becoming a good description of the statistics and the eigenvalues scaling.
4. The Shannon entropy is correlated with the local RMT structure and the eigenvalues scaling, and is substantially smaller in strongly correlated datasets compared to uncorrelated data. Additionally, it requires fewer samples to reach the distribution entropy.

2 Background and Related Work

Neural Scaling Laws Neural scaling laws are a set of empirical observations that describe the relationship between the size of a neural network, dataset, compute power, and its performance.

[1] There are systems which display multiple correlation scales, showing several bulk exponents [12].

These laws were first proposed by Kaplan et al. [3] and have since been confirmed by a number of other studies [4, 14] and studied further in [15–20]. The main finding of neural scaling laws is that the test loss of a neural network scales as a power-law with the number of parameters in the network. This means that doubling the number of parameters roughly reduces the test loss by 2^{α}. However, this relationship does not persist indefinitely, and there is a point of diminishing returns beyond which increasing the number of parameters does not lead to significant improvements in performance. One of the key challenges in understanding neural scaling laws is the complex nature of the networks themselves. The behavior of a neural network (NN) is governed by a large number of interacting parameters, making it difficult to identify the underlying mechanisms that give rise to the observed scaling behavior, and many advances have been made by appealing to the RMT framework.

Random Matrix Theory RMT is a branch of mathematics that was originally developed to study the properties of large matrices with random entries. It is particularly suited to studying numerous realizations of the same system, where the number of realizations $M \to \infty$, the dimensions of the system $d \to \infty$, and the ratio between the two tends to a constant $d/M \to \gamma \leq 1, \gamma \in \mathbb{R}^{+}$. Results from RMT calculations have been applied to a wide range of problems in Machine Learning (ML), beyond the scope of neural scaling laws, including the study of nonlinear ridge regression [21], random Fourier feature regression [22], the Hessian spectrum [23], and weight statistics [24, 25]. For a review of some of the recent developments, we refer the reader to Couillet and Liao [26] and references therein.

Universality Considerable work has been dedicated to the concept of universality, i.e. that certain features are shared between seemingly disparate systems, when the systems are sufficiently large. For instance, spectra that are generated by different dynamical processes may have similar distributions [27–30]. Universality is powerful since it often happens that System A's complex structure is difficult to analyze, and can be explained by system B, which lies in the same universality class, and is much easier to study. In our work, system A represents real-world datasets with unknown statistics generated from a complex process, while system B is our CGD, whose Gram matrix is a simple Wishart matrix. The fact that real world datasets fall in the same universality class as CGD allows us to replace its complex covariance matrix by the simple CGD one, while retaining the information encoded in its spectrum.

3 Correlations and power-law Scaling

In this section, we analyze the feature-feature covariance matrix for datasets of varying size, complexity, and origin. We consider real-world as well as correlated and uncorrelated gaussian datasets, establish a power-law scaling of their eigenvalues, and relate it to a correlation length.

3.1 Feature-Feature Empirical Covariance Matrix

We consider the data matrix $X_{ia} \in \mathbb{R}^{d \times M}$, constructed of M columns, each corresponding to a single sample, composed of d features. In this work, we focus on the empirical feature-feature covariance matrix, defined as

$$\Sigma_{ij,M} = \frac{1}{M} \sum_{a=1}^{M} X_{ia} X_{aj} \in \mathbb{R}^{d \times d} \,. \tag{1}$$

Intuitively, the correlations between the different input features, X_{ia}, should be the leading order characteristic of the dataset. For instance, if the X_{ia} are pixels of an image, we may expect that different pixels will vary similarly across similar images. Conversely, the mean value of an input feature is uninformative, and so we will assume that our data is centered in a pre-processing stage.

A random matrix ensemble is a probability distribution on the set of $d \times d$ matrices that satisfy specific symmetry properties, such as invariance under rotations or unitary transformations. In order to study Eq. (1) using the RMT approach, we define $\Sigma_{ij,a}$ as a single sample realization of the population random matrix ensemble Σ_{ij}, and thus $\Sigma_{ij,M}$ is the empirical ensemble average, i.e. $\Sigma_M = \langle \Sigma_a \rangle_{a \in M} = \frac{1}{M} \sum_{a=1}^{M} \Sigma_a$ approximating the limits of $M \to \infty, d \to \infty$. If M and d are sufficiently large, the statistical properties of Σ_M will be determined entirely by the underlying symmetry of the ensemble. We refer to this scenario as the "RMT regime".

3.2 Data Exploration

We study the following real-world datasets: MNIST [31], FMNIST [32], CIFAR10 [33], Tiny-IMAGENET [34], and CelebA [35] (downsampeld to 109 × 89 in grayscale). We proceed to center and normalize all the datasets in the pre-processing stage, to remove the uninformative mean contribution. The uncorrelated gaussian data is represented by a data matrix $X_{ia} \in \mathbb{R}^{d\times M}$, where each column is drawn from a jointly normal distribution $\mathcal{N}(0, I_{d\times d})$. We then construct the empirical covariance matrix $\Sigma_M = \frac{1}{M}\sum_{a=1}^{M} X_{ia}X_{aj} \in \mathbb{R}^{d\times d}$. To generate correlated gaussian data, we repeat the same process, changing the sample distribution such to $\mathcal{N}(0, \Sigma_{d\times d})$, where we choose a specific form for Σ which produces feature-feature correlations with and includes a natural cut-off scale, as

$$\Sigma_{ij}^{\text{Toe}} = S, \qquad T = \boldsymbol{I}_{ij} + c|i-j|^{\alpha} = U^{\dagger}SV, \qquad \alpha, c \in \mathbb{R}. \tag{2}$$

The matrix Σ_{ij}^{Toe} is a positive semi definite diagonal matrix of singular values S constructed from T, a full-band Toeplitz matrix. The sign of α dictates whether correlations decay (negative) or intensify (positive) with distance along a one-dimensional feature space[2].

3.3 Correlations Determine the Noise to Data Transition

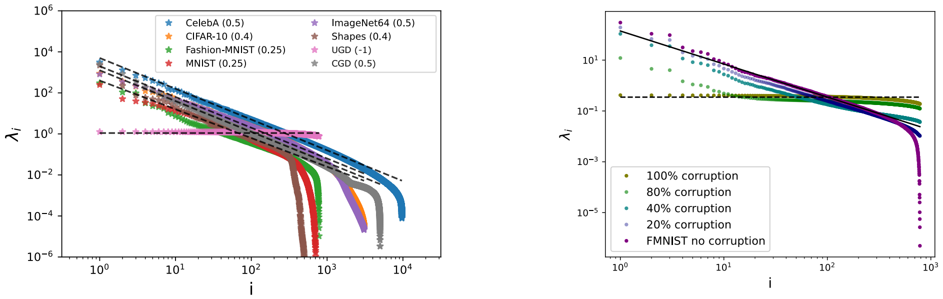

Figure 1: **Left:** Scree plot of $\Sigma_{ij,M}$ for several different vision datasets, as well as for UGD and a CGD with fixed α. Here, the number of samples is taken to be the entire dataset for each real-world dataset, and $M = 50$k for the gaussian data, where we set $c = 1$. We see a clear scaling law for the eigenvalue bulk as $\lambda_i \propto i^{-1-\alpha}$ where all real-world datasets display $\alpha \leq 1/2$. **Right:** The power-law scaling parameter α value can be tuned from $\alpha = 1/4$ to $\alpha = -1$ by corrupting the FMNIST dataset with a varying amount of normally distributed noise.

We begin by reproducing and extending some of the results from Maloney et al. [4]. In Fig. 1, we show the $\Sigma_{ij,M}$ eigenvalue scaling for the different classes of data (*i.e.* real-world, UGD and CGDs). We find that for all datasets, the eigenvalues bulk scales as a power-law

$$\lambda_i \propto i^{-1-\alpha}, \quad \alpha \in \mathbb{R}, \qquad i = 10, \ldots d_{\text{bulk}}\,, \tag{3}$$

where $i = 10$ is approximately where the scaling law behavior begins and d_{bulk} is the effective bulk size, where the power-law abruptly terminates. We stress that this behavior repeats across all datasets, regardless of origin and complexity.

The value of α can be readily explained in terms of correlations within our CGD model. Taking the Laplace Transform of the second term in Eq. (2), the bulk spectrum is given by Appendix B as

$$\lambda_i^{\text{bulk}} = c \cdot \Gamma(\alpha+1)\left(\frac{d}{i}\right)^{1+\alpha}, \tag{4}$$

where $\Gamma(x)$ is the Gamma function. This implies that the value of α determines the strength of correlations in the original data covariance matrix. For real-world data, we consistently find that $\alpha > 0$, which corresponds to increasing correlations between different features. In contrast, for UGD, the value of $\alpha \sim -1$, and the power-law behavior vanishes. Interpolating between UGD, and real-world-data, the CGD produces a power-law scaling, which can be tuned from $-1 < \alpha \leq 0$, in

[2] Correlation strength which grows with distance is a hallmark of some one-dimensional physical systems, such as the Coulomb and Riesz gases [36, 37], which display an inverse power-law repulsion, while decaying correlations are common in the 1-d Ising model [38].

the case of decaying long range correlations, or $0 \leq \alpha < \infty$ for increasing correlations, to match any real-world dataset we examined. Lastly, we can extend this statement further and verify the transition from correlated to uncorrelated features by corrupting a real-world dataset (FMNIST) and observing the continuous deterioration of the power-law from $\alpha \sim 1/4$ to $\alpha = -1$, implying that the CGD can mimic the bulk behavior of both clean and corrupted data.

4 Global and Local Statistical Structure

4.1 Random Matrix Theory

In this section, we move on from the eigenvalue scaling, to their statistical properties. We begin by describing the RMT diagnostic tools, often used to characterize RMT ensembles, with which we obtain our main results. We define the matrix ensemble under investigation, then provide an overview of each diagnostic, concluding with a summary of results for the specific matrix ensemble which both real-world and gaussian datasets converge to.

We interpret Σ_M for real world data as a single realization, drawn from the space of all possible Gram matrices which could be constructed from sampling the underlying population distribution. In that sense, Σ_M itself is a random matrix with an unknown distribution. For such a random matrix, there are several universality classes, which depend on the strength of correlations in the underlying distribution. These range from extremely strong correlations, which over-constrain the system and lead to the so called Poisson ensemble [39], to the case of no correlations, which is equivalent to sampling independent elements from a normal distribution, represented by the Gaussian Orthogonal Ensemble (GOE) [40]. These classes are the only ones allowed by the symmetry of the matrix XX^T, provided that the number of samples and the number of features are both large. Since the onset of the RMT regime depends on the population statistics, it is a priori unknown. Determining if real data Gram matrices converge to an RMT class, and to which one they converge to at finite sample size would inform the correct way to model real world Gram matrices.

Below we review the tools used in our analysis. While we provide an overview of each diagnostic, we refer the reader to Tao [41], Kim et al. [42] for a more comprehensivereview. We then apply these tools to gain insights into the statistical structure of the datasets.

Spectral Density: The empirical spectral density of a matrix Σ is defined as,

$$\rho_\Sigma(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \delta(\lambda - \lambda_i(\Sigma)), \tag{5}$$

where δ is the Dirac delta function, and the $\lambda_i(\Sigma), i = 1, ..., n$, denote the n eigenvalues of Σ, including multiplicity. The limiting spectral density is defined as the limit of Eq. (5) as $n \to \infty$.

Level Spacing Distribution and r-statistics: The level spacing distribution measures the probability density for two adjacent eigenvalues to be in the spectral distance s, in units of the mean level spacing Δ. The procedure for normalizing all distances in terms of the local mean level spacing is often referred to as unfolding. We unfold the spectrum of the empirical covariance matrix $\Sigma_M(\rho)$ by standard methods [42], reviewed in Appendix A. Ultimately, the transformation $\lambda_i \to e_i = \tilde{\rho}(\lambda_i)$ is performed such that e_i shows an approximately uniform distribution with unit mean level spacing. Once unfolded, the level spacing is given by $s_i = e_{i+1} - e_i$, and its probability density function $p(s)$ is measured.

The distribution $p(s)$ captures information about the short-range spectral correlations, demonstrating the presence of level repulsion, *i.e.*, whether $p(s) \to 0$ as $s \to 0$, which is a common trait of the GOE ensemble, as the probability of two eigenvalues being exactly degenerate is zero. Furthermore, the level spacing distribution $p(s)$ for certain systems is known. For integrable systems, it follows the Poisson distribution $p(s) = e^{-s}$, while for chaotic systems (GOE), it is given by the Wigner surmise

$$p_\beta(s) = Z_\beta s^\beta e^{-b_\beta s^2}, \tag{6}$$

where β, Z_β, and b_β depend on which universality class of random matrices the covariance matrix belongs to [40]. In this work, we focus on matrices that fall under the universality class of the GOE, for which $\beta = 1$, as we show that both real-world data and CGD Gram matrices belong to.

While the level spacing distribution depends on unfolding the eigenspectrum, which is only heuristically defined and has some arbitrariness, it is useful to have additional diagnostics of chaotic behavior that bypass the unfolding procedure. The r-statistics, first introduced in Oganesyan and Huse [43], is such a diagnostic tool for short-range correlations, defined without the need to unfold the spectrum.

Given the level spacings s_i, defined as the differences between adjacent eigenvalues $\cdots < \lambda_i < \lambda_{i+1} < \cdots$ *without* unfolding, one defines the following ratios:

$$r_i = \mathrm{Min}(s_i,\, s_{i+1})/\mathrm{Max}(s_i,\, s_{i+1}) \; , \qquad 0 \leq r_i \leq 1 \; . \tag{7}$$

The expectation value of the ratios r_i takes very specific values if the energy levels are the eigenvalues of random matrices: for matrices in the GOE the ratio is $\langle r \rangle \approx 0.536$. The value becomes typically smaller for integrable systems, approaching $\langle r \rangle \approx 0.386$ for a Poisson process [39].

Spectral Form Factor: The spectral form factor (SFF) is a long-range observable that probes the agreement of a given unfolded spectrum with RMT at energy scales much larger than the mean level spacing. It can be used to detect spectral rigidity, which is a signature of the RMT regime.

The SFF is defined as the Fourier transform of the spectral two-point correlation function [44, 45]

$$K(\tau) = |Z(\tau)|^2/Z(0)^2 \simeq \frac{1}{Z}\langle |\sum_i \rho(e_i) e^{-i2\pi e_i \tau} \|^2 \rangle \; , \tag{8}$$

where $Z(\tau) = \mathrm{Tr} e^{-i\tau\Sigma_M}$. The second equality is the numerically evaluated SFF [46], where e_i is the unfolded spectrum, and $Z = \sum_i |\rho(e_i)|^2$ is chosen to ensure that $K(\tau \to \infty) \approx 1$.

The SFF has been computed analytically for the GOE ensemble, and it reads

$$K_{\mathrm{GOE}}(\tau) = 2\tau - \tau \ln(1 + 2\tau) \text{ for } 0 < \tau < 1, \quad K_{\mathrm{GOE}}(\tau) = 1 \text{ for } 1 \leq \tau \; . \tag{9}$$

Several universal features occur in chaotic RMT ensembles, manifesting in Eq. (9) and discussed in detail in Liu [45], Kim et al. [42]. We mention here only two: (i) The constancy of $K(\tau)$ for $\tau \geq 1$ is simply a consequence of the discreteness of the spectrum. (ii) The existence of a timescale that characterizes the ergodicity of a dynamical system. It is defined as the time when the SFF of the dynamical system converges to the universal RMT computation. More concretely, it is indicated by the onset of the universal linear ramp 2τ as in equation 9, which is absent in non-ergodic systems.

4.2 Insights from the Global and Local Statistical Structure

4.2.1 Eigenvalues Distributions

While the scaling behavior of the bulk of eigenvalues is certainly meaningful, it is not the only piece of information that can be extracted from the empirical covariance matrix. Particularly, it is natural to inquire whether the origin of the power-law scaling determines also the degeneracy of each eigenvalue. We can test this hypothesis by comparing the global and local statistics of the bulk between real-world data and their CGD counterparts.

For the gaussian datasets we generate, there are known predictions for the spectral density, level spacing distribution, r-statistics and spectral form factor. In these special cases, the empirical covariance matrix in Eq. (1) is known as a Wishart matrix [47]: $\Sigma_{ij,M} \sim \mathcal{W}_d(\Sigma, M)$.

For a Wishart matrix, the spectral density $\rho(\lambda)$ is given by the generalized Marčenko-Pastur (MP) law [48, 26], which depends on the details of Σ and specified in Appendix B, for certain limits. For $\Sigma = \sigma^2 I_d$, the spectral density is given explicitly by the MP distribution as

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\gamma\lambda} \quad \text{for } \lambda \in [\lambda_{\min}, \lambda_{\max}] \text{ and 0 otherwise} \; , \tag{10}$$

where $\sigma \in \mathbb{R}^+$, $\lambda_{\max/\min} = \sigma^2(1 \pm \sqrt{\gamma})^2$, $\gamma \equiv d/M$ and $d, M \to \infty$.

In Fig. 2, we show that the CGDs capture not only the scaling behavior of the eigenvalue bulk, but also the spectral density and the distribution of eigenvalues, for ImageNet, CIFAR10, and FMNIST, measured by the Kullback–Leibler divergence (KL) [49]. We further emphasize this point by contrasting the distributions with the MP distribution, which accurately captures the spectral density of the UGD datasets. This measurement alone is insufficient to determine that the system is well approximated by RMT, and we must study several other statistical diagnostics.

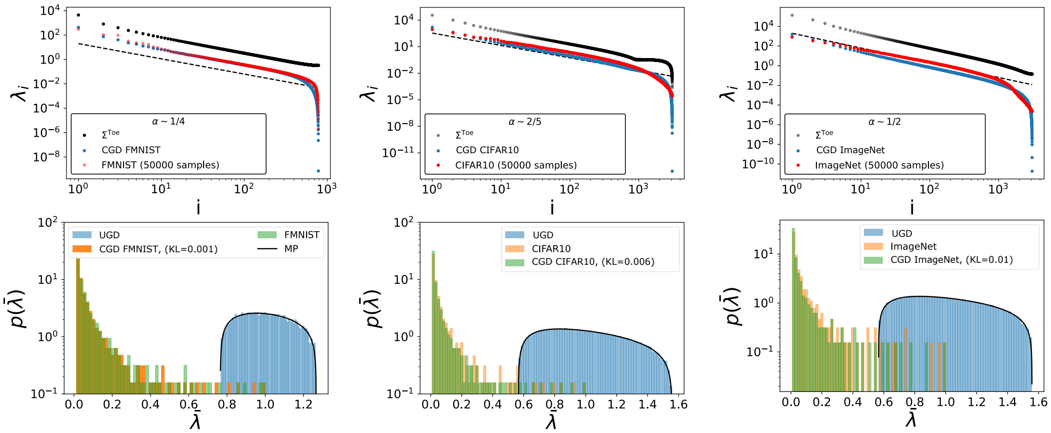

Figure 2: **Top row:** Scree plot of $\Sigma_{ij,M}$ for several different configurations and datasets. We show the eigenvalues of the population covariance matrix Σ^{Toe}, the eigenvalues for the empirical covariance of the full real-world dataset with $M = 50$k and finally the eigenvalues of the empirical covariance using the same Σ^{Toe}, with $M = 50k$. The datasets used here are (left to right): FMNIST, CIFAR10, ImageNet. **Bottom row:** Spectral density for the bulk of eigenvalues for the same datasets, as well as a comparison against UGD of the same dimensions. The $\bar{\lambda}$ indicates normalization over the maximal eigenvalue among the bulk. We also provide the KL divergence between the CGDs and the real-world data distributions.

4.2.2 Level Spacing Diagnostics

RMT predicts that certain local and global statistical properties are determined uniquely by symmetry. Therefore, the empirical covariance matrix must lie either in the GOE ensemble if it is akin to a quantum chaotic system[3] or in the Poisson ensemble, if it corresponds to an integrable system.

Both the level spacing and r statistics (the ratio of adjacent level spacings) probability distribution functions and SFF for a Wishart matrix in the limit of $d, M \to \infty$ and $d/M = \gamma$, are given by the GOE universality class:

$$p_{\mathrm{GOE}}(s) = \frac{\pi}{2} s e^{-\frac{\pi}{4}s^2}, \quad p_{\mathrm{GOE}}(r) = \frac{27}{4}\frac{(r+r^2)}{(1+r+r^2)^{5/2}}\Theta(1-r), \quad \langle r\rangle_{\mathrm{GOE}} = 4-2\sqrt{3}, \quad (11)$$

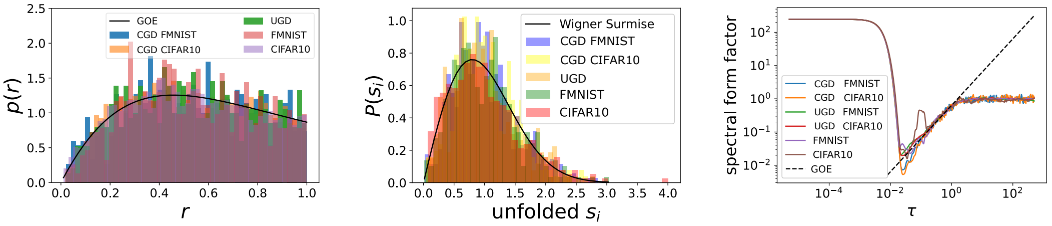

Figure 3: The r probability density (**left**), the unfolded level spacing distribution (**center**) and the spectral form factor (**right**) of Σ_M for FMNIST, CIFAR10, their CGDs, and UGD, obtained with $M = 50000$. Black curves indicate the RMT predictions for the GOE distribution from Eq. (11). These results indicate that the bulk of real-world data eigenvalues belongs to the GOE universality class, and that system has enough statistics to converge to the RMT predictions.

In Fig. 3, we demonstrate that the bulk of eigenvalues for various real-world datasets behaves as the energy eigenvalues of a quantum chaotic system described by the GOE universality class. This result is matched by both the UGD and the CGDs, as is expected of a Wishart matrix. Here, the dataset size is taken to be $M = 50000$ samples, and the results show that this sample size is sufficient to provide a proper sampling of the underlying ensemble.

[3] Large random real symmetric matrices belong in the orthogonally invariant class.

4.2.3 Effective Convergence

Having confirmed that CGDs provide a good proxy for the bulk structure for a large fixed dataset size, we may now ask how the statistical results depend on the number of samples.

As discussed in Section 3.1, Σ_M can be interpreted as an ensemble average over single realizations of the true population covariance matrix Σ. As the number of realizations M increases, a threshold value of M_{crit} is expected to appear when the space of matrices that matches the effective dimension of the true population matrix is fully explored.

The specific value of M_{crit} can be approximated without knowing the true effective dimension by considering two different evaluation metrics. Firstly, convergence of the local statistics of Σ_M, given by the point at which its level spacing distribution and r value approximately match their respective RMT ensemble expectations. Secondly, convergence of the global spectral statistics, both of Σ_M to that of Σ and of the empirical parameter α_M to its population expectation α.

Here, we define these metrics and measure them for different datasets, obtaining analytical expectations for the CGDs, which accurately mimic their real-world counterparts.

We can deduce M_{crit} from the local statistics by measuring the difference between the empirical average r value and the theoretical one given by

$$|r_M - r_{\text{RMT}}| = \delta(M) r_{\text{GOE}}, \tag{12}$$

where $r_{\text{GOE}} = 4 - 2\sqrt{3} \simeq 0.536$ for the Gaussian Orthogonal Ensemble.

Next, we compare the results obtained for M_{crit} from $\delta(M)$ to the ones obtained from the global statistics by using a spectral distance measure for the eigenvalue bulk given by

$$|\alpha_M - \alpha| = \Delta(M), \tag{13}$$

where α_M is the measured value obtained by fitting a power-law to the bulk of eigenvalues for a fixed dataset size M, while α represents the convergent value including all samples from a dataset.

Lastly, we compare the entire empirical Gram matrix Σ_M with the convergent result Σ obtained using the full dataset by taking

$$|\Sigma_M - \Sigma| = \epsilon(M)|\Sigma|, \tag{14}$$

where $|A|$ is the spectral norm of A, and $\epsilon(M)$ will be our measure of the distance between the two covariance matrices.

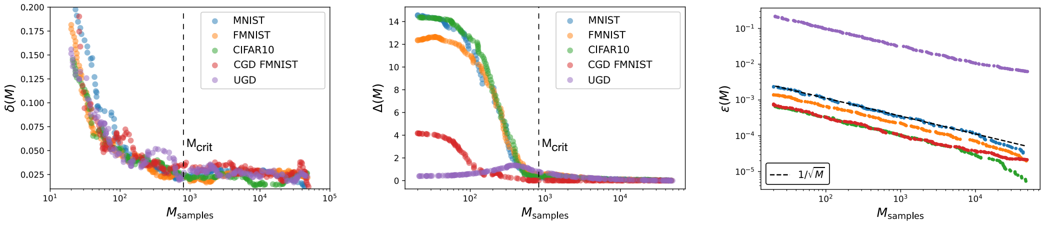

Figure 4: **Left:** The r distance metric $\delta(M)$ for the bulk of eigenvalues. **Center:** The α distance metric $\Delta(M)$ for the bulk of eigenvalues. **Right:** The full matrix comparison metric $\epsilon(M)$. We show the results for CIFAR10, FMNIST, UGD, and the FMNIST CGD as a function of the number of samples. The results show that the bulk distances decrease as $1/M$, where M is the number of samples, asymptoting to a constant value at similar values of $M_{\text{crit}} \sim d$ (**black dashed**), where d is the number of features.

In Fig. 4, we show the results for each of these metrics separately as a function of the number of samples M. We find that the $\delta(M)$ parameter, which is a measure of local statistics, converges to the expected GOE value at roughly the same M_{crit} as the entirely independent $\Delta(M)$ parameter, which measures the scaling exponent α. The combination of these two metrics confirms empirically that the system has become ergodic at sample sizes roughly $M_{\text{crit}} \sim d$, which is much smaller than the typical size of the datasets.

4.2.4 Datasets Entropy

[3] We omit UGD from the center panel, as $\alpha = -1$ regardless of M.

The Shannon entropy [50] of a random variable a measure of information, inherent to the variable's possible outcomes [51], given by $H = -\sum_{i=1}^{n} p_i \log(p_i)$ where p_i is the probability of a given outcome and n is the number of possible states. For covariance matrices, we define p_i given the spectrum as $p_i = \lambda_i / \sum_{i=1}^{n_{\text{bulk}}} \lambda_i$, where n_{bulk} is the number of bulk eigenvalues.

In Fig. 5 (left) we plot the Shannon entropies of real and gaussian datasets as a function of the number of samples. The entropies grow linearly and reach a plateau whose value is related to the correlation strength, with strong correlation corresponding to low entropy. We see the same entropy for both the gaussian and real datasets that have the same scaling exponent, implying that they also share the same eigenvalues degeneracy.

4.2.5 Entropy, Scaling and r-Statistics

In Fig. 5 (left), we see that the entropy saturation is correlated with the effective convergence in Fig. 4 as a function of the number of samples, while the middle and right plots show the correlation between the convergence of the entropy, the scaling exponent and the r-statistics, respectively. We see that real data and gaussian data with the same scaling exponent exhibit similar convergence behaviour.

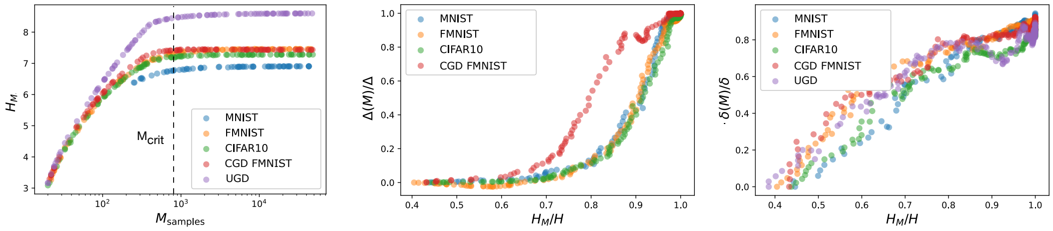

Figure 5: Convergence of the various metrics in Eqs. (12) to (14) in relation to entropy for the bulk of eigenvalues. **Left:** The Shannon entropy H_M as a function of the dataset size M. **Center:** Convergence of the normalized α metric Δ_M/Δ to its asymptotic value as a function of the normalized entropy H_M/H. **Right:** Convergence of the normalized r statistics metric δ_M/δ to its asymptotic value as a function of the normalized entropy H_M/H. We show the results for CIFAR10, FMNIST, MNIST, UGD, and the FMNIST CGD[5].

5 Conclusions

In this work, we have shown that the bulk eigenvalues of Gram matrices for real world data can be well approximated by a Wishart matrix with a shift invariant correlation structure and a defining exponent α. The fact that these Gram matrices are universally GOE, regardless of the generating process, implies that this approximation is good not only for the scaling of the eigenvalues, but also for their distribution, as the latter can be derived using RMT tools.

We believe our work bridges the gap to producing provable statements regarding NNs beyond the ubiquitous random feature model, which lacks data-data correlations. Although the random feature model is an obvious over-simplification, it has been useful in understanding certain aspects of the NN learning process, related to learning dynamics [52–54], parameter scaling limits [4], weight evolution [55], to name a few. As a basic application, we show in Appendix D how our results are required to solve the dynamics of even the simplest teacher-student model with correlated data.

We suggest an RMT model of data much closer to the real world, whereby correlations are simply introduced, but the RMT structure is maintained. This has been done to some extent in the neural scaling laws literature, but we believe that by re-affirming this approach with much stronger tools, we allow practitioners to make predictions much more aligned with behaviors found in the real world.

Our work can also aid in understanding the underlying distribution of real data; Not every distribution will converge to a GOE rather than Poisson with a finite number of samples. This sets constraints on the moments of the underlying distribution of real images, and can also help understand the data generation process which conforms to these constraints.

[5] We omit UGD from the center panel, as $\alpha = -1$ regardless of M.

In this manuscript, we focused on the Gram matrix, which, by construction ignores spatial information. We therefore do not capture the full statistics of the images. The strength of our analysis is in its generality. By proposing the simplest model for Gram matrices, we can easily extend our analysis to other domains, such as language datasets, or audio signals, offering valuable insights into the universality of scaling laws across modalities. Additionally, the interplay between eigenvectors and eigenvalues in neural networks merits further exploration, as both components likely play crucial roles in the way neural networks process information.

Finally, while our empirical results indicate that real-world data displays chaotic properties, the exact source is not evident. We believe that further work is necessary to determine whether it is due to the underlying strongly correlated structure that is manifest in real-world data, or if it stems from the chaotic sampling process that generates noise, which is captured in the finer details encoded in the eigenvalue bulk.

6 Acknowledgements

We would like to thank Yohai Bar-Sinai, Marat Freytsis, Alex Maloney, Dan Roberts, and Jamie Sully for useful discussions and comments. N.L. would like to thank the Milner Foundation for the award of a Milner Fellowship. The work of Y.O. is supported in part by Israel Science Foundation Center of Excellence. This work was performed in part at Aspen Center for Physics, which is supported by the U.S. National Science Foundation grant PHY-2210452.

References

[1] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00008-4.

[2] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found Comput Math*, 7(2):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[4] Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws, 2022.

[5] Thomas Guhr, Axel Müller–Groeling, and Hans A. Weidenmüller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(4-6):189–425, jun 1998. doi: 10.1016/s0370-1573(97)00088-4. URL `https://doi.org/10.1016%2Fs0370-1573%2897%2900088-4`.

[6] O. Bohigas, M. J. Giannoni, and C. Schmit. Characterization of chaotic quantum spectra and universality of level fluctuation laws. *Physical review letters*, 52(1):1, 1984.

[7] M. L. Mehta. *Random matrices*, volume 111. Academic Press, 1991.

[8] A. Pandey. Random matrix theory and quantum chaos. *Reviews of Modern Physics*, 55(4): 807–823, 1983.

[9] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, H. E. Stanley, and Stanley M. S. Random matrix theory and financial markets. *Physical Review E*, 60(5):6519–6532, 1999.

[10] T. A. Brody. Random matrix models in nuclear physics. *Reports on Progress in Physics*, 44(4): 1125–1191, 1981.

[11] K. B. Efetov. *Supersymmetry and disorder in quantum mechanics*. Cambridge University Press, 1997.

[12] Noam Levi and Yaron Oz. The universal statistical structure and scaling laws of chaos and turbulence, 2023.

[13] Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006. ISSN 1567-2190. doi: 10.1561/0100000006. URL `http://dx.doi.org/10.1561/0100000006`.

[14] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data, 2022.

[15] Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7354–7371. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.findings-emnlp.544`.

[16] Ibrahim M. Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/8c22e5e918198702765ecff4b20d0a90-Abstract-Conference.html`.

[17] Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *J. Mach. Learn. Res.*, 23:9:1–9:34, 2022. URL `http://jmlr.org/papers/v23/20-1111.html`.

[18] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html`.

[19] Lukasz Debowski. A simplistic model of neural scaling laws: Multiperiodic santa fe processes. *CoRR*, abs/2302.09049, 2023. doi: 10.48550/arXiv.2302.09049. URL `https://doi.org/10.48550/arXiv.2302.09049`.

[20] Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. *CoRR*, abs/2302.09650, 2023. doi: 10.48550/arXiv.2302.09650. URL `https://doi.org/10.48550/arXiv.2302.09650`.

[21] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017. URL `https://papers.nips.cc/paper/6857-nonlinear-random-matrix-theory-for-deep-learning`.

[22] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124006, dec 2021. doi: 10.1088/1742-5468/ac3a77.

[23] Zhenyu Liao and Michael W. Mahoney. Hessian eigenspectra of more realistic nonlinear models, 2021.

[24] Charles H Martin and Michael W Mahoney. Traditional and heavy-tailed self-regularization in neural network models. *arXiv preprint arXiv:1901.08276*, 2019. URL `https://arxiv.org/abs/1901.08276`.

[25] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Phys. Rev. E*, 106:054124, Nov 2022. doi: 10.1103/PhysRevE.106.054124. URL `https://link.aps.org/doi/10.1103/PhysRevE.106.054124`.

[26] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. doi: 10.1017/9781009128490.

[27] Zhigang Bao, Guangming Pan, and Wang Zhou. Universality for the largest eigenvalue of sample covariance matrices with general population. *The Annals of Statistics*, 43(1), feb 2015. doi: 10.1214/14-aos1281. URL `https://doi.org/10.1214%2F14-aos1281`.

[28] Jinho Baik, Gerard Ben Arous, and Sandrine Peche. Phase transition of the largest eigenvalue for non-null complex sample covariance matrices, 2004.

[29] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features, 2022.

[30] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Pierre Haffner. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 2010.

[32] Han Xiao, Salim Rasul, and Richard S Zemel. Fashion-mnist: a novel image classification benchmark based on fashion articles. *arXiv preprint arXiv:1708.07747*, 2017.

[33] Cifar-10. URL `https://www.cs.toronto.edu/~kriz/cifar.html`.

[34] Antonio Torralba, Andreas A Efros, and Christopher Anderson. Tiny imagenet: A benchmark for evaluation of image classification algorithms. *International Journal of Computer Vision*, 69 (2):203–228, 2008.

[35] Ziwei Liu, Zihang Luo, Xiaogang Wang, and Xiaoou Tang. Celeba: A large-scale celebrity face attribute dataset. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, 2015.

[36] T. D. Lee and C. N. Yang. Statistical mechanics of charged particles. *Zeitschrift für Physik*, 196:433–453, 1966. doi: 10.1007/BF02750405.

[37] M. A. Smorodinsky. On the classical motion of charged particles. *Journal of Mathematical Physics*, 4:1005–1011, 1953. doi: 10.1063/1.1703719.

[38] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31:853–865, 1925. doi: 10.1007/BF01343133.

[39] Y. Y. Atas, E. Bogomolny, O. Giraud, and G. Roux. Distribution of the ratio of consecutive level spacings in random matrix ensembles. *Phys. Rev. Lett.*, 110:084101, 2013. doi: 10.1103/PhysRevLett.110.084101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.110.084101`.

[40] M. L. Mehta. *Random Matrices*. 3 edition, 2004.

[41] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[42] Joonho Kim, Yaron Oz, and Dario Rosa. Quantum chaos and circuit parameter optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(2):023104, feb 2023. doi: 10.1088/1742-5468/acb52d. URL `https://doi.org/10.1088%2F1742-5468%2Facb52d`.

[43] Vadim Oganesyan and David A. Huse. Localization of interacting fermions at high temperature. *Phys. Rev. B*, 75:155111, 2007. doi: 10.1103/PhysRevB.75.155111. URL `https://link.aps.org/doi/10.1103/PhysRevB.75.155111`.

[44] Jordan S. Cotler, Guy Gur-Ari, Masanori Hanada, Joseph Polchinski, Phil Saad, Stephen H. Shenker, Douglas Stanford, Alexandre Streicher, and Masaki Tezuka. Black holes and random matrices. *J. High Energy Phys.*, 2017:118, 2017. doi: 10.1007/JHEP05(2017)118. URL `https://link.springer.com/article/10.1007/JHEP05(2017)118`.

[45] Junyu Liu. Spectral form factors and late time quantum chaos. *Physical Review D*, 98(8), oct 2018. doi: 10.1103/physrevd.98.086026. URL `https://doi.org/10.1103%2Fphysrevd.98.086026`.

[46] J. Juntajs, J. Bonca, T. Prosen, and L. Vidmar. Quantum chaos challenges many-body localization. *Phys. Rev. E*, 102:062144, 2020. doi: 10.1103/PhysRevE.102.062144. URL `https://link.aps.org/doi/10.1103/PhysRevE.102.062144`.

[47] John Wishart Wishart. Generalised product moment distribution in samples from an indefinitely large population. *Biometrika*, 20(1-2):30–52, 1928.

[48] Jack W Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–199, 1995.

[49] Solomon Kullback and Richard A Leibler. Information theory and statistics. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[50] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[51] Alfred Rényi. On measures of information and entropy. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1(4):547–561, 1956.

[52] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124013, dec 2021. doi: 10.1088/1742-5468/ac3ae6. URL `https://doi.org/10.1088%2F1742-5468%2Fac3ae6`.

[53] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, 2020.

[54] Stephane dAscoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

[55] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model, 2018.

[56] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[57] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

[58] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 426–471. PMLR, 2022.

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

[60] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23549–23588. PMLR, 2022.

[61] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *arXiv:2205.13303*, 2023.

[62] Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? extents and limits of universality in high-dimensional generalized linear estimation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of The 40th International Conference on Machine Learning*, volume 162, pages 10–15. PMLR, 2023.

[63] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[64] Timothy LH Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[65] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[66] David L Donoho. Statistical modeling of natural images with wavelets. *Proceedings of the National Academy of Sciences*, 92(12):5191–5196, 1995.

[67] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[68] 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.

[69] Lenka Zdeborov'a and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[70] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4): 935–969, 2016.

[71] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[72] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[73] Emmanuel J Cand'es, Pragya Sur, et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1): 27–42, 2020.

[74] Benjamin Aubin, Florent Krzakala, Yue M Lu, and Lenka Zdeborov'a. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[75] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The performance analysis of generalized margin maximizers on separable data. In *International Conference on Machine Learning*, pages 8417–8426. PMLR, 2020.

[76] Michael A Nielsen. A simple formula for the average gate fidelity of a quantum dynamical operation. *Physics Letters A*, 303(4):249–252, October 2002. ISSN 0375-9601. doi: 10.1016/s0375-9601(02)01272-0. URL `http://dx.doi.org/10.1016/S0375-9601(02)01272-0`.

[77] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures, 2020.

[78] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, May 2019. ISSN 1091-6490. doi: 10.1073/pnas.1820226116. URL `http://dx.doi.org/10.1073/pnas.1820226116`.

A The Unfolding Procedure

Here, we provide additional details on the unfolding procedure used to produce Fig. 3 in the main text.

Care must be taken when analyzing the eigenvalues of the empirical covariance matrix Σ_M, since they exhibit unavoidable numerical errors. To control for the effect of numerical errors, we adopted a robust phenomenological procedure that utilizes the fact that all eigenvalues of Σ_M must be non-vanishing by definition. To ensure we consider only eigenvalues of Σ_M unimpacted by edge effects, we inspect only the bulk spectrum.

Restricting to the bulk removes many eigenvalues of Σ_M as many are zero for small M. However, for larger M when Σ_M's structure is clearly visible, this is not the case. The procedure ensures the eigenvalues kept are robust and not significantly impacted by numerical precision. From the significant eigenvalues of the empirical covariance matrix Σ_M, we compute the spectrum λ_i.

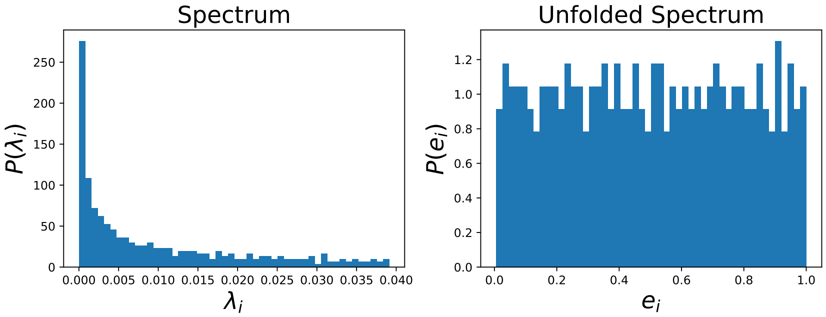

Figure 6: Bulk eigenvalue distribution for the empirical covariance matrix constructed from $M = 50000$ samples of FMNIST, before unfolding (**left**), and after unfolding (**right**) The unfolded spectrum displays approximately unit mean, and defined on the interval $[0, 1]$.

The unfolding procedure used to derive the unfolded spectrum is as follows:

1. Arrange the non-degenerate eigenvalues, λ_i , of the empirical covariance matrix (Σ_M) in ascending order.
2. Compute the staircase function $S(\lambda)$ that enumerates all eigenstates of the empirical covariance matrix (Σ_M) whose eigenvalues are smaller than or equal to λ.
3. Fit a smooth curve, denoted by $\tilde{\rho}(\lambda)$, to the staircase function. Specifically, we used a 12^{th}-order polynomial as the smooth approximation.
4. Rescale the eigenvalues λ_i as follows:
$$\lambda_i \to e_i = \tilde{\rho}(\lambda_i) \qquad (15)$$
5. By construction, the unfolded eigenvalues e_i should show an approximately uniform distribution with mean level spacing 1. This can be used to check if the procedure was successful by plotting the unfolded levels and checking the flatness of the distribution.

In Fig. 6, we show an example of the unfolding procedure for the FMNIST dataset. Specifically, we show the eigenvalue distribution before ($P(\lambda_i)$) and after ($P(e_i)$) unfolding. Up to the quality of the smoothing function $\tilde{\rho}(\lambda_i)$, the unfolded eigenvalue distribution displays a uniform distribution on the unit interval.

B Spectral Density for Wishart Matrices with a Correlated Features

For $z \in \mathbb{C}\backslash\text{supp}(\rho_\Sigma)$, the Stieltjes transform G and inverse Stieljes transform ρ_Σ are defined as

$$G(z) = \int \frac{\rho_\Sigma(t)}{z-t}dt = -\frac{1}{n}\mathbb{E}\big[\text{Tr}(\Sigma - zI_n)^{-1}\big], \qquad \rho_\Sigma(\lambda) = -\frac{1}{\pi}\lim_{\epsilon\to 0^+}\Im G(\lambda + i\epsilon), \qquad (16)$$

where $\mathbb{E}[\ldots]$ is taken with respect to the random variable X and $(\Sigma - zI_n)^{-1}$ is the resolvent of Σ.

For the construction, discussed in the main text, and general α, there is no closed form for the spectral density. However, in certain limits, analytical expressions can be derived from the Stieljes transform using Eq. (16). Specifically, given a determinstic expression for Σ, the spectral density can be derived by evoking Theorem 2.6 found in Couillet and Liao [26], which uses the following result by Silverstein and Bai [48]

$$G(z) = \frac{1}{\gamma}\tilde{G}(z) + \frac{1-\gamma}{\gamma z}, \qquad \tilde{G}(z) = \left(-z + \frac{1}{M}\mathrm{Tr}\left[\Sigma(I_d + \tilde{G}(z)\Sigma)^{-1}\right]\right)^{-1}, \tag{17}$$

where $\gamma \equiv d/M$ and $d, M \to \infty$, and we substitute C from the original theorem with Σ.

The empirical covariance matrix of the Gaussian correlated datasets discussed in the main text, is a Wishart matrix with a deterministic covariance, and thus fits the requirements of Theorem 2.6, where $\Sigma^{\mathrm{Toe}} = S$, $S = V^\dagger T U$, and $T_{i,j} = \boldsymbol{I}_{ij} + c|i-j|^\alpha$. In order to use Eq. (17), it is useful to first find the singular values of $T_{i,j}$. This can be done by using the discrete Laplace transform (extension of the Fourier transform), leading to

$$\Sigma^{\mathrm{Toe}}(s) = S(s) = 1 + c\mathrm{Li}_{-\alpha}\left(e^{-\frac{s}{d}}\right) - ce^{-s}\Phi\left(e^{-\frac{s}{d}}, -\alpha, d\right), \tag{18}$$

where $s = 1 \ldots d$, $\Phi(x,k,a)$ is the Lerch transcendent, and $\mathrm{Li}(x)$ is the Poly-log function. Note that by the definition of S, Eq. (18) is a non-negative function of s. Because the identity matrix commutes with Σ^{Toe}, we may substitute Eq. (18) in Eq. (17) to obtain

$$\tilde{G}(z) = \frac{1}{-z + \frac{\gamma}{d}S_d(\alpha)}, \tag{19}$$

where we define the sum $S_d(\alpha)$ to be

$$S_d(\alpha) = \sum_{s=1}^{d} \frac{\Sigma^{\mathrm{Toe}}(s)}{1 + \tilde{G}(z)\Sigma^{\mathrm{Toe}}(s)}. \tag{20}$$

Since the behavior of $\Sigma^{\mathrm{Toe}}(s)$ is intrinsically different for positive and very negative α, we separate the two cases. First, consider the case of $\alpha < -1$, where correlations decay very quickly. In this scenario, the covariance matrix reduces to $\tilde{\Sigma}^{\mathrm{Toe}}(s) \simeq 1$.

Here, $S_d(\alpha)$ is given simply by the $\alpha \to \infty$ limit

$$S_d(\alpha \to -\infty) = \sum_{k=1}^{d} \frac{1}{1 + \tilde{G}(z)} = \frac{d}{1 + \tilde{G}(z)}, \tag{21}$$

which is precisely the case of $\Sigma = I_d$.

Solving Eq. (19) using the above result yields the following expression for $\tilde{G}(z)$

$$\tilde{G}(z) = \frac{-1 - z + \gamma - \sqrt{(-1 - z + \gamma)^2 - 4z}}{2z}. \tag{22}$$

Finally, substituting Eq. (22) into Eq. (16) leads to the known Marčenko-Pastur (MP) law [26]

$$\rho(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_{\mathrm{max}} - \lambda)(\lambda - \lambda_{\mathrm{min}})}}{\gamma\lambda} \quad \text{for } \lambda \in [\lambda_{\mathrm{min}}, \lambda_{\mathrm{max}}] \text{ and 0 otherwise }, \tag{23}$$

where $\lambda_{\mathrm{max/min}} = (1 \pm \sqrt{\gamma})^2$.

The other interesting limit is that of $\alpha > -1$, in which the correlations do not decay quickly, and for $d \to \infty$ the Laplace transform of the Toeplitz matrix simplifies to

$$\Sigma^{\mathrm{Toe}}(s) \simeq c \cdot \Gamma(1+\alpha)\left(\frac{d}{s}\right)^{1+\alpha}. \tag{24}$$

Using this approximation for the population covariance in Eq. (20) we obtain

$$\tilde{G}(z) = \left(-z + \frac{\gamma}{d}\sum_{s=1}^{d}\frac{c(s/d)^{-1-\alpha}}{1 + \tilde{G}(z)c(s/d)^{-1-\alpha}}\right)^{-1}, \quad \hat{c} = c\Gamma(1+\alpha). \tag{25}$$

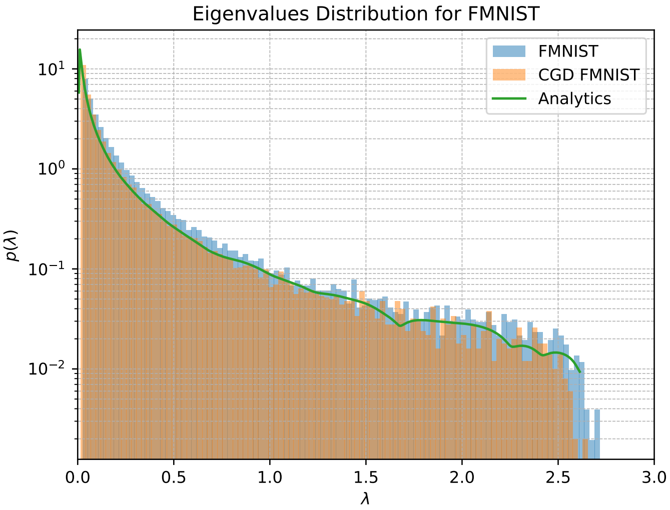

Figure 7: Theoretical predictions for the bulk spectral density of a CGD matrix against the empirical densities of FMNIST and the CGD. The **green** curve represents the generalized MP distribution, given by the solution to the inverse Steiljes transform in Eq. (27). The CGD curve has a value of $c = 1.14$, and $\gamma = 380/1000$ since that is the approximate number of bulk eigenvalues.

In the $d \to \infty$ limit,we can convert the sum to an integral using the Riemann definition

$$\lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} f(i/d) = \lim_{d\to\infty} \sum_{i=1}^{d} f(x_i)\Delta x = \int_a^b dx f(x), \quad \Delta x = \frac{b-a}{d} = \frac{1}{d}, \tag{26}$$

allowing us to write the equation for $\tilde{G}(z)$ as

$$\tilde{G}(z) = \left(-z + \frac{\gamma}{d} \sum_{s=1}^{d} \frac{\hat{c}(s/d)^{-1-\alpha}}{1 + \tilde{G}(z)\hat{c}(s/d)^{-1-\alpha}}\right)^{-1} \simeq \left(-z + \gamma \int_0^1 \frac{\hat{c}x^{-1-\alpha}}{1 + \tilde{G}(z)\hat{c}x^{-1-\alpha}} dx\right)^{-1} \tag{27}$$

$$= \left(-z + \gamma \frac{{}_2F_1\left(1, \frac{1}{\alpha+1}; 1 + \frac{1}{\alpha+1}; -\frac{1}{\hat{c}\tilde{G}(z)}\right)}{\tilde{G}(z)}\right)^{-1},$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function. Eq. (27) is an algebraic equation which can be solved numerically, or analytically approximated in certain limits.

In Fig. 7, we show the theoretical results for the spectral density of a Wishart matrix with Σ^{Toe} covariance, for a value of α and c matching FMNIST, against the empirical densities for FMNIST and the matching CGD. The green curve shows the generalized MP distribution given by the inverse Stieljes transform of Eq. (27).

C Robustness of our results

Here, we discuss some details regarding the robustness of our local and global statistical analyses.

For all of our analyses, we focused on the full Gram matrix, consisting of every sample in a given dataset. This implies that we only have access to a single realization of a Σ_M empirical Gram matrix, per dataset, thus limiting our ability to perform standard statistics, for instance averaging over an ensemble of Σ_M, and obtaining confidence bands. This is not an issue in the RMT regime, as the matrix itself is thought of as an ensemble on to itself, and its eigenvalues have an interesting structure due to a generalization of the Central Limit Theorem (CLT).

We can still attempt to persuade the reader that our results are robust a posteriori, by noting that the number of samples required to reach the RMT regime is approximately $M_{\text{crit}} \sim d$. This

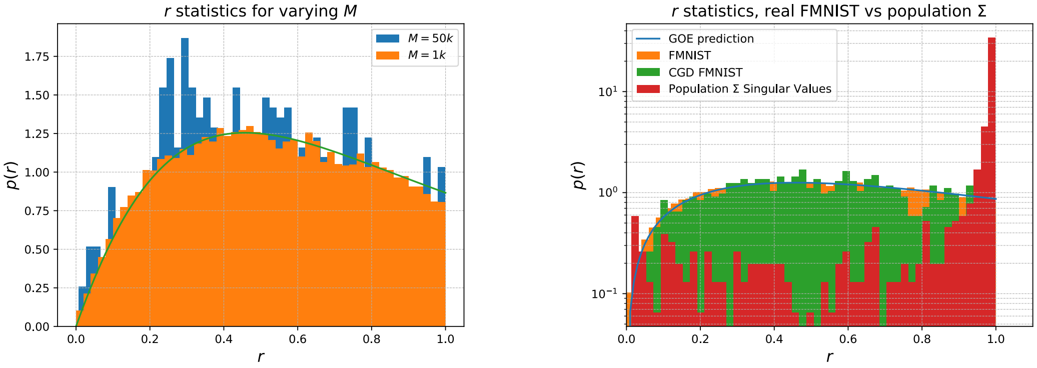

Figure 8: **Left:** The r statistics distribution for FashionMNIST, comparing $M = 50000$ with $M = 1000$ subsets. In the first case, we obtain only a single realization of the Gram matrix, and so the r statistics appear more noisy, however, when taking 40 realizations of a smaller subset, still above M_{crit}, we see that the fit to the GOE prediction (green) improves. We will add these figures, either in the main text or appendices, including goodness of fit measures on the rest of the datasets studied in the paper. **Right:** The r statistics distribution for FashionMNIST and its CGD. In red, we show the singular values of the population covariance, Σ^{Toe} used in the main text. In Orange, the true FMNIST r distribution, obtained by taking 40 different realizations of a $M = 1000$ subset of the full dataset, leading to a perfect fit to the GOE prediction (blue). In green, we show the CGD using a 1000 samples as well. This figure illustrates that the deterministic population covariance does not sufficiently capture all the information that resides in the Gram matrix, while the CGD does.

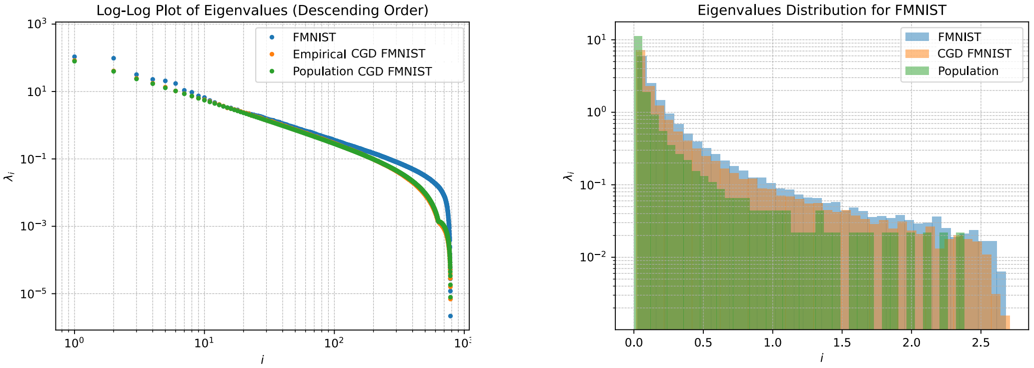

Figure 9: **Left:** Scree plot for the eigenvalues of the FashionMNIST Gram matrix (blue), its CGD (orange) using $M = 1000$ for 50 runs, and the Toeplitz population covariance matrix (green). Here, we show that the population and empirical covariance matrices match precisely in spectral scaling. The Gram matrices for FMNIST and its analogue are obtained by first normalizing the samples (mean subtraction and dividing by the standard deviation) and the population covariance is rescaled by a constant factor that depends only on the input dimension d. **Right:** The eigenvalue distribution for FashionMNIST, Gram matrix (blue), its CGD (orange) using $M = 1000$ for 50 runs, and the Toeplitz population covariance matrix (green). We see that the three distributions are similar, as can be expected, but that certain local features (such as the spacing between eigenvalues) is poorly captured by the deterministic population covariance.

implies that we can consider sub-samples of the full empirical Gram matrix, consisting of M_{crit}, as $\Sigma_{M_{\text{crit}}} = 1/M_{\text{crit}} \sum_{a=1}^{M_{\text{crit}}} X_{ia} X_{aj}$, and average over multiple sub-sample matrices.

In Fig. 8, we see an implementation of this process for FMNIST, demonstrating that additional sampling pushes the distribution to a perfect fit for the GOE r-statistics, while in Fig. 9, we see the same type of convergence for the eigenvalue distribution.

D Universality in Neural Network Analysis - A Toy Example

Universality laws have been employed in various ways to study error universality in neural networks, for instance in [56–58, 29]. In this context, one looks directly at the universality of the training and generalisation errors instead of the features, taking into account the labels and the task. It has also been observed to hold for correlated Gaussian data in teacher-student settings in [59]. Furthermore, it has been shown that for a simple regression task, the computation of the error reduces to an RMT problem [60], which is linked to the work presented in the main text regarding the data features themselves. In particular, it has been noted that in some cases the structure of the bulk fully characterises the error, even for multi-modal distributions, see [61, 62].

As a self contained example of applying universality in neural network analysis, we call upon an exceedingly simple machine learning setup, namely, training a linear network using gradient descent to learn a teacher-student mapping. We show that even in this basic example, it is necessary to apply the results demonstrated in the main text in order to correctly analyze the system, when data correlations are present in the underlying population covariance.

Teacher-student models have been the subject of a long line of works [63–70] , and have experienced a resurgence of interest in recent years [71–75] as a powerful tool to study the high-dimensional asymptotic performance of learning problems with synthetic data.

The teacher-student model can be described as follows: The teacher uses ground truth information along with a probabilistic model to generate data which is then passed to the student who is supposed to recover the ground truth as well as possible only from the knowledge of the data and the model.

Here, we consider a linear teacher-student model, where the data inputs $x_i \in \mathbb{R}^{d_{\text{in}}}$ are identical independently distributed (iid) normal variables drawn from a Gaussian distribution with non-trivial population covariance $x_i \sim \mathcal{N}(0, \Sigma_{\text{pop}})$. We draw N_{tr} training samples, and the teacher model generates output labels by computing a vector product on each input $y = w^* \cdot x$, where $w* \in \mathbb{R}^{d_{\text{in}}}$, assuming a perfect, noiseless teacher. The student, which shares the same model as the teacher, generates predictions $\hat{y} = w \cdot x$, where $w \in \mathbb{R}^{d_{\text{in}}}$ as well. The loss function which measures convergence of the student to the teacher outputs is the standard MSE loss. Our analysis is done in the regime of large input dimension and large sample size, i.e., $d_{\text{in}}, N_{\text{tr}} \to \infty$, where the ratio $\lambda \equiv d_{\text{in}}/N_{\text{tr}} \in \mathbb{R}^+$ kept constant. The student model is trained with the full batch Gradient Descent (GD) optimizer for t steps with a learning rate η. The training loss function is given by

$$\mathcal{L}_{\text{tr}} = \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \|(w - w^*)^T x_i\|^2 = \text{Tr}\left[\Delta^T \Sigma_{\text{tr}} \Delta\right], \tag{28}$$

where we define $\Delta \equiv w - w^*$ as the difference between the student and teacher vectors. Here, $\Sigma_{\text{tr}} \equiv \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} x_i x_i^T$ is the $d_{\text{in}} \times d_{\text{in}}$ empirical data covariance, or Gram matrix for the *training* set. The elements of w^* and w are drawn at initialization from a normal distribution $w_0, w^* \sim \mathcal{N}(0, 1/(2d_{\text{in}}))$. We do not include biases in the student or teacher weights, as they have no effect on centrally distributed data.

The generalization loss function is defined as its expectation value over the input distribution, which can be approximated by the empirical average over N_{gen} randomly sampled points

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{x \sim \mathcal{N}}\left[\|(w - w^*)^T x\|^2\right] = \text{Tr}\left[\Delta^T \Sigma_{\text{gen}} \Delta\right] . \tag{29}$$

Here Σ_{gen} is the covariance of the generalization distribution. Note that in practice the generalization loss is computed by a sample average over an independent set, which is not equal to the analytical expectation value. The gradient descent equations at training step t are

$$\Delta_{t+1} = \left(\boldsymbol{I} - 2\eta\Sigma_{\text{tr}}\right)\Delta_t, \tag{30}$$

where $\gamma \in \mathbb{R}^+$ is the weight decay parameter, and $\boldsymbol{I} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ is the identity.

Eq. (30) can be solved in the gradient flow limit, setting $\eta = \eta_0 dt$ and $dt \to 0$, resulting in

$$\dot{\Delta}(t) = -2\eta_0 \Sigma_{\text{tr}} \Delta(t) \quad \to \quad \Delta(t) = e^{-2\eta_0 \Sigma_{\text{tr}} t} \Delta_0, \tag{31}$$

where Δ_0 is simply the difference between teacher and student vectors at initialization. It follows that the empirical losses, calculated over a dataset admit closed form expressions as

$$\mathcal{L}_{\text{tr}} = \Delta_0^T e^{-4\eta_0 \Sigma_{\text{tr}} t} \Sigma_{\text{tr}} \Delta_0, \qquad \mathcal{L}_{\text{gen}} = \Delta_0^T e^{-2\eta_0 \Sigma_{\text{tr}} t} \Sigma_{\text{pop}} e^{-2\eta_0 \Sigma_{\text{tr}} t} \Delta_0. \tag{32}$$

Since the directions of both Δ and the eigenvectors of $\Sigma_{\rm tr}$ are uniformly distributed, we make the approximation that the projection of Δ on all eigenvectors is the same, which transforms Eq. (32) to the simple form

$$\mathcal{L}_{\rm tr} \approx \frac{1}{d_{\rm in}} \sum_i e^{-4\eta_0 \nu_i t} \nu_i \ , \tag{33}$$

while the calculation for the generalization loss amounts to

$$\mathcal{L}_{\rm gen} \approx \frac{1}{d_{\rm in}} \sum_{i,j} e^{-2\eta_0 (\nu_i + \nu_j) t} (U \Sigma_{\rm pop} U^\dagger)_{ij} \ , \tag{34}$$

where U is a random unitary matrix used to rotate to the basis of $\Sigma_{\rm tr}$.

Now we turn to the choice of $\Sigma_{\rm pop}$ and the implication for $\Sigma_{\rm tr}$. As demonstrated in the main text, the empirical covariance matrix of many real world data-sets can be faithfully modelled by a Wishart matrix with long range correlations, where the bulk of eigenvalues is described by the population covariance $\Sigma_{\rm pop} = \Gamma(1+\alpha)(i/d)^{-1-\alpha}\delta_{ij}$. As we discussed in Appendix B, we can utilize our RMT observations to give a closed formula expression for the empirical Gram matrix eigenvalues and spectral density, in terms of a generalized MP law.

Following this choice of data modelling, and focusing on the bulk eigenvalues alone, it is clear that the sums Eqs. (33) and (34) are the empirical averages over the function $e^{-4\eta_0 \nu t} f(\nu)$, if ν follows the spectral density derived in Appendix B. We can the solve the training dynamics by approximating the sum by its respective expectation value,

$$\mathcal{L}_{\rm tr}(\eta_0, \lambda, \alpha, t) \approx \mathbb{E}_{\nu \sim \rho_{\Sigma_{\rm pop}}(\lambda, \alpha)} \left[\nu e^{-4\eta_0 \nu t} \right] . \tag{35}$$

In order to proceed further for the generalization loss, we note that the rotation matrices which form the basis for $\Sigma_{\rm tr}$ are random unitary matrices, drawn from the Haar measure. This implies that we can glean further information by averaging over training realizations, which will not change the training trajectory at all, but will provide with an average generalization loss $\langle \mathcal{L} \rangle_U$. We utilize the following property of ensemble averaging over unitary random matrices [76]

$$\Phi(X) \equiv \mathbb{E}_U[UXU^\dagger] \equiv \int_{\mathcal{U}} d\mu(U) UXU^\dagger = \frac{1}{d_{\rm in}} {\rm Tr}(X) \boldsymbol{I}, \tag{36}$$

where $d\mu(U)$ is the Haar measure. Since the eigenvalue distribution does not change upon this averaging, the average generalization loss can be expressed as

$$\langle \mathcal{L}_{\rm gen} \rangle_U \approx {\rm Tr}(\Sigma_{\rm pop}) \times \frac{1}{d_{\rm in}} \sum_i e^{-4\eta_0 \nu_i t} \ , \tag{37}$$

which can be approximated by its expectation value

$$\langle \mathcal{L}_{\rm gen}(\eta_0, \lambda, \alpha, t) \rangle_U \approx {\rm Tr}(\Sigma_{\rm pop}) \mathbb{E}_{\nu \sim \rho_{\Sigma_{\rm pop}}(\lambda, \alpha)} \left[e^{-4\eta_0 \nu t} \right] , \tag{38}$$

completing the dynamical analysis of the loss curves for the model at hand.

Above we gave a toy example of how one may use our results to obtain justified theoretical predictions. Namely, we solved a simple teacher-student model with power law correlated data, and showed that the training dynamics and convergence both depend on the spectral density of the Gram matrices studied in the main text. We obtained analytical expressions for the training and generalization losses.

We stress that on their own, these findings do not attempt to fully explain many aspects of neural network dynamics and generalization, which depend on additional factors beyond the bulk spectrum, such as the large outlier eigenvalues, eigenvectors and higher moments.

Analyzing the interaction between these elements and learning dynamics/generalization remains an important open question, as recent works have started to demonstrate how outliers impact early gradient steps and network collapse.

For instance, as shown by [77] and verified by our results, the outliers can also be described by a Gaussian model, but simply not the CGD that we presented in this work, as the largest eigenvalues are expected to describe the most shared features in the entire data, and do not demonstrate the local correlation structure of the bulk. They are certainly important in classification tasks, and in particular their effect, as well as the effect of the different class mean values are the most important for linear classifiers, as shown in [78].

Our approach focused more on the regime where one would like to understand improved performance using more and more data, where the largest eigenvalues have long been well captured, and the only performance gain that can be achieved is squeezed out of the bulk alone. This has proven a sufficient path to construct solvable models which approximate real-world generalization curves [3, 4, 56].