

Applicability of Random Matrix Theory in Deep Learning

Nicholas P. Baskerville^{*1} Diego Granziol^{*2} Jonathan P. Keating³

Abstract

We investigate the local spectral statistics of the loss surface Hessians of artificial neural networks, where we discover excellent agreement with Gaussian Orthogonal Ensemble statistics across several network architectures and datasets. These results shed new light on the applicability of Random Matrix Theory to modelling neural networks and suggest a previously unrecognised role for it in the study of loss surfaces in deep learning. Inspired by these observations, we propose a novel model for the true loss surfaces of neural networks, consistent with our observations, which allows for Hessian spectral densities with rank degeneracy and outliers, extensively observed in practice, and predicts a growing independence of loss gradients as a function of distance in weight-space. We further investigate the importance of the true loss surface in neural networks and find, in contrast to previous work, that the exponential hardness of locating the global minimum has practical consequences for achieving state of the art performance.

1. Introduction

Artificial Neural Networks (ANNs) continually advance state of the art computer vision and natural language processing. However, we do not have a precise theoretical understanding of their training and generalisation dynamics. The observation that gradient based optimisation methods (Bottou, 2012) with different random initialisations do not seem to get stuck in poor quality local minima, despite the high dimensionality and non-convexity of the loss surfaces, has led to a significant focus on the neural network loss surface.

^{*}Equal contribution ¹School of Mathematics, University of Bristol, Bristol, United Kingdom ²Machine Learning Research Group, University of Oxford, Oxford, United Kingdom ³Institute of Mathematics, University of Oxford, Oxford, United Kingdom. Correspondence to: Nicholas P. Baskerville <n.p.baskerville@bristol.ac.uk>.

The loss surface is typically investigated through the matrix of second derivatives of the loss with respect to the weights, the *Hessian*. Under strong simplifying assumptions, such as independence of the neural network inputs and weights (Choromanska et al., 2015a;b; Pennington and Bahri, 2017), the Hessian at critical points of the loss (where the gradient is zero), are described by certain important classes of random matrices, such as the Gaussian Orthogonal Ensemble (GOE) (Tao, 2012) or the Wishart Ensemble (Bun et al., 2017) of Random Matrix Theory (RMT). The average spectral density (taken over an ensemble) of these matrices, in the limit of infinite dimension, can be calculated; for the GOE the result is known as the *Wigner semicircle law*, and for the Wishart Ensemble it is the *Marchenko-Pastur law*. Hence with these assumptions, one can make quantitative predictions about the nature of the critical points and aspects of the geometry of the loss landscape. Another line of enquiry has been the study of similarity between neural networks and *spin-glass* models from statistical physics (Amit et al., 1985; Gardner and Derrida, 1988), extended recently to ANNs (Baity-Jesi et al., 2018; Sagun et al., 2014), where the use of weight decay has been shown to be analogous to a magnetic field in disordered systems (Chaudhari and Soatto, 2015). The Hessian of a spin-glass at a given energy level, or the loss value for a Deep Neural Network (DNN) is given by a random matrix (Auffinger et al., 2013; Castellani and Cavagna, 2005) and hence spin glass models and Random Matrix Theory are closely related.

Choromanska et al. (2015a) showed, assuming i.i.d Gaussian inputs and network path independence, that a multi-layer ReLU neural network’s loss is equivalent to that of a spin-glass model. Its conditional Hessian spectrum is thus given by a GOE calculation (Auffinger et al., 2013) involving real-symmetric matrices with otherwise independent Gaussian random entries. It follows that under these assumptions local minima are located within a narrow band, bounded below by the global minimum. The practical implication is that for a sufficient number of hidden layers (more than 2) all local minima are *close* in loss to the global minimum. Baskerville et al. (2020) extend this line of work to networks with general activation functions. Baskerville et al. (2021) show for General Adversarial Networks, using a spin-glass model for both the generator and discriminator, that the structure of local optima encourages collapse to a

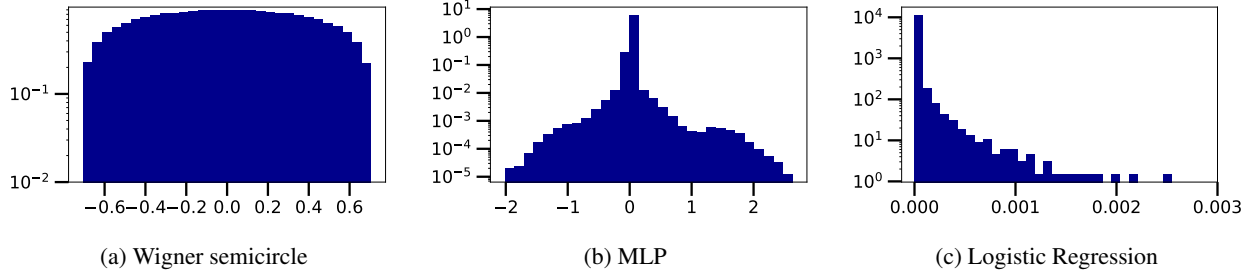


Figure 1. Deep neural network spectra do not match those predicted by those of theoretical spin glass models: Comparison of different global spectral statistics (spectral densities). We show actual GOE data to demonstrate the form of the Wigner semicircle, data from MLP and logistic regression models on MNIST (see Section 4). Note the log-scale on the y-axis. A few outliers have been clipped from logistic regression to aid visualisation.

narrow band of the loss for at least one of the networks but not necessarily both simultaneously. Similarly, Pennington and Bahri (2017) use the Gauss Newton decomposition of a squared loss Hessian, assuming independence and normality of both the data and weights along with free addition of the resulting Wigner/Wishart ensembles, to derive a functional form for the critical index (the fraction of the eigenvalues that are negative) as a function of the loss. They show that below a certain critical energy threshold *all critical points are minima*. In Ba et al. (2020) the authors assume Gaussian inputs, i.i.d Gaussian weights and a linear teacher model and use Random Matrix Theory to derive the generalisation properties of two layer neural networks, such as the population risk¹ of the regularised least squares problem to explain the *double descent phenomenon*: increasing the network size initially leads to over-fitting, but beyond a critical point, further increasing the network size decreases the test error to a lower level than the optimal small network.

An important and fundamental problem with the aforementioned works is that typically the average spectral density of the Hessian of neural networks does not in fact match that of the associated random matrix ensembles. This is illustrated in Figure 1. Put simply, *we do not observe the Wigner semicircle or Marchenko-Pastur eigenvalue distributions, implied by the Gaussian Orthogonal or Wishart Ensembles for ANNs*. As shown extensively in Granziol (2020); Granziol et al. (2019a); Pappan (2018; 2019); Ghorbani et al. (2019); Sagun et al. (2016; 2017) the spectral density of ANN Hessians contain outliers and a large number of near zero eigenvalues, features not seen in canonical random matrix ensembles. Furthermore, even allowing for this, as shown in (Granziol et al., 2020a) by specifically embedding outliers as a low rank perturbation to a random matrix, the remaining bulk spectral density still does not match the Wigner semicircle or Marchenko-Pastur distributions (Granziol, 2020), bringing into question the validity of the underlying modeling.

¹Loss under the expectation of the data generating distribution.

The fact that the experimental results differ markedly from the theoretical predictions calls into question the validity of ANN analyses based on canonical random matrix ensembles. Moreover, the compelling results of works such as (Choromanska et al., 2015a; Pennington and Bahri, 2017) are obtained using very particular properties of the canonical ensembles, such as large deviation principles, as pointed out in Granziol (2020). The extent to which such results can be generalised is an open question. Hence, further work is required to better understand to what extent Random Matrix Theory can be used to analyse the loss surfaces of ANNs.

The main novel contributions of this paper are

- We show that the *local spectral statistics* (i.e. those measuring correlations on the scale of the mean eigenvalue spacing) of ANN Hessians are well modelled by those of GOE random matrices, even when the mean spectral density is different from the semicircle law. We display this on MNIST trained multi-layer perceptrons and on the final layer of a ResNet-34 on CIFAR-10.
- Based on these observations we propose a novel model of the loss surface under the data generating distribution i.e. *The True Loss*. Such a model allows for outliers and rank degeneracy in the empirical Hessian extensively observed in practice. We show in particular that several global spectral densities that match those observed in real neural networks give rise to GOE statistics in the local spectral statistics.
- We show that properties such as the exponential hardness of attaining the global minimum familiar from spin glass theory re-emerge for deep neural networks provided we consider the minimum of the True Loss not the Empirical Loss. This implies that for training procedures in which the empirical and true loss do not significantly deviate, superior empirical results can be continually achieved by increasing training time. This in contrast to prior work, shows that spin glass models

of neural networks can be very useful in predicting behaviour which is relevant for achieving state of the art performance, opening up the avenue of further research in understanding the surface of the True Loss.

2. Preliminaries

Consider a neural network with weights $\mathbf{w} \in \mathbb{R}^P$ and a dataset with distribution \mathbb{P}_{data} . Let $L(\mathbf{w}, \mathbf{x})$ the loss of the network for a single datum \mathbf{x} and let \mathcal{D} denote any finite sample of data points from \mathbb{P}_{data} . The *true loss* is given by

$$\mathcal{L}_{true}(\mathbf{w}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{data}} L(\mathbf{w}, \mathbf{x}) \quad (1)$$

and the *empirical loss* (or training loss) is given by

$$\mathcal{L}_{emp}(\mathbf{w}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} L(\mathbf{w}, \mathbf{x}). \quad (2)$$

Where \mathcal{D} denotes the dataset. The true loss is a deterministic function of the weights, while the empirical loss is a random function with the randomness coming from the random sampling of the finite dataset \mathcal{D} . The empirical Hessian $\mathbf{H}_{emp}(\mathbf{w}) = \nabla^2 R_{emp}(\mathbf{w})$, describes the loss curvature at the point \mathbf{w} in weight space. By the spectral theorem, the Hessian can be written in terms of its eigenvalue/eigenvector pairs $\mathbf{H}_{emp} = \sum_i \lambda_i \phi_i \phi_i^T$, where the dependence on \mathbf{w} has been dropped to keep the notation simple. The eigenvalues of the Hessian are particularly important, being explicitly required in second-order optimisation methods, and characterising the stationary points of the loss as local minima, local maxima or generally saddle points of some other index.

For a matrix drawn from a probability distribution, its eigenvalues are random variables. The eigenvalue distribution is described by the joint probability density function (j.p.d.f) $p(\lambda_1, \lambda_2, \dots, \lambda_P)$, also known as the P -point correlation function. The simplest example is the *empirical spectral density* (ESD), $\rho^{(P)}(\lambda) = \frac{1}{P} \sum_i \delta(\lambda - \lambda_i)$. Integrating $\rho^{(P)}(\lambda)$ over an interval with respect to λ gives the fraction of the eigenvalues in that interval. Taking an expectation over the random matrix ensemble, we obtain the *mean spectral density* $\mathbb{E}\rho^{(P)}(\lambda)$, which is a deterministic probability distribution on \mathbb{R} . Alternatively, taking the $P \rightarrow \infty$ limit, assuming it exists, gives the *limiting spectral density* (LSD) ρ , another deterministic probability distribution on \mathbb{R} . A key feature of many random matrix ensembles is *self-averaging* or *ergodicity*, meaning that the leading order term (for large P) in $\mathbb{E}\rho^{(P)}$ agrees with ρ . Given the j.p.d.f, one can obtain the mean spectral density, known as the the 1-point correlation function (or any other k -point correlation function) by marginalisation

$$\mathbb{E}\rho^{(P)}(\lambda) = \int p(\lambda, \lambda_2, \dots, \lambda_P) d\lambda_2 \dots d\lambda_P. \quad (3)$$

A GOE matrix is an example of a *Wigner random matrix*, namely a real-symmetric (or complex-Hermitian) matrix with otherwise i.i.d. entries and off-diagonal variance σ^2 .² The mean spectral density for Wigner matrices is known to be Wigner's semicircle (Mehta, 2004)

$$\rho_{SC}(\lambda) = \frac{1}{2\pi\sigma^2 P} \sqrt{4P\sigma^2 - \lambda^2} \mathbf{1}_{|\lambda| \leq 2\sigma\sqrt{P}}. \quad (4)$$

The radius of the semicircle³ is proportional to $\sqrt{P}\sigma$, hence scaling Wigner matrices by $1/\sqrt{P}$ leads to a limit distribution when $P \rightarrow \infty$. This is the LSD. With this scaling, there are, on average, $\mathcal{O}(P)$ eigenvalues in any open subset of the compact spectral support. In this sense, the mean (or limiting) spectral density is *macroscopic*, meaning that, as $P \rightarrow \infty$, one ceases to see individual eigenvalues, but rather a continuum with some given density.

3. Motivation: Microscopic Universality

Random Matrix Theory was first developed in physics to explain the statistical properties of nuclear energy levels, and later used to describe the spectral statistics in atomic spectra, condensed matter systems, quantum chaotic systems etc; see, for example (Weidenmuller and Mitchell, 2008; Beenakker, 1997; Berry et al., 1987; Bohigas, 1991). *None of these physical systems exhibits a semicircular empirical spectral density.* Similarly, neither MLP nor Softmax Regression Hessians are described by the Wigner semicircle law which holds for GOE matrices, shown in Figure 1a: their spectra contain outliers, large peaks near the origin and the remaining components of the histogram also do not match the semicircle.

However all the physical systems show agreement with RMT at the level of the mean eigenvalue spacing when local spectral statistics are compared. Physics RMT calculations re-scale the eigenvalues to have a mean level spacing of 1 and then typically look at the *nearest neighbour spacings distribution* (NNSD), i.e. the distribution of the distances between adjacent pairs of eigenvalues. One theoretical motivation for considering the NNSD is that it is independent of the Gaussianity assumption and reflects the symmetry of the underlying system. It is the NNSD that is universal (for systems of the same symmetry class) and not the average spectral density, which is best viewed as a parameter of the system. The aforementioned transformation to give mean spacing 1 is done precisely to remove the effect of the average spectral density on the pair correlations leaving behind only the universal correlations. To the best of our knowledge no prior work has evaluated the NNSD of artificial neural networks and this is a central focus of this paper.

²The GOE corresponds to taking the independent matrix entries to be normal random variables.

³Using the Frobenius norm identity $\sum_i \lambda_i^2 = P^2 \sigma^2$

In contrast to the LSD, other k -point correlation functions are normalised such that the mean spacing between adjacent eigenvalues is unity. At this *microscopic* scale, the LSD is locally constant and equal to 1 meaning that its effect on the eigenvalues' distribution has been removed and only microscopic correlations remain. In the case of Wigner random matrices, for which the LSD varies slowly across the support of the eigenvalue distribution, this corresponds to scaling by \sqrt{P} . On this scale the limiting eigenvalue correlations when $P \rightarrow \infty$ are *universal*; that is, they are the same for wide classes of random matrices, depending only on symmetry (Guhr et al., 1998). For example, this universality is exhibited by the NNSD. Consider a 2×2 GOE matrix, in which case the j.p.d.f has a simple form:

$$p(\lambda_1, \lambda_2) \propto |\lambda_1 - \lambda_2| e^{-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)}. \quad (5)$$

Making the change of variables $\nu_1 = \lambda_1 - \lambda_2, \nu_2 = \lambda_1 + \lambda_2$, integrating out ν_2 and setting $s = |\nu_1|$ results in a density $\rho_{Wigner}(s) = \frac{\pi s}{2} e^{-\frac{\pi}{4}s^2}$, known as the *Wigner surmise* (see Figure 2). For larger matrices, the j.p.d.f must include an indicator function $\mathbb{1}\{\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_P\}$ before marginalisation so that one is studying pairs of *adjacent* eigenvalues. While the Wigner surmise can only be proved exactly, as above, for the 2×2 GOE, it holds to high accuracy for the NNSD of GOE matrices of any size provided that the eigenvalues have been scaled to give mean spacing 1.⁴ The Wigner surmise density vanishes at 0, capturing ‘repulsion’ between eigenvalues that is characteristic of RMT statistics, in contrast to the distribution of entirely independent eigenvalues given by the *Poisson law* $\rho_{Poisson}(s) = e^{-s}$. The Wigner surmise is universal in that the same density formula applies to all real-symmetric random matrices, not just the GOE or Wigner random matrices.

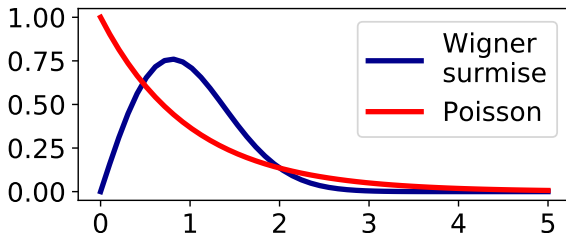


Figure 2. The density of the Wigner surmise.

4. Methodology

Prior work (Granzoli et al., 2019b; Pappan, 2018; Ghorbani et al., 2019) focusing on the Hessian empirical spectral density has utilised fast Hessian vector products (Pearlmutter,

⁴An exact formula for the NNSD of GOE matrices of any size, and one that holds in the large P limit, can be found in Mehta (2004).

1994) in conjunction with Lanczos (Meurant and Strakoš, 2006) methods. However, these methods approximate only macroscopic quantities like the spectral density, not microscopic statistics such as nearest neighbour spectral spacings. For modern neural networks, the $\mathcal{O}(P^3)$ Hessian eigendecomposition cost will be prohibitive, e.g. for a Residual Network with 34 layers $P = 10^7$. Hence, We restrict to models small enough to perform exact full Hessian computation and eigendecomposition.

We consider, single layer neural networks for classification (softmax regression), 2-hidden-layer multi-layer perceptrons (MLPs)⁵ and 3 hidden-layer MLPs⁶. On MNIST (Deng, 2012), the Hessians are of size 7850×7850 for logistic regression, 9860×9860 for the small MLP and 20060×20060 for the larger 3 hidden-layer MLP, so can be computed exactly by simply applying automatic differentiation twice, and the eigenvalues can be computed exactly in a reasonable amount of time. We also consider a single layer applied to CIFAR10 (Krizhevsky et al., 2009) classification with pre-trained Resnet-34 embedding features. While we cannot at present study the full Hessian of, for example, a Resnet-34, we can study the common transfer learning use-case of training only the final layer on some particular task. The Hessians can be computed at any data point or over any collection of data points. We consider Hessians computed over the entire datasets in question, and over batches of size 64. We separately consider test and train sets.

Training details: All networks were trained using SGD for 300 epochs with initial learning rate 0.003, linear learning rate decay to 0.00003 between epoch 150 and 270, momentum 0.9 and weight decay 5×10^{-4} . We use a PyTorch (Paszke et al., 2017) implementation. Full code to reproduce our results is made available ⁷.

5. Spectral spacing statistics in RMT

Consider a random $P \times P$ matrix M_P with ordered $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_P$. Let I_{ave} be the mean spectral cumulative density function for the random matrix ensemble from which M_P is drawn. The *unfolded spectrum* is defined as

$$l_i = I_{ave}(\lambda_i). \quad (6)$$

The unfolded spacings are then defined as

$$s_i = l_i - l_{i-1}, \quad i = 2, \dots, P. \quad (7)$$

With this definition, the mean of the s_i is unity, which means that this transformation has brought the eigenvalues on to the

⁵Hidden layer widths: 10, 100.

⁶Hidden layer widths: 10, 100, 100.

⁷<https://github.com/npbaskerville/dnn-rmt-spacings>

microscopic scale on which universal spectral spacing statistics emerge. We are investigating the presence of Random Matrix Theory statistics in neural networks by considering the nearest neighbour spectral spacings of their Hessians. Within the Random Matrix Theory literature, it has been repeatedly observed (Bohigas, 1991; Berry et al., 1987) that the unfolded spacings of a matrix with RMT pair correlations follow universal distributions determined only by the symmetry class of the M_P . Hessians are real symmetric, so the relevant universality class is GOE and therefore the unfolded neural network spacings should be compared to the Wigner surmise

$$\rho_{Wigner}(s) = \frac{\pi s}{2} e^{-\frac{\pi}{4}s^2}. \quad (8)$$

A collection of unfolded spacings s_2, \dots, s_P from a matrix with GOE spacing statistics should look like a sample of i.i.d. draws from the Wigner surmise density (8). For some known random matrix distributions, I_{ave} may be available explicitly, or at least via highly accurate quadrature methods from a known mean spectral density. For example, for the $P \times P$ GOE (Abuelenin and Abul-Magd, 2012) $I_{ave}^{GOE}(\lambda)$ is given by:

$$P \left[\frac{1}{2} + \frac{\lambda}{2\pi P} \sqrt{2P - \lambda^2} + \frac{1}{\pi} \arctan \left(\frac{\lambda}{\sqrt{2P - \lambda^2}} \right) \right]. \quad (9)$$

However, when dealing with experimental data where the mean spectral density is unknown, one must resort to using an approximation to I_{ave} . Various approaches are used in the literature, including polynomial spline interpolation (Abuelenin and Abul-Magd, 2012). The approach of (Scholak et al., 2014; Scholak, 2015) is most appropriate in our case, since computing Hessians over many mini-batches of data results in a large pool of spectra which can be used to accurately approximate I_{ave} simply by the empirical cumulative density. Suppose that we have m samples $(M_P^{(i)})_{i=1}^m$ from a random matrix distribution over symmetric $P \times P$ matrices. Fix some integers $m_1, m_2 > 0$ such that $m_1 + m_2 = m$. The spectra of the matrices $(M_P^{(i)})_{i=1}^{m_1}$ can then be used to construct an approximation to I_{ave} . More precisely, let Λ_1 be the set of all eigenvalues of the $(M_P^{(i)})_{i=1}^{m_1}$, then we define

$$\tilde{I}_{ave}(\lambda) = \frac{1}{|\Lambda_1|} |\{\lambda' \in \Lambda_1 \mid \lambda' < \lambda\}|. \quad (10)$$

For each of the matrices $(M_P^{(i)})_{i=m_1+1}^m$, one can then use \tilde{I}_{ave} to construct their unfolded spacings. When the matrix size P is small, one can only study the spectral spacing distribution by looking over multiple matrix samples. However, the same spacing distribution is also present for a single matrix in the large P limit. A clear disadvantage of

studying unfolded nearest neighbour spectral spacings with the above methods is the need for a reasonably large number of independent matrix samples. This rules-out studying the unfolded spacings of a single large matrix. Another obvious disadvantage is the introduction of error by the approximation of I_{ave} , giving the opportunity for local spectral statistics to be distorted or destroyed. An alternative statistic is the consecutive spacing ratio of (Atas et al., 2013). In the above notation, the ratios for a single $P \times P$ matrix are defined as

$$r_i = \frac{\lambda_i - \lambda_{i-1}}{\lambda_{i-1} - \lambda_{i-2}}, \quad 2 \leq i \leq P. \quad (11)$$

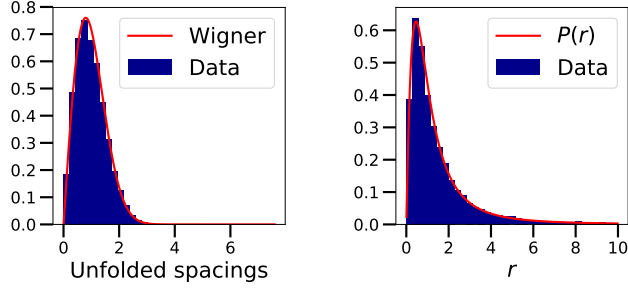
Atas et al. (2013) proved a ‘Wigner-like surmise’ for the spacing ratios, which for the GOE is

$$P(r) = \frac{27(r + r^2)}{8(1 + r + r^2)^{5/2}}. \quad (12)$$

In our experiments, we can compute the spacing ratios for Hessians computed over entire datasets or over batches, whereas the unfolded spacing ratios can only be computed in the batch setting, in which case $\frac{2}{3}$ of the batch Hessians are reserved for computing \tilde{I}_{ave} and the remaining $\frac{1}{3}$ are unfolded and analysed.

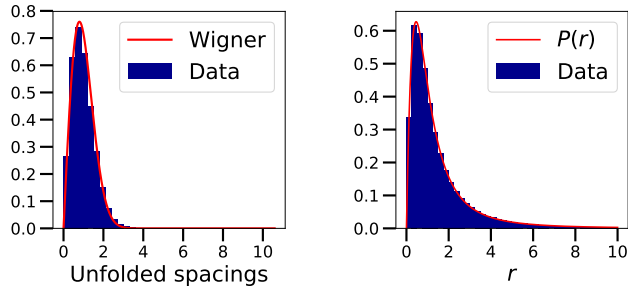
6. Results

We display results as histograms of data along with a plot of the Wigner (or the Wigner-like) surmise density. We make a few practical adjustments to the plots. Spacing ratios are truncated above some value, as the presence of a few extreme outliers makes visualisation difficult. We choose a cut-off at 10. Note that around 0.985 of the mass of the Wigner-like surmise is below 10, so this is a reasonable adjustment. The Hessians have degenerate spectra. The Wigner surmise is not a good fit to the observed unfolded spectra if the zero eigenvalues are retained. Imposing a lower cut-off of 10^{-20} in magnitude is sufficient to obtain agreement with Wigner. This is below the machine precision, so these omitted eigenvalues are indistinguishable from 0. We show results in Figures 3 and 4, with further plots in the supplementary material. We also considered randomly initialised networks and we evaluated the Hessians over train and test datasets separately in all cases. Unfolded spacings were computed only for Hessians evaluated on batches of 64 data points, while spacing ratios were computed in batches and over the entire dataset. We observe a striking level of agreement between the observed spectra and the GOE. The agreement is arguably stronger for the spacing ratio statistic than the unfolded spacings, as expected due to the approximation necessary in unfolding the spectra. There was no discernible difference between the train and test conditions, nor between batch and full dataset conditions, nor between



(a) Unfolded spacings. Batch-size 64. (b) Spacing ratios. Entire dataset.

Figure 3. Spacing distributions for the Hessian of a logistic regression trained Resnet-34 embeddings of CIFAR10. Hessians computed over the test set.



(a) Unfolded spacings. Batch-size 64. (b) Spacing ratios. Batch-size 64.

Figure 4. Spacing distributions for the Hessian of a 3-hidden-layer MLP trained on MNIST. Hessians computed over the test set.

trained and untrained models. Note that the presence of GOE statistics for the untrained models is not a foregone conclusion. Of course, the weights of the model are indeed random Gaussian, but the Hessian is still a function of the data set, so it is not the case the Hessian eigenvalue statistics are bound to be GOE a priori. Overall, the very close agreement between Random Matrix Theory predictions and our observations for several different architectures, model sizes and datasets demonstrates a clear presence of RMT statistics in neural networks.

7. A Model for the True Loss

Having introduced the empirical and true loss in Section 2, we make a further definition

$$\epsilon_{\mathcal{D}}(\mathbf{w}) = \mathcal{L}_{emp}(\mathbf{w}, \mathcal{D}) - \mathcal{L}_{true}(\mathbf{w}). \quad (13)$$

$\epsilon_{\mathcal{D}}$ is defined to be the random component of \mathcal{L}_{emp} induced by the sample \mathcal{D} . Dropping the dependence on \mathcal{D} gives

$$\mathcal{L}_{emp}(\mathbf{w}) = \mathcal{L}_{true}(\mathbf{w}) + \epsilon(\mathbf{w}) \quad (14)$$

where \mathcal{L}_{true} is a deterministic function and ϵ is random. This model results in the following Hessian structure

$$\mathbf{H}_{emp}(\mathbf{w}) = \mathbf{H}_{true}(\mathbf{w}) + \mathbf{E}(\mathbf{w}) \quad (15)$$

where $\mathbf{H}_{true} \in \mathbb{R}^{P \times P}$ is a deterministic symmetric matrix and $\mathbf{E} \in \mathbb{R}^{P \times P}$ is a random symmetric matrix. Our investigations have revealed that, whatever the spectral density of deep network Hessians may be, the microscopic correlations between nearby eigenvalues do appear to follow GOE statistics. This gives us significant freedom when constructing putative models for the loss surfaces of neural networks. The presence of GOE spacings is a critical feature and RMT teaches us to expect these statistics to be universal across *all* neural networks and data sets. Any model for neural network loss surfaces that we construct must exhibit GOE Hessian spacing statistics, but this is a far less restrictive condition than any assumptions about the Hessian’s spectral density. These considerations lead us to propose the following model:

1. \mathbf{H}_{true} has some fixed rank r which is not extensive in P ;
2. ϵ is a Gaussian process $\mathcal{GP}(0, k)$, where k is some kernel function.

With appropriate scaling of \mathcal{L}_{true} vs ϵ , this model will produce a bulk-and-spikes spectral density for \mathbf{H}_{emp} . By choosing k to act on a some low-dimensional subspace of \mathbb{R}^P (but dimension still extensive in P), rank degeneracy in the bulk can be obtained. Finally, any choice of k will result in GOE spacing statistics in the bulk. As an example, taking $k(\mathbf{w}, \mathbf{w}') \propto (\mathbf{w}^T \mathbf{w}')^p$ and restricting \mathbf{w} to a hypersphere results in ϵ taking the exact form of a spherical p -spin glass, a model repeatedly studied in the context of neural network loss surfaces and gradient descent in high-dimensions (Choromanska et al., 2015a; Gardner and Derrida, 1988; Mezard et al., 1987; Ros et al., 2019; Mannelli et al., 2019). Following from our Gaussian process definition, the covariance of derivatives of the empirical loss can be computed using a well-known result (see Adler and Taylor (2009) equation 5.5.4), e.g.

$$\text{Cov}(\partial_i \mathcal{L}_{emp}(\mathbf{w}), \partial_j \mathcal{L}_{emp}(\mathbf{w}')) = \partial_{w_i} \partial_{w'_j} k(\mathbf{w}, \mathbf{w}')$$

and further, assuming a stationary kernel $k(\mathbf{w}, \mathbf{w}') = k(-\frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2)$ (note abuse of notation)

$$\begin{aligned} & \text{Cov}(\partial_i \mathcal{L}_{emp}(\mathbf{w}), \partial_j \mathcal{L}_{emp}(\mathbf{w}')) \\ &= (w_i - w'_i)(w'_j - w_j) k'' \left(-\frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2 \right) \\ &+ \delta_{ij} k' \left(-\frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2 \right). \end{aligned} \quad (16)$$

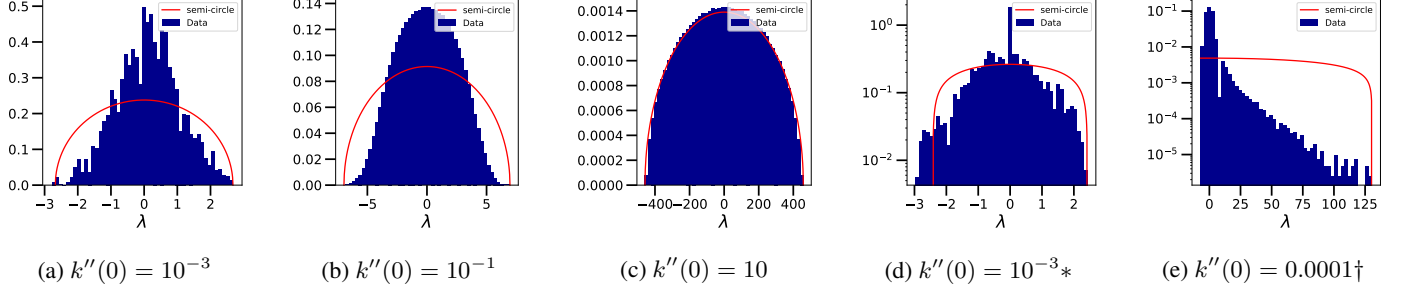


Figure 5. Spectral densities of Gaussian process Hessians with various kernel choices. All use $k'(0) = 0$. The dimension is 300 in all cases except (d), in which the Hessian is padded to 400 dimensions with zeros. All histograms are produced with 100 independent Hessian samples. * = 100 degenerate directions. † = 20 outliers

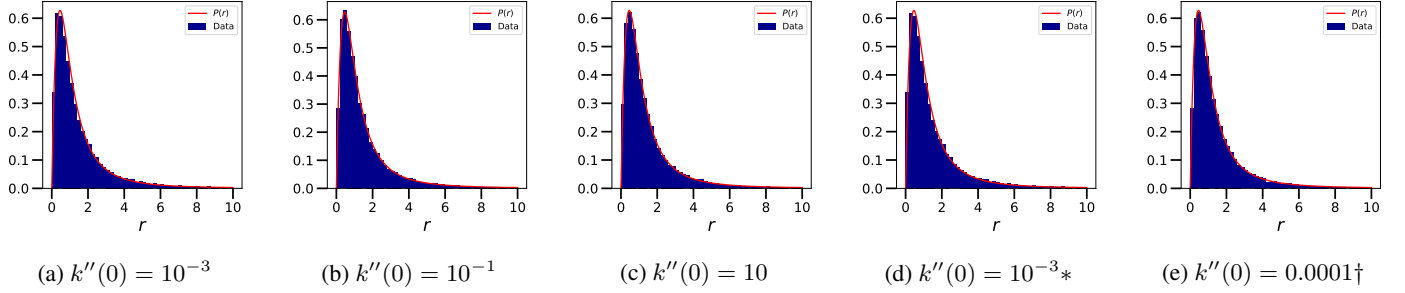


Figure 6. Consecutive spacing ratios of Gaussian process Hessians with various kernel choices. All use $k'(0) = 1$. The dimension is 300 in all cases except (d), in which the Hessian is padded to 400 dimensions with zeros. * = 100 degenerate directions. † = 20 outliers.

Differentiating (16) further, we obtain

$$\begin{aligned} \text{Cov}(\partial_{ij}\mathcal{L}_{\text{emp}}(\mathbf{w}), \partial_{kl}\mathcal{L}_{\text{emp}}(\mathbf{w})) \\ = k''(0) (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + k'(0)^2 \delta_{ij}\delta_{kl} \end{aligned} \quad (17)$$

\mathbf{E} has Gaussian entries with mean zero, so the distribution of \mathbf{E} is determined entirely by $k'(0)$ and $k''(0)$. Neglecting to choose k explicitly, we vary the values of $k'(0)$ and $k''(0)$ to produce nearest neighbour spectral spacings ratios and spectral densities. The histograms for spectral spacing ratios are indistinguishable and agree very well with the GOE, as shown in Figure 6. The spectral densities are shown in Figure 5, including examples with rank degeneracy and outliers.

8. Iterate Averaging & Generalisation

The Iterate Average (IA) (Polyak and Juditsky, 1992) is defined as the average of the model parameters over the model optimisation trajectory $\mathbf{w}_{\text{IA}} = \frac{1}{n} \sum_i^n \mathbf{w}_i$. It is a classical variance reducing technique in optimisation with optimal asymptotic convergence rates and greater robustness to the choice of learning rate (Kushner and Yin, 2003). Popular regret bounds that form the basis of gradient-based convergence proofs (Duchi et al., 2011; Reddi et al., 2019) only imply convergence for the Iterate Average (Duchi, 2018).

One of the problems with IA is that for networks with Batch Normalisation (Ioffe and Szegedy, 2015), simply averaging

the Batch Normalisation statistics is known to lead to poor results (Defazio and Bottou, 2019). However, by computing the batch normalisation statistics for the iterate average using a forward pass of the data at the IA point, Izmailov et al. (2018) show that the performance of small-scale image experiments such as CIFAR-10/100 and pre-trained ImageNet fine-tuning can be significantly improved. They and Merity et al. (2017) show that by combining *tail averaging* (where averaging is only done at the late stages of training) along with high learning rates, generalisation performance is significantly improved. An open theoretical question is *why does IA along with large learning rates gives such improved generalisation?*. The authors of Izmailov et al. (2018) argue that IA brings the solution to areas which are flatter and generalise better, however Dinh et al. (2017) show that sharpness can be arbitrarily manipulated without altering the loss value. Another theoretical argument due to Granzio et al. (2020b) is,

Theorem 8.1. Suppose $\nabla \mathcal{L}_{\text{emp}} - \nabla \mathcal{L}_{\text{true}} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then

$$\begin{aligned} \mathbb{P}\left\{\|\mathbf{w}_n\| - \sqrt{\sum_i^P w_{0,i}^2 e^{-2n\alpha\lambda_i} + \frac{P\alpha\sigma^2}{B} \left\langle \frac{1}{\lambda(2-\alpha\lambda)} \right\rangle} \geq t\right\} \leq \nu \\ \mathbb{P}\left\{\|\mathbf{w}_{\text{avg}}\| - \sqrt{\sum_i^P \frac{w_{0,i}^2}{\lambda_i^2 n^2 \alpha^2} + \frac{P\alpha\sigma^2}{Bn} \left\langle \frac{1}{\lambda^2} \right\rangle} \geq t\right\} \leq \nu \end{aligned} \quad (18)$$

where $\nu = 2 \exp(-ct^2)$, and $\langle \lambda^k \rangle = \frac{1}{P} \text{Tr} \mathbf{H}^k$. $\mathbf{H}_{emp} = \nabla^2 \mathcal{L}_{emp}$ is the Hessian of the empirical loss w.r.t weights and B is the batch size.

However, as pointed out in [Granziol et al. \(2020b\)](#), the assumption of an independent perturbation between the true gradient and empirical gradient is hard to justify. Intuitively under this assumption we would never experience the problem of overfitting. However they argue intuitively that for large learning rates, the perturbation between the true loss gradient and empirical gradient is approximately independent. We derive this assumption from first principles using our Gaussian process model introduced in the previous section. Our model results in a relaxation of this independence assumption. We can form a more explicit link with these notions of generalisation by establishing the following result pertaining to variance reduction in the presence of the covariance structure (16).

Theorem 8.2. Let $(\mathbf{w}_i)_{i=1}^T$ be a sequence of weights in \mathbb{R}^P . Let $(\mathbf{A}_i)_{i=1}^T$ be any sequence of symmetric matrices in $\mathbb{R}^{P \times P}$, with bounded trace norm $\|\mathbf{A}_i\|_1 < a \forall i$. Define $d_{ij} = \|\mathbf{w}_i - \mathbf{w}_j\|_2$. Assume the covariance structure (16), let $\mathbf{g}_i = \partial \mathcal{L}_{emp}(\mathbf{w}_i)$ and let $\mathbf{g}_{avg} = \frac{1}{n} \sum_{i=1}^T \mathbf{A}_i \mathbf{g}_i$, then

$$\begin{aligned} & P^{-1} \text{Tr} \text{Cov}(\mathbf{g}_{avg}) \\ & \leq \frac{a^2 k'(0)}{T} + \frac{2a^2}{T^2} \sum_{1 \leq i < j \leq T} \left\{ k'(-\frac{d_{ij}^2}{2}) - \frac{1}{P} k''(-\frac{d_{ij}^2}{2}) d_{ij}^2 \right\}. \end{aligned} \quad (19)$$

Proof. Each of the \mathbf{g}_i is Gaussian distributed with covariance matrix $\text{Cov}(\mathbf{g}_i)$ given by (16) and the covariance between different gradients $\text{Cov}(\mathbf{g}_i, \mathbf{g}_j)$ is similarly given by (16). By standard multivariate Gaussian properties

$$\begin{aligned} \text{Cov}(\mathbf{g}_{avg}) &= \frac{1}{T^2} \sum_{i=1}^T \mathbf{A}_i \text{Cov}(\mathbf{g}_i) \mathbf{A}_i^T \\ &+ \frac{1}{T^2} \sum_{i \neq j} \mathbf{A}_i \text{Cov}(\mathbf{g}_i, \mathbf{g}_j) \mathbf{A}_j^T, \end{aligned}$$

then taking the trace and recalling \mathbf{A}_i are symmetric

$$\begin{aligned} \text{Tr} \text{Cov}(\mathbf{g}_{avg}) &= \frac{1}{T^2} \sum_{i=1}^T \text{Tr}(\mathbf{A}_i^2 \text{Cov}(\mathbf{g}_i)) \\ &+ \frac{2}{T^2} \sum_{1 \leq i < j \leq T} \text{Tr}(\mathbf{A}_i \mathbf{A}_j \text{Cov}(\mathbf{g}_i, \mathbf{g}_j)). \end{aligned}$$

Using the trace norm bounds

$$\begin{aligned} & \text{Tr} \text{Cov}(\mathbf{g}_{avg}) \\ & \leq \frac{a^2}{T^2} \sum_{i=1}^T \text{Tr}(\text{Cov}(\mathbf{g}_i)) + \frac{2a^2}{T^2} \sum_{1 \leq i < j \leq T} \text{Tr}(\text{Cov}(\mathbf{g}_i, \mathbf{g}_j)). \end{aligned} \quad (20)$$

Using the covariance structure (16), noting that

$$\text{Tr}[(\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^T] = \|\mathbf{w}_i - \mathbf{w}_j\|_2^2$$

and dividing through by P , the result follows. \square

As a corollary, consider large T and P and suppose all but $o(T^2)$ of the d_{ij} are extensive in P , i.e. $d_{ij} \sim P^\eta$ for some $0 < \eta \leq \frac{1}{2}$. The first term in (19) scales simply with T^{-1} , whereas the second term scales like

$$\frac{a^2(T-1)}{T} \left[k'(-\frac{1}{2}P^{2\eta}) - P^{2\eta-1} k''(-\frac{1}{2}P^{2\eta}) \right]. \quad (21)$$

For k with appropriately decaying first and second derivatives, this term decays with large P uniformly in T . Taking $\mathbf{A}_i = \alpha(1 - \lambda\alpha)^i \mathbf{I}$, for learning rate α and weight decay λ , we recover the iterate averaging variance reduction of ([Granziol et al., 2020b](#)), but only for large P , and with a greater reduction with larger P uniformly in T and vice-versa. The d_{ij} size condition can be satisfied by using a large learning rate, and so our model predicts a variance reducing effect of iterate averaging with large learning rates and and better variance reduction for larger networks. Similar arguments could be made for adaptive methods by taking $\mathbf{A}_i \propto \mathbf{H}_{emp}^{-1}$, which are shown in [Granziol et al. \(2020b\)](#) to give very strong results.

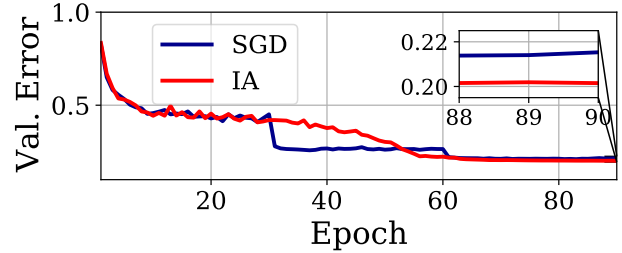


Figure 7. WideResNet-101 Val. Error for ImageNet SGD & IA. Our results for IA underpin very strong results experimental results, as shown in Fig 7 for ImageNet on the WideResNet-101, where for the same number of training epochs we achieve nearly a 2% reduction in validation error by using tail averaging.

8.1. Exponential Hardness of the True Loss Minimum

Previous work considering spin glass models of neural networks ([Choromanska et al., 2015a](#); [Baity-Jesi et al., 2018](#)), has focused exclusively on the training loss. Practitioners typically closely follow and indirectly optimise the validation or held out test set loss/error, which can be seen as an unbiased estimate of the true loss/error. Whilst [Sagun et al. \(2014\)](#) argue using the central limit theorem that both the training and test loss converge to the same quantity, it is well known that training and testing dynamics can differ significantly and hence unclear if the limiting assumptions even approximately hold.

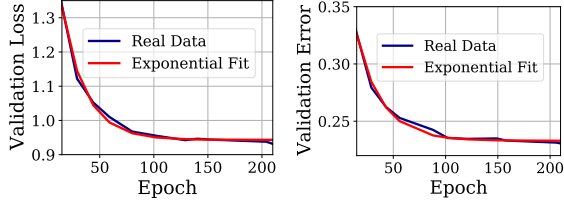


Figure 8. **Exponential Hardness of Attaining the True Loss Minimum:** ResNet-50 on ImageNet Lowest Validation Loss/Error against Epoch attained

ImageNet vs MNIST: For small scale datasets such as MNIST/CIFAR with only 50,000 samples and a minimal amount of data augmentation⁸, since the total number of augmentations due to cropping is given by $(P_I - P_C + 1)^2$, where P_I, P_C refer to the number of pixels in the original image and cropped version respectively, the total number of samples seen by the optimiser is very small and hence the problem of *overfitting* can be severe. For large scale datasets such as ImageNet (Deng et al., 2009), with over 1m images with $P_I, P_C = [256, 224]$, the number of samples seen is vastly larger. Whilst due to the sample dependence it is hard to characterise exactly how many effective samples are seen by the optimiser, we expect similarly sized networks to have very different dynamics on such large scale datasets. Specifically we expect much less divergence between the training and true loss. This opens up an interesting avenue of research. Given that for very small datasets, Choromanska et al. (2015a) show that there are many minima all approximately equivalent on the test set and Baity-Jesi et al. (2018) show that the final stages of the training loss do not display the exponential time-scale typical of barrier crossing, *do we get different results for the true loss?*

We scale the training procedure of He et al. (2016), on the ResNet-50 using ImageNet and SGD with runs containing 15, 30, 45, 60, 90, 120, 150, 180, 210, 270, 300 epochs. The best validation loss closely follows an exponential drop, as predicted by spin-glass theory. The test error also closely follows an exponential drop, further motivating the importance of studying the true loss. We note that the best validation error of the 300 epoch regime is 22.8% as opposed to 24.2% for the 90 epoch regime and hence *not all local minima are approximately equivalent on the test set*. In fact by considering spin glass theory on the true instead of empirical loss we recover a well known tenet of deep learning practice. *Whilst the returns are diminishing, by continually training neural networks results can continue to improve.*

⁸usually images are cropped to 28×28 from 32×32

9. Conclusion and future work

We have demonstrated experimentally the existence of random matrix statistics in small neural networks on the scale of the mean eigenvalue separation. This provides the first direct evidence of universal RMT statistics present in neural networks trained on real datasets. It means that, in practice, when working with a neural network on some dataset, one has information a priori about the local correlations between Hessian eigenvalues. We focus on small neural networks where Hessian eigendecomposition is feasible. Future research that our work motivates could develop methods to approximate the level spacing distribution of large deep neural networks for which exact Hessian spectra cannot be computed. If the same RMT statistics are found, this would constitute a profound universal property of neural networks models; conversely, a break-down in these RMT statistics would be a fascinating indication of some fundamental separation between different network sizes or architectures. One intriguing possible avenue is the relation to chaotic systems. Quantum systems with chaotic classical limits are known to display RMT spectral pairwise correlations, whereas Poisson statistics correspond to integrable systems. We suggest that the presence of GOE pairwise correlations in neural network Hessians, as opposed to Poisson, indicates that neural network training dynamics cannot be reduced to some simpler, smaller set of dynamical equations. Furthermore we have shown how by evaluating the *true* instead of *empirical* loss, known results from spin-glass systems can be recovered (exponential hardness of attaining the global minimum) and how this has practical implications for achieving state of the art performance. Our model for the empirical loss, which unlike prior theoretical work is fully compatible with outliers and rank degeneracy extensively observed in the literature, whilst still being compatible with GOE level spacings. Furthermore, this model predicts the independence of gradients with increased weight-space distance which justifies the generalisation benefit of iterate averaging.

Acknowledgements

JPK is pleased to acknowledge support from a Royal Society Wolfson Research Merit Award and ERC Advanced Grant 740900 (LogCorRM). DMG is grateful for the support from the JADE computing facility and in particular the extensive support of Andrew Gittings. NPB is grateful for the support of the Advanced Computing Research Centre of the University of Bristol. Furthermore the authors would like to thank Samuel Albanie for extensive discussions on the exponential hardness of the true loss.

References

- Sherif M Abuelenin and Adel Y Abul-Magd. Effect of unfolding on the spectral statistics of adjacency matrices of complex networks. *Procedia Computer Science*, 12: 69–74, 2012.
- Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.
- Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.
- YY Atas, E Bogomolny, O Giraud, and G Roux. Distribution of the ratio of consecutive level spacings in random matrix ensembles. *Physical review letters*, 110(8):084101, 2013.
- Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. *risk*, 1(1.5):2–0, 2020.
- Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gérard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2018.
- Nicholas P. Baskerville, Jonathan P. Keating, Francesco Mezzadri, and Joseph Najnudel. The loss surfaces of neural networks with general activation functions. *arXiv preprint arXiv:2004.03959*, 2020.
- Nicholas P Baskerville, Jonathan P Keating, Francesco Mezzadri, and Joseph Najnudel. A spin-glass model for the loss surfaces of generative adversarial networks. *arXiv preprint arXiv:2101.02524*, 2021.
- Carlo WJ Beenakker. Random-matrix theory of quantum transport. *Reviews of modern physics*, 69(3):731, 1997.
- Michael V Berry et al. Quantum chaos. *Proc. Roy. Soc. London A*, 413:183–198, 1987.
- Oriol Bohigas. Random matrix theories and chaotic dynamics. Technical report, Paris-11 Univ., 1991.
- Léon Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL https://doi.org/10.1007/978-3-642-35289-8_25.
- Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.
- Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, 2005.
- Pratik Chaudhari and Stefano Soatto. On the energy landscape of deep networks. *arXiv preprint arXiv:1511.06485*, 2015.
- Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015a.
- Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760, 2015b.
- Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pages 1753–1763, 2019.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive sub-gradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul): 2121–2159, 2011.
- John C Duchi. Introductory lectures on stochastic optimization. *The Mathematics of Data*, 25:99, 2018.
- Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.
- Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. *arXiv preprint arXiv:1901.10159*, 2019.

- Diego Granzio. Beyond random matrix theory for deep networks. *arXiv preprint arXiv:2006.07721*, 2020.
- Diego Granzio, Timur Garipov, Dmitry Vetrov, Stefan Zohren, Stephen Roberts, and Andrew Gordon Wilson. Towards understanding the true loss surface of deep neural networks using random matrix theory and iterative spectral methods. 2019a.
- Diego Granzio, Xingchen Wan, Timur Garipov, Dmitry Vetrov, and Stephen Roberts. MLRG deep curvature. *arXiv preprint arXiv:1912.09656*, 2019b.
- Diego Granzio, Timur Garipov, Dmitry Vetrov, Stefan Zohren, Stephen Roberts, and Andrew Gordon Wilson. Towards understanding the true loss surface of deep neural networks using random matrix theory and iterative spectral methods, 2020a. URL <https://openreview.net/forum?id=H1gza2NtwH>.
- Diego Granzio, Xingchen Wan, and Stephen Roberts. Iterate averaging helps: An alternative perspective in deep learning. *arXiv preprint arXiv:2003.01247*, 2020b.
- Thomas Guhr, Axel Müller-Groeling, and Hans A Weidenmüller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(4-6):189–425, 1998.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.
- Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. *arXiv preprint arXiv:1902.00139*, 2019.
- Madan Lal Mehta. *Random matrices*. Elsevier, 2004.
- Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*, 2017.
- Gérard Meurant and Zdeněk Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542, 2006.
- Marc Mezard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.
- Vardan Papayan. The full spectrum of deepnet Hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.
- Vardan Papayan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet Hessians. *arXiv preprint arXiv:1901.08244*, 2019.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in Pytorch. 2017.
- Barak A Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1):147–160, 1994.
- Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2798–2806. JMLR. org, 2017.
- Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
- Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019.
- Levent Sagun, V Ugur Guney, Gerard Ben Arous, and Yann LeCun. Explorations on high dimensional landscapes. *arXiv preprint arXiv:1412.6615*, 2014.
- Levent Sagun, Léon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Torsten Scholak. unfoldr, 2015. URL <https://github.com/tscholak/unfoldr>. Accessed 30/10/2020.

Torsten Scholak, Thomas Wellens, and Andreas Buchleitner. Spectral backbone of excitation transport in ultracold rydberg gases. *Physical Review A*, 90(6):063415, 2014.

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

HA Weidenmuller and GE Mitchell. Random matrices and chaos in nuclear physics. *arXiv preprint arXiv:0807.1070*, 2008.