Geometry of Neural Network Loss Surfaces via Random Matrix Theory

Jeffrey Pennington[1] Yasaman Bahri[1]

Abstract

Understanding the geometry of neural network loss surfaces is important for the development of improved optimization algorithms and for building a theoretical understanding of why deep learning works. In this paper, we study the geometry in terms of the distribution of eigenvalues of the Hessian matrix at critical points of varying energy. We introduce an analytical framework and a set of tools from random matrix theory that allow us to compute an approximation of this distribution under a set of simplifying assumptions. The shape of the spectrum depends strongly on the energy and another key parameter, ϕ, which measures the ratio of parameters to data points. Our analysis predicts and numerical simulations support that for critical points of small index, the number of negative eigenvalues scales like the $3/2$ power of the energy. We leave as an open problem an explanation for our observation that, in the context of a certain memorization task, the energy of minimizers is well-approximated by the function $\frac{1}{2}(1-\phi)^2$.

1. Introduction

Neural networks have witnessed a resurgence in recent years, with a smorgasbord of architectures and configurations designed to accomplish ever more impressive tasks. Yet for all the successes won with these methods, we have managed only a rudimentary understanding of *why* and *in what contexts* they work well. One difficulty in extending our understanding stems from the fact that the neural network objectives are generically non-convex functions in high-dimensional parameter spaces, and understanding their loss surfaces is a challenging task. Nevertheless, an improved understanding of the loss surface could have a large impact on optimization (Saxe et al.; Dauphin et al., 2014; Choromanska et al., 2015; Neyshabur et al., 2015), architecture design, and generalization (Keskar et al.).

1.1. Related work

There is no shortage of prior work focused on the loss surfaces of neural networks. Choromanska et al. (2015) and Dauphin et al. (2014) highlighted the prevalence of saddle points as dominant critical points that plague optimization, as well as the existence of many local minima at low loss values. Dauphin et al. (2014) studied the distribution of critical points as a function of the loss value empirically and found a trend which is qualitatively similar to predictions for random Gaussian landscapes (Bray & Dean, 2007). Choromanska et al. (2015) argued that the loss function is well-approximated by a spin-glass model studied in statistical physics, thereby predicting the existence of local minima at low loss values and saddle points at high loss values as the network increases in size. Goodfellow et al. observed that loss surfaces arising in practice tend to be smooth and seemingly convex along low-dimensional slices. Subsequent works (Kawaguchi, 2016; Safran & Shamir, 2016; Freeman & Bruna, 2016) have furthered these and related ideas empirically or analytically, but it remains safe to say that we have a long way to go before a full understanding is achieved.

1.2. Our contributions

One shortcoming of prior theoretical results is that they are often derived in contexts far removed from practical neural network settings – for example, some work relies on results for generic random landscapes unrelated to neural networks, and other work draws on rather tenuous connections to spin-glass models. While there is a lot to be gained from this type of analysis, it leaves open the possibility that characteristics of loss surfaces specific to neural networks may be lost in the more general setting. In this paper, we focus narrowly on the setting of neural network loss surfaces and propose an analytical framework for studying the spectral density of the Hessian matrix in this context.

Our bottom-up construction assembles an approximation of the Hessian in terms of blocks derived from the weights, the data, and the error signals, all of which we assume to

[1] Google Brain. Correspondence to: Jeffrey Pennington <jpennin@google.com>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

be random variables[1]. From this viewpoint, the Hessian may be understood as a *structured* random matrix and we study its eigenvalues in the context of random matrix theory, using tools from free probability. We focus on single-hidden-layer networks, but in principle the framework can accommodate any network architecture. After establishing our methodology, we compute approximations to the Hessian at several levels of refinement. One result is a prediction that for critical points of small index, the index scales like the energy to the 3/2 power.

2. Preliminaries

Let $W^{(1)} \in \mathbb{R}^{n_1 \times n_0}$ and $W^{(2)} \in \mathbb{R}^{n_2 \times n_1}$ be weight matrices of a single-hidden-layer network without biases. Denote by $x \in \mathbb{R}^{n_0 \times m}$ the input data and by $y \in \mathbb{R}^{n_2 \times m}$ the targets. We will write $z^{(1)} = W^{(1)} x$ for the pre-activations. We use $[\cdot]_+ = \max(\cdot, 0)$ to denote the ReLU activation, which will be our primary focus. The network output is

$$\hat{y}_{i\mu} = \sum_{k=1}^{n_1} W^{(2)}_{ik} [z^{(1)}_{k\mu}]_+ \,, \tag{1}$$

and the residuals are $e_{i\mu} = \hat{y}_{i\mu} - y_{i\mu}$. We use Latin indices for features, hidden units, and outputs, and μ to index examples. We consider mean squared error, so that the loss (or energy) can be written as,

$$\mathcal{L} = n_2 \epsilon = \frac{1}{2m} \sum_{i,\mu=1}^{n_2,m} e^2_{i\mu} \,. \tag{2}$$

The Hessian matrix is the matrix of second derivatives of the loss with respect to the parameters, i.e. $H_{\alpha\beta} = \frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta}$, where $\theta_\alpha \in \{W^{(1)}, W^{(2)}\}$. It decomposes into two pieces, $H = H_0 + H_1$, where H_0 is positive semi-definite and where H_1 comes from second derivatives and contains all of the explicit dependence on the residuals. Specifically,

$$[H_0]_{\alpha\beta} \equiv \frac{1}{m} \sum_{i,\mu=1}^{n_2,m} \frac{\partial \hat{y}_{i\mu}}{\partial\theta_\alpha} \frac{\partial \hat{y}_{i\mu}}{\partial\theta_\beta} \equiv \frac{1}{m}[JJ^T]_{\alpha\beta} \,, \tag{3}$$

where we have introduced the Jacobian matrix, J, and,

$$[H_1]_{\alpha\beta} \equiv \frac{1}{m} \sum_{i,\mu=1}^{n_2,m} e_{i\mu} \left(\frac{\partial^2 \hat{y}_{i\mu}}{\partial\theta_\alpha \partial\theta_\beta} \right) . \tag{4}$$

We will almost exclusively consider square networks with $n \equiv n_0 = n_1 = n_2$. We are interested in the limit of large networks and datasets, and in practice they are typically of the same order of magnitude. A useful characterization of the network capacity is the ratio of the number of parameters to the effective number of examples[2], $\phi \equiv 2n^2/mn = 2n/m$. As we will see, ϕ is a critical parameter which governs the shape of the distribution. For instance, eqn. (3) shows that H_0 has the form of a covariance matrix computed from the Jacobian, with ϕ governing the rank of the result.

We will be making a variety of assumptions throughout this work. We distinguish *primary* assumptions, which we use to establish the foundations of our analytical framework, from *secondary* assumptions, which we invoke to simplify computations. To begin with, we establish our primary assumptions:

1. The matrices H_0 and H_1 are *freely independent*, a property we discuss in sec. 3.
2. The residuals are i.i.d. normal random variables with tunable variance governed by ϵ, $e_{i\mu} \sim \mathcal{N}(0, 2\epsilon)$. This assumption allows the gradient to vanish in the large m limit, specifying our analysis to critical points.
3. The data features are i.i.d. normal random variables.
4. The weights are i.i.d. normal random variables.

We note that normality may not be strictly necessary. We will discuss these assumptions and their validity in sec. 5.

3. Primer on Random Matrix Theory

In this section we highlight a selection of results from the theory of random matrices that will be useful for our analysis. We only aim to convey the main ideas and do not attempt a rigorous exposition. For a more thorough introduction, see, e.g., (Tao, 2012).

3.1. Random matrix ensembles

The theory of random matrices is concerned with properties of matrices M whose entries M_{ij} are random variables. The degree of independence and the manner in which the M_{ij} are distributed determine the type of random matrix ensemble to which M belongs. Here we are interested primarily in two ensembles: the real *Wigner* ensemble for which M is symmetric but otherwise the M_{ij} are i.i.d.; and the real *Wishart* ensemble for which $M = XX^T$ where X_{ij} are i.i.d.. Moreover, we will restrict our attention to studying the *limiting spectral density* of M. For a random matrix $M_n \in \mathbb{R}^{n \times n}$, the *empirical spectral density* is defined as,

$$\rho_{M_n}(\lambda) = \frac{1}{n} \sum_{j=1}^{n} \delta(\lambda - \lambda_j(M_n)) \,, \tag{5}$$

[1] Although we additionally assume the random variables are independent, our framework does not explicitly require this assumption, and in principle it could be relaxed in exchange for more technical computations.

[2] In our context of squared error, each of the n_2 targets may be considered an effective example.

where the $\lambda_j(M_n)$, $j = 1, \ldots, n$, denote the n eigenvalues of M_n, including multiplicity, and $\delta(z)$ is the Dirac delta function centered at z. The limiting spectral density is defined as the limit of eqn. (5) as $n \to \infty$, if it exists.

For a matrix M of the real Wigner matrix ensemble whose entries $M_{ij} \sim \mathcal{N}(0, \sigma^2)$, Wigner (1955) computed its limiting spectral density and found the semi-circle law,

$$\rho_{\mathrm{SC}}(\lambda; \sigma, \phi) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2} & \text{if } |\lambda| \le 2\sigma \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

Similarly, if $M = XX^T$ is a real Wishart matrix with $X \in \mathbb{R}^{n\times p}$ and $X_{ij} \sim \mathcal{N}(0, \sigma^2/p)$, then the limiting spectral density can be shown to be the Marchenko-Pastur distribution (Marčenko & Pastur, 1967),

$$\rho_{\mathrm{MP}}(\lambda; \sigma, \phi) = \begin{cases} \rho(\lambda) & \text{if } \phi < 1 \\ (1 - \phi^{-1})\delta(\lambda) + \rho(\lambda) & \text{otherwise} \end{cases} , \tag{7}$$

where $\phi = n/p$ and,

$$\begin{aligned} \rho(\lambda) &= \frac{1}{2\pi\lambda\sigma\phi}\sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)} \\ \lambda_\pm &= \sigma(1 \pm \sqrt{\phi})^2 . \end{aligned} \tag{8}$$

The Wishart matrix M is low rank if $\phi > 1$, which explains the delta function density at 0 in that case. Notice that there is an eigenvalue gap equal to λ_-, which depends on ϕ.

3.2. Free probability theory

Suppose A and B are two random matrices. Under what conditions can we use information about their individual spectra to deduce the spectrum of $A + B$? One way of analyzing this question is with free probability theory, and the answer turns out to be exactly when A and B are *freely independent*. Two matrices are freely independent if

$$\mathbf{E} f_1(A) g_1(B) \cdots f_k(A) g_k(B) = 0 \tag{9}$$

whenever f_i and g_i are such that

$$\mathbf{E} f_i(A) = 0 , \quad \mathbf{E} g_i(B) = 0 . \tag{10}$$

Notice that when $k = 1$, this is equivalent to the definition of classical independence. Intuitively, the eigenspaces of two freely independent matrices are in "generic position" (Speicher, 2009), i.e. they are not aligned in any special way. Before we can describe how to compute the spectrum of $A + B$, we must first introduce two new concepts, the Stieltjes transform and the $\mathcal{R}$ transform.

3.2.1. The Stieltjes transform

For $z \in \mathbb{C} \setminus \mathbb{R}$ the Stieltjes transform G of a probability distribution ρ is defined as,

$$G(z) = \int_{\mathbb{R}} \frac{\rho(t)}{z - t} dt , \tag{11}$$

from which ρ can be recovered using the inversion formula,

$$\rho(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G(\lambda + i\epsilon) . \tag{12}$$

3.2.2. The $\mathcal{R}$ transform

Given the Stieltjes transform G of a probability distribution ρ, the $\mathcal{R}$ transform is defined as the solution to the functional equation,

$$\mathcal{R}\big(G(z)\big) + \frac{1}{G(z)} = z . \tag{13}$$

The benefit of the $\mathcal{R}$ transform is that it linearizes free convolution, in the sense that,

$$\mathcal{R}_{A+B} = \mathcal{R}_A + \mathcal{R}_B , \tag{14}$$

if A and B are freely independent. It plays a role in free probability analogous to that of the log of the Fourier transform in commutative probability theory.

The prescription for computing the spectrum of $A + B$ is as follows: 1) Compute the Stieltjes transforms of ρ_A and ρ_B; 2) From the Stieltjes transforms, deduce the $\mathcal{R}$ transforms $\mathcal{R}_A$ and $\mathcal{R}_B$; 3) From $\mathcal{R}_{A+B} = \mathcal{R}_A + \mathcal{R}_B$, compute the Stieltjes transform G_{A+B}; and 4) Invert the Stieltjes transform to obtain ρ_{A+B}.

4. Warm Up: Wishart plus Wigner

Having established some basic tools from random matrix theory, let us now turn to applying them to computing the limiting spectral density of the Hessian of a neural network at critical points. Recall from above that we can decompose the Hessian into two parts, $H = H_0 + H_1$ and that $H_0 = JJ^T/m$. Let us make the *secondary* assumption that at critical points, the elements of J and H_1 are i.i.d. normal random variables. In this case, we may take H_0 to be a real Wishart matrix and H_1 to be a real Wigner matrix. We acknowledge that these are strong requirements, and we will discuss the validity of these and other assumptions in sec. 5.

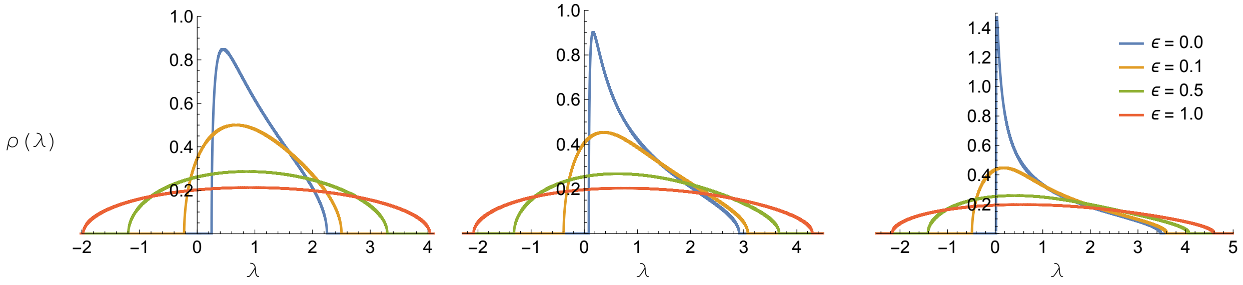
4.1. Spectral distribution

We now turn our attention to the spectral distribution of H and how that distribution evolves as the energy changes. For this purpose, only the relative scaling between H_0 and H_1 is relevant and we may for simplicity take $\sigma_{H_0} = 1$ and $\sigma_{H_1} = \sqrt{2\epsilon}$. Then we have,

$$\rho_{H_0}(\lambda) = \rho_{\mathrm{MP}}(\lambda; 1, \phi), \quad \rho_{H_1}(\lambda) = \rho_{\mathrm{SC}}(\lambda; \sqrt{2\epsilon}, \phi) , \tag{15}$$

which can be integrated using eqn. (11) to obtain G_{H_0} and G_{H_1}. Solving eqn. (13) for the $\mathcal{R}$ transforms then gives,

$$\mathcal{R}_{H_0}(z) = \frac{1}{1 - z\phi}, \quad \mathcal{R}_{H_1}(z) = 2\epsilon z . \tag{16}$$



(a) $\phi = 1/4$ (b) $\phi = 1/2$ (c) $\phi = 3/4$

Figure 1. Spectral distributions of the Wishart + Wigner approximation of the Hessian for three different ratios of parameters to data points, ϕ. As the energy ϵ of the critical point increases, the spectrum becomes more semicircular and negative eigenvalues emerge.

We proceed by computing the $\mathcal{R}$ transform of H,

$$\mathcal{R}_H = \mathcal{R}_{H_0} + \mathcal{R}_{H_1} = \frac{1}{1 - z\phi} + 2\epsilon z\,, \tag{17}$$

so that, using eqn. (13), we find that its Stieltjes transform G_H solves the following cubic equation,

$$2\epsilon\phi\, G_H^3 - (2\epsilon + z\phi)G_H^2 + (z + \phi - 1)G_H - 1 = 0\,. \tag{18}$$

The correct root of this equation is determined by the asymptotic behavior $G_H \sim 1/z$ as $z \to \infty$ (Tao, 2012). From this root, the spectral density can be derived through the Stieltjes inversion formula, eqn. (12). The result is illustrated in fig. 1. For small ϵ, the spectral density resembles the Marchenko-Pastur distribution, and for small enough ϕ there is an eigenvalue gap. As ϵ increases past a critical value ϵ_c, the eigenvalue gap disappears and negative eigenvalues begin to appear. As ϵ gets large, the spectrum becomes semicircular.

4.2. Normalized index

Because it measures the number of descent directions[3], one quantity that is of importance in optimization and in the analysis of critical points is the fraction of negative eigenvalues of the Hessian, or the *index* of a critical point, α. Prior work (Dauphin et al., 2014; Choromanska et al., 2015) has observed that the index of critical points typically grows rapidly with energy, so that critical points with many descent directions have large loss values. The precise form of this relationship is important for characterizing the geometry of loss surfaces, and in our framework it can be readily computed from the spectral density via integration,

$$\alpha(\epsilon, \phi) \equiv \int_{-\infty}^{0} \rho(\lambda; \epsilon, \phi)\, d\lambda = 1 - \int_{0}^{\infty} \rho(\lambda; \epsilon, \phi)\, d\lambda\,. \tag{19}$$

Given that the spectral density is defined implicitly through equation eqn. (18), performing this integration analytically is nontrivial. We discuss a method for doing so in the supplementary material. The full result is too long to present here, but we find that for small α,

$$\alpha(\epsilon, \phi) \approx \alpha_0(\phi) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2}, \tag{20}$$

where the critical value of ϵ,

$$\epsilon_c = \frac{1}{16}(1 - 20\phi - 8\phi^2 + (1 + 8\phi)^{3/2})\,, \tag{21}$$

is the value of the energy below which all critical points are minimizers. We note that ϵ_c can be found in a simpler way: it is the value of ϵ below which eqn. (18) has only real roots at $z = 0$, i.e. it is the value of ϵ for which the discriminant of the polynomial in eqn. (18) vanishes at $z = 0$. Also, we observe that ϵ_c vanishes cubically as ϕ approaches 1,

$$\epsilon_c \approx \frac{2}{27}(1 - \phi)^3 + \mathcal{O}(1 - \phi)^4\,. \tag{22}$$

The $3/2$ scaling in eqn. (20) is the same behavior that was found in (Bray & Dean, 2007) in the context of the field theory of Gaussian random functions. As we will see later, the $3/2$ scaling persists in a more refined version of this calculation and in numerical simulations.

5. Analysis of Assumptions

5.1. Universality

There is a wealth of literature establishing that many properties of large random matrices do not depend on the details of how their entries are distributed, i.e. many results are *universal*. For instance, Tao & Vu (2012) show that the spectrum of Wishart matrices asymptotically approaches the Marcenko-Pastur law regardless of the distribution of the individual entries, so long as they are independent, have mean zero, unit variance, and finite k-th moment for $k > 2$. Analogous results can be found for Wigner matrices (Aggarwal). Although we are unaware of any existing analy-

[3] Here we ignore degenerate critical points for simplicity.

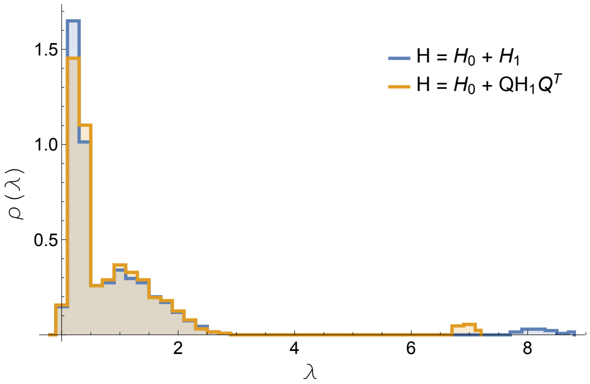


Figure 2. Testing the validity of the free independence assumption by comparing the eigenvalue distribution of $H_0 + H_1$ (in blue) and $H_0 + QH_1Q^T$ (in orange) for randomly generated orthogonal Q. The discrepancies are small, providing support for the assumption. Data is for a trained single-hidden-layer ReLU autoencoding network with 20 hidden units and no biases on 150 4×4 downsampled, grayscaled, whitened CIFAR-10 images.

ses relevant for the specific matrices studied here, we believe that our conclusions are likely to exhibit some universal properties – for example, as in the Wishart and Wigner cases, normality is probably not necessary.

On the other hand, most universality results and the tools we are using from random matrix theory are only exact in the limit of infinite matrix size. Finite-size corrections are actually largest for small matrices, which (counter-intuitively) means that the most conservative setting in which to test our results is for small networks. So we will investigate our assumptions in this regime. We expect agreement with theory only to improve for larger systems.

5.2. Primary assumptions

In sec. 2 we introduced a number of assumptions we dubbed *primary* and we now discuss their validity.

5.2.1. Free independence of H_0 and H_1

Our use of free probability relies on the free independence of H_0 and H_1. Generically we may expect some alignment between the eigenspaces of H_0 and H_1 so that free independence is violated; however, we find that this violation is often quite small in practice. To perform this analysis, it suffices to examine the discrepancy between the distribution of $H_0 + H_1$ and that of $H_0 + QH_1Q^T$, where Q is an orthogonal random matrix of Haar measure (Chen et al., 2016). Fig. 2 show minimal discrepancy for a network trained on autoencoding task; see the supplementary material for further details and additional experiments. More precise methods for quantifying partial freeness exist (Chen et al., 2016), but we leave this analysis for future work.

5.2.2. Residuals are i.i.d. random normal

First, we note that $e_{i\mu} \sim \mathcal{N}(0, 2\epsilon)$ is consistent with the definition of the energy ϵ in eqn. (2). Furthermore, because the gradient of the loss is proportional to the residuals, it vanishes in expectation (i.e. as $m \to \infty$), which specializes our analysis to critical points. So this assumptions seems necessary for our analysis. It is also consistent with the priors leading to the choice of the squared error loss function. Altogether we believe this assumption is fairly mild.

5.2.3. Data are i.i.d. random normal

This assumption is almost never strictly satisfied, but it is approximately enforced by common preprocessing methods, such as whitening and random projections.

5.2.4. Weights are i.i.d. random normal

Although the i.i.d. assumption is clearly violated for a network that has learned any useful information, the weight distributions of trained networks often appear random, and sometimes appear normal (see, e.g., fig. S1 of the supplementary material).

5.3. Secondary assumptions

In sec. 4 we introduced two assumptions we dubbed *secondary*, and we discuss their validity here.

5.3.1. J and H_1 are i.i.d. random normal

Given that J and H_1 are constructed from the residuals, the data, and the weights – variables that we assume are i.i.d. random normal – it is not unreasonable to suppose that their entries satisfy the universality conditions mentioned above. In fact, if this were the case, it would go a long way to validate the approximations made in sec. 4. As it turns out, the situation is more complicated: both J and H_1 have substructure which violates the independence requirement for the universality results to apply. We must understand and take into account this substructure if we wish to improve our approximation further. We examine one way of doing so in the next section.

6. Beyond the Simple Case

In sec. 4 we illustrated our techniques by examining the simple case of Wishart plus Wigner matrices. This analysis shed light on several qualitative properties of the spectrum of the Hessian matrix, but, as discussed in the previous section, some of the assumptions were unrealistic. We believe that it is possible to relax many of these assumptions to produce results more representative of practical networks. In this section, we take one step in that direction and focus on the specific case of a single-hidden-layer ReLU network.

6.1. Product Wishart distribution and H_1

In the single-hidden-layer ReLU case, we can write H_1 as,

$$H_1 = \begin{pmatrix} 0 & H_{12} \\ {H_{12}}^T & 0 \end{pmatrix} ,$$

where its off-diagonal blocks are given by the matrix H_{12},

$$\begin{aligned} [H_{12}]_{ab;cd} &= \frac{1}{m} \sum_{i\mu} e_{i\mu} \frac{\partial}{\partial W^{(2)}_{cd}} \frac{\partial \hat{y}_{i\mu}}{\partial W^{(1)}_{ab}} \\ &= \frac{1}{m} \sum_{\mu=1}^{m} e_{c\mu} \delta_{ad} \theta(z^{(1)}_{a\mu}) x_{b\mu} \, . \end{aligned} \tag{23}$$

Here it is understood that the matrix H_{12} is obtained from the tensor $[H_{12}]_{ab;cd}$ by independently vectorizing the first two and last two dimensions, δ_{ad} is the Kronecker delta function, and θ is the Heaviside theta function. As discussed above, we take the error to be normally distributed,

$$e_{c\mu} = \big(\sum_j W^{(2)}_{cj} [z^{(1)}_{j\mu}]_+ - y_{c\mu}(x) \big) \sim \mathcal{N}(0, 2\epsilon) \, . \tag{24}$$

We are interested in the situation in which the layer width n and the number of examples m are large, in which case $\theta(z^{(1)}_{a\mu})$ can be interpreted as a mask that eliminates half of the terms in the sum over μ. If we reorder the indices so that the surviving ones appear first, we can write,

$$H_{12} \quad \approx \quad \frac{1}{m} \delta_{ad} \sum_{\hat{\mu}=1}^{m/2} e_{c\hat{\mu}} x_{b\hat{\mu}} = \frac{1}{m} I_n \otimes \hat{e} \hat{x}^T \, ,$$

where we have written $\hat{e}$ and $\hat{x}$ to denote the $n \times \frac{m}{2}$ matrices derived from e and x by removing the vanishing half of their columns.

Owing to the block off-diagonal structure of H_1, the squares of its eigenvalues are determined by the eigenvalues of the product of the blocks,

$$H_{12} {H_{12}}^T = \frac{1}{m^2} I_n \otimes (\hat{e}\hat{x}^T)(\hat{e}\hat{x}^T)^T \, . \tag{25}$$

The Kronecker product gives an n-fold degeneracy to each eigenvalue, but the spectral density is the same as that of

$$M \equiv \frac{1}{m^2} (\hat{e}\hat{x}^T)(\hat{e}\hat{x}^T)^T \, . \tag{26}$$

It follows that the spectral density of H_1 is related to that of M,

$$\rho_{H_1}(\lambda) = |\lambda| \rho_M(\lambda^2) \, . \tag{27}$$

Notice that M is a Wishart matrix where each factor is itself a product of real Gaussian random matrices. The spectral density for matrices of this type has been computed using the cavity method (Dupic & Castillo, 2014). As the specific form of the result is not particularly enlightening, we defer its presentation to the supplementary material. The result may also be derived using free probability theory (Muller, 2002; Burda et al., 2010). From either perspective it is possible to derive the the $\mathcal{R}$ transform, which reads,

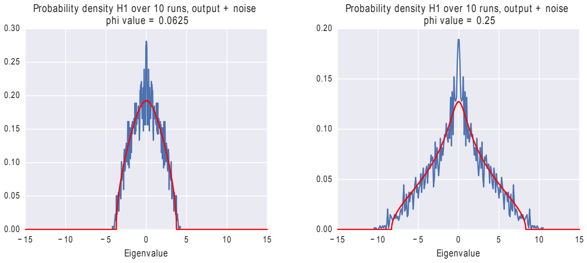


Figure 3. Spectrum of H_1 at initialization, using 4×4 downsampled, grayscaled, whitened CIFAR-10 in a single layer ReLU autoencoder with 16 hidden units. Error signals have been replaced by noise distributed as $\mathcal{N}(0, 100)$, i.e. $\epsilon = 50$. Theoretical prediction (red) matches well. Left $\phi = 1/16$, right $\phi = 1/4$.

$$\mathcal{R}_{H_1}(z) = \frac{\epsilon \phi z}{2 - \epsilon \phi^2 z^2} \, . \tag{28}$$

See fig. 3 for a comparison of our theoretical prediction for the spectral density to numerical simulations[4].

6.2. H_0 and the Wishart assumption

Unlike the situation for H_1, to the best of our knowledge the limiting spectral density of H_0 cannot easily be deduced from known results in the literature. In principle it can be computed using diagrammatics and the method of moments (Feinberg & Zee, 1997; Tao, 2012), but this approach is complicated by serious technical difficulties – for example, the matrix has both block structure and Kronecker structure, and there are nonlinear functions applied to the elements. Nevertheless, we may make some progress by focusing our attention on the smallest eigenvalues of H_0, which we expect to be the most relevant for computing the smallest eigenvalues of H.

The empirical spectral density of H_0 for an autoencoding network is shown in fig. 4. At first glance, this distribution does not closely resemble the Marchenko-Pastur distribution (see, e.g., the $\epsilon = 0$ curve in fig. 1) owing to its heavy tail and small eigenvalue gap. On the other hand, we are not concerned with the large eigenvalues, and even though the gap is small, its scaling with ϕ appears to be well-approximated by λ_- from eqn. (8)

[4] As the derivation requires that the error signals are random, in the simulations we manually overwrite the network's error signals with random noise.

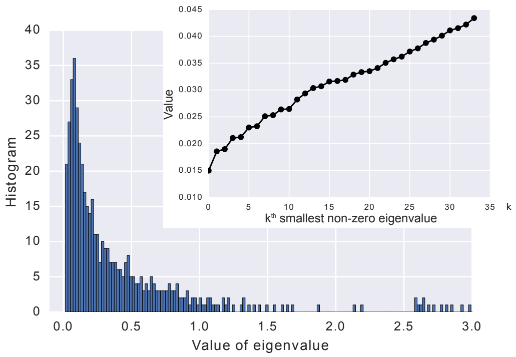


Figure 4. The spectrum of H_0 at initialization of a single layer ReLU autoencoder with 16 hidden units and 256 4×4 downsampled, grayscaled, whitened CIFAR-10 images. There are 16 null directions, and the corresponding zero eigenvalues have been removed. The inset shows the values of the smallest 35 nonzero eigenvalues. The positive value of the first datapoint reflects the existence of a nonzero gap.

for appropriate σ. See fig. 5. This observation suggests that as a first approximation, it is sensible to continue to represent the limiting spectral distribution of H_0 with the Marchenko-Pastur distribution.

6.3. Improved approximation to the Hessian

The above analysis motivates us to propose an improved approximation to $\mathcal{R}_H(z)$,

$$\mathcal{R}_H(z) = \frac{\sigma}{1 - \sigma z \phi} + \frac{\epsilon \phi z}{2 - \epsilon \phi^2 z^2}\,, \tag{29}$$

where the first term is the $\mathcal{R}$ transform of the Marchenko-Pastur distribution with the σ parameter restored. As before, an algebraic equation defining the Stieltjes transform of ρ_H can be deduced from this expression,

$$\begin{aligned} 2 = 2\big(z - \sigma(1-\phi)\big)G - \phi\big(2\sigma z + \epsilon(1-\phi)\big)G_H^2 \\ - \epsilon\phi^2\big(z + \sigma(\phi - 2)\big)G_H^3 + z\epsilon\sigma\phi^3 G_H^4\,. \end{aligned} \tag{30}$$

Using the same techniques as in sec. 4, we can use this equation to obtain the function $\alpha(\epsilon, \phi)$. We again find the same $^3/_2$ scaling, i.e.,

$$\alpha(\epsilon, \phi) \approx \tilde{\alpha}_0(\phi) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2}\,. \tag{31}$$

Compared with eqn. (20), the coefficient function $\tilde{\alpha}_0(\phi)$ differs from $\alpha_0(\phi)$ and,

$$\epsilon_c = \frac{\sigma^2\big(27 - 18\chi - \chi^2 + 8\chi^{3/2}\big)}{32\phi(1-\phi)^3}\,, \quad \chi = 1 + 16\phi - 8\phi^2\,. \tag{32}$$

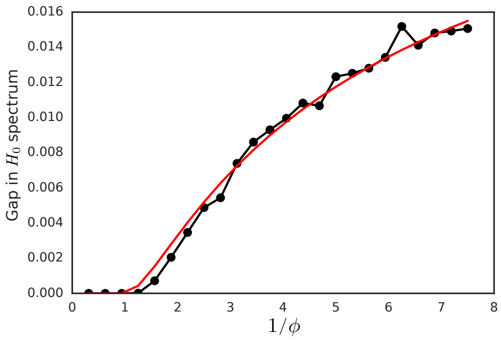


Figure 5. Evolution of the H_0 spectral gap as a function of $1/\phi$, comparing the Marcenko-Pastur (red) against empirical results for a 16-hidden unit single-layer ReLU autoencoder at initialization (black dots, each averaged over 30 independent runs). Dataset was taken from 4×4 downsampled, grayscaled, whitened CIFAR-10 images. A single fit parameter, characterizing the variance of the matrix elements in the Wishart ensemble, is used.

Despite the apparent pole at $\phi = 1$, ϵ_c actually vanishes there,

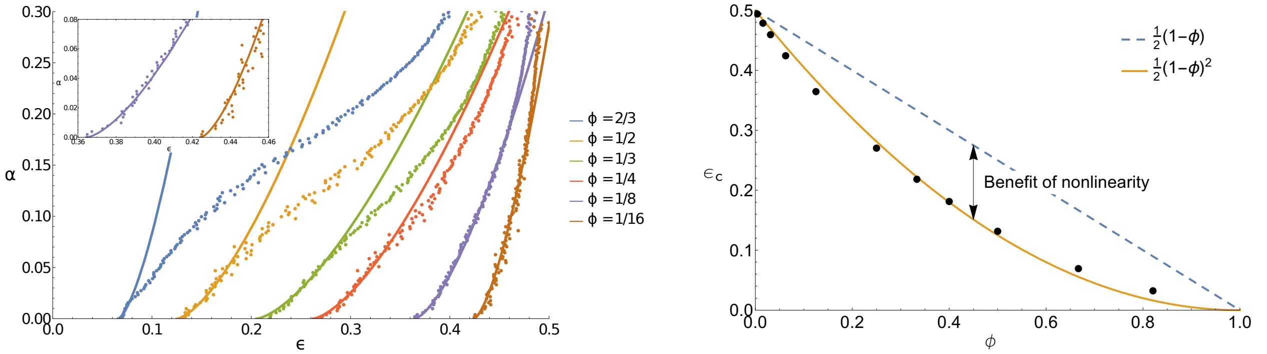
$$\epsilon_c \approx \frac{8}{27}\sigma^2(1-\phi)^3 + \mathcal{O}(1-\phi)^4\,. \tag{33}$$

Curiously, for $\sigma = 1/2$, this is precisely the same behavior we found for the behavior of ϵ_c near 1 in the Wishart plus Wigner approximation in sec. 4. This observation, combined with the fact that the $^3/_2$ scaling in eqn. (31) is also what we found in sec. 4, suggest that it is H_0, rather than H_1, that is governing the behavior near ϵ_c.

6.4. Empirical distribution of critical points

We conduct large-scale experiments to examine the distribution of critical points and compare with our theoretical predictions. Uniformly sampling critical points of varying energy is a difficult problem. Instead, we take more of a brute force approach: for each possible value of the index, we aim to collect many measurements of the energy and compute the mean value. Because we cannot control the index of the obtained critical point, we run a very large number of experiments (~50k) in order to obtain sufficient data for each value of α. This procedure appears to be a fairly robust method for inferring the $\alpha(\epsilon)$ curve.

We adopt the following heuristic for finding critical points. First we optimize the network with standard gradient descent until the loss reaches a random value between 0 and the initial loss. From that point on, we switch to minimizing a new objective, $J_g = |\nabla_\theta \mathcal{L}|^2$, which, unlike the primary objective, is attracted to saddle points. Gradient descent on J_g only requires the computation of Hessian-vector products and can be executed efficiently.



(a) Index of critical points versus energy

(b) Energy of minimizers versus parameters/data points

Figure 6. Empirical observations of the distribution of critical points in single-hidden-layer $\tanh$ networks with varying ratios of parameters to data points, ϕ. (a) Each point represents the mean energy of critical points with index α, averaged over ~200 training runs. Solid lines are best fit curves for small $\alpha \approx \alpha_0|\epsilon - \epsilon_c|^{3/2}$. The good agreement (emphasized in the inset, which shows the behavior for small α) provides support for our theoretical prediction of the $^3\!/_2$ scaling. (b) The best fit value of ϵ_c from (a) versus ϕ. A surprisingly good fit is obtained with $\epsilon_c = \frac{1}{2}(1-\phi)^2$. Linear networks obey $\epsilon_c = \frac{1}{2}(1-\phi)$. The difference between the curves shows the benefit obtained from using a nonlinear activation function.

We discard any run for which the final $J_g > 10^{-6}$; otherwise we record the final energy and index.

We consider relatively small networks and datasets in order to run a large number of experiments. We train single-hidden-layer $\tanh$ networks of size $n = 16$, which also equals the input and output dimensionality. For each training run, the data and targets are randomly sampled from standard normal distributions, which makes this a kind of memorization task. The results are summarized in fig. 6. We observe that for small α, the scaling $\alpha \approx |\epsilon - \epsilon_c|^{3/2}$ is a good approximation, especially for smaller ϕ. This agreement with our theoretical predictions provides support for our analytical framework and for the validity of our assumptions.

As a byproduct of our experiments, we observe that the energy of minimizers is well described by a simple function, $\epsilon_c = \frac{1}{2}(1-\phi)^2$. Curiously, a similar functional form was derived for linear networks (Advani & Ganguli, 2016), $\epsilon_c = \frac{1}{2}(1-\phi)$. In both cases, the value at $\phi = 0$ and $\phi = 1$ is understood simply: at $\phi = 0$, the network has zero effective capacity and the variance of the target distribution determines the loss; at $\phi = 1$, the matrix of hidden units is no longer rank constrained and can store the entire input-output map. For intermediate values of ϕ, the fact that the exponent of $(1-\phi)$ is larger for $\tanh$ networks than for linear networks is the mathematical manifestation of the nonlinear network's better performance for the same number of parameters. Inspired by these observations and by the analysis of Zhang et al. (2016), we speculate that this result may have a simple information-theoretic explanation, but we leave a quantitative analysis to future work.

7. Conclusions

We introduced a new analytical framework for studying the Hessian matrix of neural networks based on free probability and random matrix theory. By decomposing the Hessian into two pieces $H = H_0 + H_1$ one can systematically study the behavior of the spectrum and associated quantities as a function of the energy ϵ of a critical point. The approximations invoked are on H_0 and H_1 separately, which enables the analysis to move beyond the simple representation of the Hessian as a member of the Gaussian Orthogonal Ensemble of random matrices.

We derived explicit predictions for the spectrum and the index under a set of simplifying assumptions. We found empirical evidence in support of our prediction that that small $\alpha \approx |\epsilon - \epsilon_c|^{3/2}$, raising the question of how universal the $^3\!/_2$ scaling may be, especially given the results of (Bray & Dean, 2007). We also showed how some of our assumptions can be relaxed at the expense of reduced generality of network architecture and increased technical calculations. An interesting result of our numerical simulations of a memorization task is that the energy of minimizers appears to be well-approximated by a simple function of ϕ. We leave the explanation of this observation as an open problem for future work.

Acknowledgements

We thank Dylan J. Foster, Justin Gilmer, Daniel Holtmann-Rice, Ben Poole, Maithra Raghu, Samuel S. Schoenholz, Jascha Sohl-Dickstein, and Felix X. Yu for useful comments and discussions.

References

Advani, Madhu and Ganguli, Surya. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

Aggarwal, Amol. Bulk universality for generalized wigner matrices with few moments. *arXiv preprint arXiv:1612.00421*.

Bray, Alan J. and Dean, David S. Statistics of critical points of gaussian fields on large-dimensional spaces. *Phys. Rev. Lett.*, 98:150201, Apr 2007. doi: 10.1103/PhysRevLett.98.150201. URL http://link.aps.org/doi/10.1103/PhysRevLett.98.150201.

Burda, Z, Jarosz, A, Livan, G, Nowak, MA, and Swiech, A. Eigenvalues and singular values of products of rectangular gaussian random matrices. *Physical Review E*, 82(6):061114, 2010.

Chen, Jiahao, Van Voorhis, Troy, and Edelman, Alan. Partial freeness of random matrices. *arXiv preprint arXiv:1204.2257*, 2016.

Choromanska, Anna, Henaff, Mikael, Mathieu, Michaël, Arous, Gérard Ben, and LeCun, Yann. The loss surface of multilayer networks. *JMLR*, 38, 2015.

Dauphin, Yann N, Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, Ganguli, Surya, and Bengio, Yoshua. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems 27*, pp. 2933–2941, 2014.

Dupic, Thomas and Castillo, Isaac Pérez. Spectral density of products of wishart dilute random matrices. part i: the dense case. *arXiv preprint arXiv:1401.7802*, 2014.

Feinberg, Joshua and Zee, Anthony. Renormalizing rectangles and other topics in random matrix theory. *Journal of statistical physics*, 87(3-4):473–504, 1997.

Freeman, C. Daniel and Bruna, Joan. Topology and geometry of half-rectified network optimization. 2016. URL http://arxiv.org/abs/1611.01540.

Goodfellow, Ian J., Vinyals, Oriol, and Saxe, Andrew M. Qualitatively characterizing neural network optimization problems. *ICLR 2015*. URL http://arxiv.org/abs/1412.6544.

Kawaguchi, Kenji. Deep learning without poor local minima. *Advances in Neural Information Processing Systems 29*, pp. 586–594, 2016.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, and Tang, Ping Tak Peter. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR 2017*. URL http://arxiv.org/abs/1609.04836.

Laisant, C-A. Intégration des fonctions inverses. *Nouvelles annales de mathématiques, journal des candidats aux écoles polytechnique et normale*, 5:253–257, 1905.

Marčenko, Vladimir A and Pastur, Leonid Andreevich. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Muller, Ralf R. On the asymptotic eigenvalue distribution of concatenated vector-valued fading channels. *IEEE Transactions on Information Theory*, 48(7):2086–2091, 2002.

Neyshabur, Behnam, Salakhutdinov, Ruslan, and Srebro, Nathan. Path-sgd: Path-normalized optimization in deep neural networks. pp. 2413–2421, 2015.

Safran, Itay and Shamir, Ohad. On the quality of the initial basin in overspecified neural networks. *JMLR*, 48, 2016. URL http://arxiv.org/abs/1511.04210.

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

Speicher, Roland. Free probability theory. *arXiv preprint arXiv:0911.0087*, 2009.

Tao, T. and Vu, V. Random covariance matrices: universality of local statistics of eigenvalues. *Annals of Probability*, 40(3):1285–1315, 2012.

Tao, Terence. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

Wigner, Eugene P. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, pp. 548–564, 1955.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.