Use of Random Matrix Theory on Understanding Deep Learning

20170058 Keonwoo Kim

July 3, 2020

Abstract

Deep learning using neural networks is ubiquitous in these days, but their analyses are inadequate due to the big and complicated, high-dimensional, highly non-convex nature of the neural networks. This report surveys some of the studies of neural networks using random matrix theory, by approximating a complicated neural network as a nice random matrix system. Naively, a multilayer fully connected feed forward deep neural network can be seen as an approximate spherical spin-glass model. With less conditions, the spectral densities of the Hessian, the Gram matrix, and the Fisher information matrix of a single-hidden-layer network can be obtained. These tell us about some geometrical properties of the loss surface, which are necessary to optimize the loss function.

1 Introduction

Nowadays, machine learning researches are rapidly growing. Especially, the deep learning is one of the hottest research area in not just computer science, but a bunch of other studies. Although the machine learning with deep neural networks have certainly succeeded until now, there are few results on the mathematical understanding of why and which situation they work well. This is mainly because most of the networks are pretty big and complicated, their underlying space is very high-dimensional, and the the loss is highly non-convex for practical neural networks. Furthermore, the data and the weights usually seem to be quite *random* at first glance.

From such a point of view, some researchers started to bring the random matrix theory into the studies of complicated neural networks. In this report, I'll introduce some researches which try to explain (simplified) neural networks using random matrices.

A common strategy used in the studies is regarding unknown entries as being random and/or Gaussian. This is because for easy calculations, some sort of randomness and independence are necessary. Also, many of the basic results made in random matrix theory contains Gaussian condition: such as GOE, a Wishart ensemble, and so on.

With these observations, we may explain why certain phenomena occur during the training process. We may also find an (approximately) efficient choices of parameters or functions (e.g., activation functions), which are strongly related to the convergence of the optimization problem. Moreover, one can guess why certain tricks used to make the convergence faster works, e.g., the batch normalization.

2 Preliminaries

2.1 Random Matrix Theory

Let me recap the concepts discussed in the lectures. Let $M \in \mathbb{R}^{N \times N}$ be a random matrix. Then, the empirical spectral measure of M is defined by

$$\rho_M \coloneqq \frac{1}{N} \sum_{j=1}^{N} \delta_{\lambda_j}$$

where $\lambda_1, \dots, \lambda_N$ are the eigenvalues of M. When M is a GOE, the limiting spectral distribution is the Wigner semicircle distribution:

$$\rho_M \rightharpoonup \mu_{\mathrm{sc}}, \qquad d\mu_{\mathrm{sc}}(x) = \frac{1}{2\pi} \sqrt{(4 - x^2)_+}\, dx, \qquad m_{\mu_{sc}}(z) = \frac{-z + \sqrt{z^2 - 4}}{2}.$$

When M is a Wishart ensemble, that is,

$$X = (X_{ij})_{\substack{1 \le i \le P \\ 1 \le j \le N}}, \qquad \mathbb{E}[X_{ij}] = 0, \quad \mathbb{E}[X_{ij}^2] = 1, \qquad M = \frac{1}{N} X X^T$$

with $P/N \to \alpha \in [0, 1]$, the Marchenko–Pastur law tells us

$$\rho_M \rightharpoonup \mu_{MP(\alpha)}, \qquad d\mu_{MP(\alpha)}(x) = \frac{1}{2\pi \alpha x} \sqrt{[(a_+ - x)(x - a_-)]_+}\, dx, \qquad a_\pm = (1 \pm \sqrt{\alpha})^2.$$

Equivalently,

$$z \alpha m_{\mu_{MP(\alpha)}}(z)^2 + (z - 1 + \alpha) m_{\mu_{MP(\alpha)}}(z) + 1 = 0.$$

2.2 Free Probability Theory

Suppose A and B are two random matrices. Even if A and B are independent, we may not know the spectrum of $A + B$ with ones of A and B. Here the notion of *free independence* comes in: A and B are freely independent if for any k,

$$\mathbb{E}[f_1(A) g_1(B) \cdots f_k(A) g_k(B)] = 0$$

whenever f_i and g_i satisfies $\mathbb{E} f_i(A) = 0$ and $\mathbb{E} g_i(B) = 0$. When $k = 1$, this notion agrees with the classical independence.

Given the Stieltjes transform $m(z)$ of a measure ρ, the $\mathcal{R}$*-transform* is defined as follows: $\mathcal{R}(z)$ satisfying

$$\mathcal{R}(m(z)) + \frac{1}{m(z)} = z.$$

When A and B are freely independent, we have $\mathcal{R}_{A+B} = \mathcal{R}_A + \mathcal{R}_B$ [Spe09].

Using the free independence and the $\mathcal{R}$-transform, we can compute the spectral density of $A + B$ from one of A and B:

1. compute the Stieltjes transform and the $\mathcal{R}$-transform of A and B;
2. obtain $\mathcal{R}_{A+B} = \mathcal{R}_A + \mathcal{R}_B$;
3. compute the Stieltjes transform of $A + B$ from the $\mathcal{R}$-transform of it;
4. invert the Stieltjes transform in order to get the spectral density of $A + B$.

3 Naïve Approximation

3.1 A Connection of a Multilayer Network and the Spherical Spin-Glass Model [Cho+14]

Consider the following fully-connected feed-forward deep neural network $\mathcal{N}$ of depth H for a binary classification, that is, the output Y for the input X is as follows:

$$Y = \sigma(W^{(H)}\sigma(W^{(H-1)}\cdots\sigma(W^{(1)}X)\cdots))$$

where $W^{(i)}$ is a $n_i \times n_{i-1}$ matrix with $n_0 =: d$ and $n_H = 1$, and σ is the ReLU function (applied componentwisely): $\sigma(x) = \max(0, x)$. This can be written as the sum of $X_i A_{ij} \prod_{k=1}^{H} w_{ij}^{(k)}$, where $i = 1, \ldots, n_0$, j runs over the set of all paths from a given network input to its output, and A_{ij} represents whether the path (i, j) is activated or not (by the ReLU function), and $w_{ij}^{(k)}$ represents the contribution of $W^{(k)}$ to the path (i, j).

To make the problem simple, they made lots of simplifications. First, replace X_i in the above formula for Y by $X_{ij} \sim N(0, 1)$ iid, and assume that A_{ij} to be Bernoulli with the common probability of success ρ, and all A's are independent of X's. Then, taking expectation on A variables, $\mathbb{E}_A[Y]$ becomes

$$\mathbb{E}_A[Y] = \sum_{i=1}^{n_0}\sum_{j=1}^{\gamma} X_{ij}\rho \prod_{k=1}^{H} w_{ij}^{(k)}$$

where $\gamma = n_1 \cdots n_H$ is the number of total paths described above. In practice, the number of network parameters $N = \sum_{i=1}^{H} n_{i-1}n_i$ can be reduced to $s \ll N$, with at most ϵ loss in accuracy, by the *redundancy assumption.* Also, the *uniformity assumption* requires the weight parameters in the reduced network to be almost evenly distributed: for a sequence $(w_{i_1}, \ldots, w_{i_H})$ of weight parameters of length H, letting $t_{i_1,\ldots,i_H}$ be the number of paths having the same weight configuration with $(w_{i_1}, \ldots, w_{i_H})$, the uniformity assumption implies there exists a positive constant $c \geq 1$ such that

$$\frac{1}{c} \cdot \frac{\Psi}{s^H} \leq t_{i_1,\ldots,i_H} \leq c \cdot \frac{\Psi}{s^H}$$

where $\Psi = n_0\gamma$ is the total number of paths. By the redundancy assumption, $\mathbb{E}_A[Y]$ can be approximated by

$$Y_s := \sum_{i_1,\ldots,i_H=1}^{s} \sum_{j=1}^{t_{i_1,\ldots,i_H}} X_{i_1,\ldots,i_H}^{(j)} \rho \prod_{k=1}^{H} w_{i_k};$$

($X_{i_1,\ldots,i_H}^{(j)}$ is the j-th variable yielding such a weight configuration) and by the uniformity assumption, this can be again approximated by

$$\hat{Y}_s := \sum_{i_1,\ldots,i_H=1}^{s} \sum_{j=1}^{\Psi/s^H} X_{i_1,\ldots,i_H}^{(j)} \rho \prod_{k=1}^{H} w_{i_k};$$

and this can be further approximated by ($\Lambda := \Psi^{1/H}$)

$$\hat{Y} = \hat{Y}_s|_{s=\Lambda} := \sum_{i_1,\ldots,i_H=1}^{\Lambda} X_{i_1,\ldots,i_H} \rho \prod_{k=1}^{H} w_{i_k}.$$

Consider the following absolute loss:

$$\mathcal{L}_{\Lambda,H}(w) = \mathbb{E}_A\left[|Y_{\text{true}} - Y|\right] \approx S - \text{sgn}(Y_{\text{true}})\hat{Y} = S + \sum_{i_1,\ldots,i_H=1}^{\Lambda} X_{i_1,\ldots,i_H}\rho \prod_{k=1}^{H} w_{i_k}$$

after rewriting $X_{\ldots} \overset{d}{=} -X_{\ldots}$ if necessary, where $S = \sup_w \hat{Y}$, $Y_{\text{true}} = \pm S$, and Y_{true} is the true data labeling. With the final assumption that $\frac{1}{\Lambda}\sum_{i=1}^{\Lambda} w_i^2 = C$ for some constant $C > 0$, we have

$$\mathcal{L}_{\Lambda,H}(w) \approx S - \text{sgn}(Y_{\text{true}})\hat{Y} = S + \sum_{i_1,\ldots,i_H=1}^{\Lambda} X_{i_1,\ldots,i_H}\rho \prod_{k=1}^{H} w_{i_k}$$

so that it becomes the Hamiltonian of the H-spin spherical spin-glass model after controlling constants as they do not affect the loss minimization procedure:

$$\mathcal{L}_{\Lambda,H}(w) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1,\ldots,i_H=1}^{\Lambda} X_{i_1,\ldots,i_H} \prod_{k=1}^{H} w_{i_k}, \qquad \frac{1}{\Lambda}\sum_{i=1}^{\Lambda} w_i^2 = 1.$$

This result connects a deep FC feed-forward network to the spin-glass model, but it assumes a bunch of unrealistic assumption.

4 Single-Hidden-Layer Case

Let's investigate the neural networks more carefully. I'll focus on the single-hidden-layer neural network case with the ReLU as the activation function in this section.

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 neural network, and consider the network $\hat{y} = W^{(2)}\sigma(W^{(1)}x)$ where σ is the activation function. I'll consider the mean squared error here, that is,

$$\mathcal{L} = \frac{1}{2m}\sum_{i=1}^{n_2}\sum_{\mu=1}^{m} e_{i\mu}^2, \qquad e_{i\mu} = \hat{y}_{i\mu} - y_{i\mu}$$

where μ is the index for the data examples and m is the total number of the examples. For $\alpha, \beta \in \{1, 2\}$, the hessian of the loss $\mathcal{L}$ is defined as follows:

$$H_{\alpha\beta} = \frac{\partial^2 \mathcal{L}}{\partial W^{(\alpha)} \partial W^{(\beta)}} = \frac{1}{m}\sum_{i=1}^{n_2}\sum_{\mu=1}^{m} \frac{\partial \hat{y}_{i\mu}}{\partial W^{(\alpha)}}\frac{\partial \hat{y}_{i\mu}}{\partial W^{(\beta)}} + \frac{1}{m}\sum_{i=1}^{n_2}\sum_{\mu=1}^{m} e_{i\mu}\frac{\partial^2 \hat{y}_{i\mu}}{\partial W^{(\alpha)}\partial W^{(\beta)}} =: [H_0]_{\alpha\beta} + [H_1]_{\alpha\beta}$$

where the first term of the RHS can be written as:

$$[H_0]_{\alpha\beta} \coloneqq \frac{1}{m}\sum_{i=1}^{n_2}\sum_{\mu=1}^{m} \frac{\partial \hat{y}_{i\mu}}{\partial W^{(\alpha)}}\frac{\partial \hat{y}_{i\mu}}{\partial W^{(\beta)}} = \left[\frac{1}{m}JJ^T\right]_{\alpha\beta}, \qquad J_\alpha = \frac{\partial \mathcal{L}}{\partial W^{(\alpha)}}.$$

Three subsections below deal with the Hessian matrix, the Gram matrix, and the Fisher information matrix, respectively. The Hessian matrix is obviously important, since it determines the local nature of the critical points of the loss surface. And the Gram matrix is basically the sample covariance matrix of the output. So, as we stack the layers up, the distribution of the eigenvalues of the sample covariance, i.e., the singular values of the output (of the ℓ-th layer), detemine the extent to which the input become skewed while propagating the network. Since highly distorted network implies being poor conditioned, it should be avoided. By inverstigating the Gram matrix, we hope that we can resolve those situations. Finally, the Fisher information matrix is another fundamental quantity for understanding the neural network, because it describes a local metric of the loss surface concerning the KL-divergence:

Fisher information matrix is the Hessian of the KL-divergence between two distributions $p(x|\theta)$ and $p(x|\theta')$, w.r.t. θ', evaluated at $\theta' = \theta$.

The proposition above can be shown easily. Since it is a Hessian matrix of the KL-divergence, it can be used to decide the direction of the gradient descent process used in the optimization. Therefore, the spectrum of the Fisher information matrix is also a central quantity we need to observe.

4.1 The Spectrum of the Hessian (ReLU) [PB17]

For the sake of simplicity, assume $\sigma = \text{ReLU}$ and $n_0 = n_1 = n_2 =: n$. The following are the primary assumptions that the authors of the paper made to start the discussion:

1. H_0 and H_1 are freely independent. The authors showed an empirical evidence to this, using the discrepancy between the moments of $H_0 + H_1$ from $H_0 + OH_1O^T$ (O is an orthogonal random matrix of Haar measure) to check the freeness.

2. $e_{i\mu} \sim N(0, 1)$, i.i.d.

3. X_i's are i.i.d. Gaussian. With some common preprocesses, such as whitening and random projections, this assumption is approximately satisfied.

4. $W_{ij}^{(\alpha)}$'s are i.i.d. Gaussian without biases. Empirically, the weights often appear random, and sometimes appear Gaussian.

Note that $[H_1]_{\alpha\alpha} = \frac{1}{m}\sum \frac{\partial \hat{y}_{i\mu}}{(\partial W^{(\alpha)})^2} = 0$ since $\hat{y}_{i\mu}$ is piecewisely linear for $W^{(\alpha)}$. Thus, H_1 can be written as the following block diagonal matrix:

$$H_1 = \begin{pmatrix} 0 & \tilde{H}_1 \\ \tilde{H}_1^T & 0 \end{pmatrix}, \qquad [\tilde{H}_1]_{ab;cd} = \frac{1}{m}\sum_{i,\mu} e_{i\mu} \frac{\partial^2 \hat{y}_{i\mu}}{\partial W_{cd}^{(2)} \partial W_{ab}^{(1)}} = \frac{1}{m}\sum_{\mu=1}^{m} e_{c\mu}\delta_{ad}\theta(z_{a\mu})x_{b\mu}$$

seeing $\tilde{H}_1$ as an $n^2 \times n^2$ matrix where index is obtained from vectorizing the first two and the last two dimensions. Here θ means the Heaviside theta function and $z = W^{(1)}x$.

As n and m grow, $\theta(z_{a\mu})$ can be interpreted as a mask that eliminates half of the terms in the sum over μ. Rearranging the indices, we can write

$$[\tilde{H}_1]_{ab;cd} \approx \frac{1}{m}\delta_{ad}\sum_{\mu=1}^{m/2} e_{c\mu}x_{b\mu} =: \left[\frac{1}{m} I_n \otimes \hat{e}\hat{x}^T\right]_{ad;cb},$$

hence

$$\tilde{H}_1\tilde{H}_1^T \approx \frac{1}{m^2} I_n \otimes (\hat{e}\hat{x}^T)(\hat{e}\hat{x}^T)^T.$$

Thus, $\tilde{H}_1\tilde{H}_1^T$ has the same limiting spectral density with $M = \frac{1}{m^2}(\hat{e}\hat{x}^T)(\hat{e}\hat{x}^T)^T$. Therefore, the limiting spectral density of H_1 is:

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

where M is a Wishart matrix. From 4.2 of [DC14], we have

$$\rho_{H_1}(\lambda) = \left(1 - \min\left(1, \frac{\alpha}{2}\right)\right)\delta(\lambda) + \frac{\alpha^2|\lambda|}{2\epsilon}\rho_c\left(\frac{\alpha^2\lambda^2}{2\epsilon}, \frac{\alpha}{2}\right)$$

where $\alpha = m/n$ and

$$\rho_c(x, \alpha) = \frac{\sqrt{3}}{6\pi x \sqrt[3]{2}}(r_+ r_-)\mathbf{1}_{[\theta(1-\alpha)\,x_-, x_+]}(x), \quad r_\pm = \sqrt[3]{9(2+\alpha)(x-\xi_0) \pm 6\sqrt{3(x-x_-)x(x_+-x)}},$$

$$x_\pm = \frac{8 + 20\alpha - \alpha^2 \pm \sqrt{\alpha}(8+\alpha)^{3/2}}{8}, \qquad \xi_0 = -\frac{2(\alpha-1)^3}{9(2+\alpha)}.$$

This yields the following $\mathcal{R}$-transform:

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

On the other hand, computing the spectral density of H_0 is difficult, so they propose another assumption here: the Jacobian matrix, $J_\alpha = \partial\mathcal{L}/\partial W^{(\alpha)}$, consist of i.i.d. normal random variable. Then, H_0 becomes a Wishart matrix so that the limiting spectral density follows the Marchenko–Pastur law. In conclusion, for some ς,

$$\mathcal{R}_{H_0}(z) = \frac{\varsigma}{1 - \varsigma\phi z}, \qquad \therefore \mathcal{R}_{H_0+H_1}(z) = \frac{\varsigma}{1 - \varsigma\phi z} + \frac{\epsilon\phi z}{2 - \epsilon\phi^2 z^2}.$$

Using this result, we can restore the spectral density of $H = H_0 + H_1$, as desired. The Stieltjes transform of H satisfies the following polynomial equation:

$$2 = 2(z - \varsigma(1-\phi))m_H(z) - \phi(2\varsigma z + \epsilon(1-\phi))m_H(z)^2 - \epsilon\phi^2(z + \varsigma(\phi-2))m_H(z)^3 + z\epsilon\varsigma\phi^3 m_H(z)^4.$$

Note that the spectral density of the Hessian is related to the descent directions of the loss surface. Therefore, one quantity that is of significant importance in optimization is the fraction of negative eigenvalues of the Hessian, a.k.a. the *index* of a critical point:

$$\alpha(\epsilon, \phi) \coloneqq \int_{-\infty}^{p} \rho(\lambda; \epsilon, \phi)\, d\lambda = 1 - \int_0^\infty \rho(\lambda; \epsilon, \phi)\, d\lambda.$$

Letting $f(w) \coloneqq \mathcal{R}_H(w) + 1/w$, then $m(z) = f^{-1}(z)$ and since

$$\int f^{-1}(z)\, dz = z f^{-1}(z) - F \circ f^{-1}(z) + C$$

for an antiderivative F of f, using

$$\rho(\lambda) = -\frac{1}{\pi}\lim_{\epsilon\downarrow 0} \operatorname{Im} m(\lambda + i\epsilon),$$

we may get the formula for $\alpha(\epsilon, \phi)$. In particular, for small α,

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

for some critical $\epsilon_c \approx \frac{8}{27}\varsigma^2(1-\phi)^3 + O((1-\phi)^4)$.

4.2 The Spectrum of the Gram Matrix [PW17]

In this paper, assume $W^{(2)} = I_{n_1}$; i.e., consider the single-layered network

$$Y = \sigma(WX).$$

(In fact, for the Gram matrix we have $M = W^{(2)}(\frac{1}{m}\sigma(WX)\sigma(WX)^T)(W^{(2)})^T = W^{(2)}\tilde{M}(W^{(2)})^T$ for $M = \frac{1}{m}\sigma(WX)\sigma(WX)^T$. Thus, we may assume $W^{(2)} = I$ by pre-/post-multiplying proper inverses of some matrices.) Further, suppose the Gaussian assumptions: $X \in \mathbb{R}^{n_0 \times m}$, $X_{i\mu} \sim N(0, \sigma_x^2)$ and $W \in \mathbb{R}^{n_1 \times n_0}$ with $W_{ij} \sim N(0, \sigma_w^2/n_0)$. Let σ be an arbitrary activation function, with zero mean and finite moments:

$$\int \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \sigma(\sigma_w \sigma_x z)\, dz = 0, \qquad \left| \int \frac{1}{\sqrt{2\pi}} e^{-z^2/2} [\sigma(\sigma_w \sigma_x z)]^k\, dz \right| < \infty \quad \text{for} \quad k > 1.$$

In this subsection, we will consider the following Gram matrix: $M = \frac{1}{m} YY^T \in \mathbb{R}^{n_1 \times n_1}$.

The Stieltjes transform of $M = \frac{1}{m} YY^T$ can be obtained from the moments method:

$$m(z) = \frac{1}{n_1} \sum_{k=0}^{\infty} \frac{\mathbb{E}[\operatorname{tr} M^k]}{z^{k+1}} = \frac{1}{n_1 m^k} \sum_{k=0}^{\infty} \frac{1}{z^{k+1}} \mathbb{E}\left[\sum_{\substack{1 \le i_1, \ldots, i_k \le n_1 \\ 1 \le \mu_1, \ldots, \mu_k \le m}} Y_{i_1\mu_1} Y_{i_2\mu_1} Y_{i_2\mu_2} Y_{i_3\mu_2} \cdots Y_{i_k\mu_k} Y_{i_1\mu_k} \right].$$

Basically, the strategy is the same when we calculated the Stieltjes transform of the limiting distribution of the GOE: make a graph, whose vertices are coloured with two colors (for i's and for μ's), which represents each term of the sum above, and count the multiplicities of patterns. The authors divided the calculation of moments into two subproblems: one of enumerating certain connected outer-planar graphs, and another of evaluating integrals that correspond to cycles in those graphs. Carefully counting the moments, the authors obtained the following:

$$m(z) = \frac{\psi}{z} P\left(\frac{1}{z\psi}\right) + \frac{1-\psi}{z},$$

where η, ζ, P are defined as follows: (P is defined as a solution of a quartic polynomial equation)

$$\eta = \int \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sigma(\sigma_w \sigma_x z)^2\, dz, \qquad \zeta = \left[\sigma_w \sigma_x \int \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sigma'(\sigma_w \sigma_x z)^2\, dz \right]^2,$$

$$P(t) = 1 + (\eta - \zeta) t P_\phi(t) P_\psi(t) + \frac{P_\phi(t) P_\psi(t) t\zeta}{1 - P_\phi(t) P_\psi(t) t\zeta}, \quad P_\phi = 1 + (P-1)\phi, \quad P_\psi = 1 + (P-1)\psi.$$

In particular, when $\eta = \zeta$, which is equivalent to that σ is linear, we have $M = \frac{1}{m}(\tilde{W}X)(\tilde{W}X)^T$ so that M is a Wishart matrix, following the Marchenko–Pastur law. Note that the quartic equation becomes degerated to a cubic equation when $\eta = \zeta$, whose solution is exactly the Stieltjes transform of the Marchenko–Pastur distribution, as we already know.

4.3 The Spectrum of the Fisher Information [PW18]

In this paper, the authors concentrates on the Fisher Information matrix:

$$H_0 = \mathbb{E}_X[J^T J], \quad J_{i\alpha} = \frac{\partial \hat{y}_i}{\partial \theta_\alpha}.$$

As in §4.2, the authors proceed by the moments method.

To simplify the calculation of $\mathrm{tr}[H_0^d]$, the authors impose the following simple proposition:

$$A = [A_1 \quad A_2] \implies \mathrm{tr}((A^T A)^d) = \sum_{b \in \{1,2\}^d} \mathrm{tr} \prod_{i=1}^{d} A_{b_i} A_{b_i}^T = \sum_{b \in \{1,2\}^d} \mathrm{tr} \prod_{i=1}^{d} A_{b_{i-1}}^T A_{b_i}. \quad (b_0 \coloneqq b_d)$$

Dividing J as $[J^{(1)} \quad J^{(2)}]$,

$$\mathrm{tr}[H_0^d] = \sum_{b \in \{1,2\}^d} \mathrm{tr} \prod_{i=1}^{d} \mathbb{E}\left[(J^{(b_{i-1})})^T J^{(b_i)}\right].$$

For the sake of simplicity, assume $\sigma(z) = z$, i.e., the linear case. The general case is covered with a sequence of long and complicated procedures, similar with one in §4.2. With the linearity assumption, $J^{(1)} = (W^{(2)})^T \otimes X$ and $J^{(2)} = I \otimes W^{(1)} X$. Therefore, with the normality assumption

$$X \sim N(0, I_n) \qquad \text{and} \qquad W_{ij}^{(\alpha)} \sim N\left(0, \frac{1}{n}\right),$$

the observation above gives

$$\mathrm{tr}[H_0^d] = \mathbb{E}_W\left[\sum_{k=0}^{d} \binom{d}{k} \mathrm{tr}((W^{(2)})^T W^{(2)})^{d-k} \, \mathrm{tr}((W^{(1)})^T W^{(1)})^k\right] = \sum_{k=0}^{d} \binom{d}{k} C_{d-k} C_k.$$

Therefore, the Stieltjes transform of H_0 is as follows:

$$m(z) = \sum_{k=0}^{\infty} \frac{1}{z^{k+1}} \sum_{k=0}^{d} \binom{d}{k} C_{d-k} C_k$$

so that

$$\rho_{H_0}(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \downarrow 0} \mathrm{Im}\, m(\lambda + i\epsilon) = \frac{1}{2}\delta(\lambda) + \left[\frac{1}{2\pi^2} E\left(\frac{(8-\lambda)\lambda}{16}\right) + \frac{4-\lambda}{8\pi^2} K\left(\frac{(8-\lambda)\lambda}{16}\right)\right] \mathbf{1}_{[0,8]}(\lambda)$$

where the convergence is the weak convergence of measures, and K and E are the complete elliptic integrals of the first and the second-kind:

$$K(k) = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k \sin^2 \theta}}, \qquad E(k) = \int_0^{\pi/2} \sqrt{1 - k \sin^2 \theta}\, d\theta.$$

The result for the general case is as follows: Let σ be an arbitrary activation function, with zero mean and finite moments. Then the Stieltjes transform of the spectral density of the Fisher information matrix H_0 of a single-hidden-layer neural network described above is given as

$$m(z) = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{\lambda_1 + \lambda_2 - 2z}{2\zeta^2\big((\eta - \zeta)(\eta' - \zeta) + \lambda_1(z - \eta + \zeta) + \lambda_2(z - \eta' + \zeta) - z^2\big)} \, d\mu_1(\lambda_1)\, d\mu_2(\lambda_2)$$

where

$$\eta = \int_{\mathbb{R}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \sigma(x)^2 \, dx, \qquad \eta' = \int_{\mathbb{R}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \sigma'(x)^2 \, dx, \qquad \zeta = \left[\int_{\mathbb{R}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \sigma'(x)\, dx\right]^2,$$

$$d\mu_1(\lambda_1) = \frac{1}{2\pi} \sqrt{\frac{\eta' + 3\zeta - \lambda_1}{\lambda_1 - \eta' + \zeta}} \mathbf{1}_{[\eta' - \zeta, \eta' + 3\zeta]}(\lambda_1)\, d\lambda_1, \qquad d\mu_2(\lambda_2) = \frac{1}{2\pi} \sqrt{\frac{\eta + 3\zeta - \lambda_2}{\lambda_2 - \eta + \zeta}} \mathbf{1}_{[\eta - \zeta, \eta + 3\zeta]}(\lambda_2)\, d\lambda_2.$$

5 Conclusion

In this report, I have surveyed several studies connecting (fully connected feed forward) neural networks to some random matrix models. [Cho+14] tells the similarity of a multilayer neural network with the spherical spin-glass model, under many independence, normality, uniformity, and redundancy conditions. [PB17] found the spectral distribution of the hessian of the loss function, and [PW17] and [PW18] presented the spectral distributions of the Gram matrix and the Fisher information matrix of the network loss.

Random matrix theory has provided a new analytical framework for the analysis of the neural network. Apart from studies quoted here, there are many other researches making connection of the neural networks to the random matrix theory.

However, these approches surely have some limitations. The most important one is the assumption on (asymptotic) randomness or free independence of the matrices. For a neural network that has learnt some meaningful pattern, the weights must not be independent or random, and Gaussian neither. Despite those limitations, however, some numerical results show some of them are practically valid. I hope more general methods for those networks to be developed in the future.

References

[Cho+14] Anna Choromanska et al. "The Loss Surface of Multilayer Networks". In: *CoRR* abs/1412.0233 (2014). arXiv: `1412.0233`. URL: `http://arxiv.org/abs/1412.0233`.

[DC14] Thomas Dupic and Isaac P'erez Castillo. "Spectral density of products of Wishart dilute random matrices. Part I: the dense case". In: 2014.

[PB17] 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.* Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 2798–2806. URL: `http://proceedings.mlr.press/v70/pennington17a.html`.

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning". In: *Advances in Neural Information Processing Systems 30.* Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 2637–2646. URL: `http://papers.nips.cc/paper/6857-nonlinear-random-matrix-theory-for-deep-learning.pdf`.

[PW18] Jeffrey Pennington and Pratik Worah. "The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network". In: *Advances in Neural Information Processing Systems 31.* Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 5410–5419. URL: `http://papers.nips.cc/paper/7786-the-spectrum-of-the-fisher-information-matrix-of-a-single-hidden-layer-neural-network.pdf`.

[Spe09] Roland Speicher. "Free Probability Theory". In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 119 (2009), pp. 3–30.