Standard Condition Number of Hessian Matrix for Neural Networks

Lei Zhang[1], Wensheng Zhang[1], Yunzeng Li[1], Jian Sun[1], and Cheng-Xiang Wang[2]
[1]Shandong Provincial Key Lab of Wireless Communication Technologies,
School of Information Science and Engineering, Shandong University, Qingdao, Shandong, 266237, P.R.China.
[2]National Mobile Communications Research Laboratory,
School of Information Science and Engineering, Southeast University, Nanjing, Jiangsu, 210096, China.
Email: 18487240854@163.com, zhangwsh@sdu.edu.cn, 17865197576@163.com, sunjian@sdu.edu.cn, chxwang@seu.edu.cn

***Abstract*—Neural networks are becoming more and more important for intelligent communications and their theoretical research has become a top priority. Loss surfaces are crucial to understand and improve performance in neural networks. In this paper, the Hessian matrix of second order optimization method is analyzed through the analytical framework of random matrix theory (RMT) in order to understand the geometry of loss surfaces. The limited spectrum distribution, extreme eigenvalue distribution, and standard condition number (SCN) of Hessian matrix are analyzed to understand their asymptotic characteristics. Moreover, the relationships among the extreme eigenvalue distribution, SCN, and the convergence of loss surfaces are investigated. The above analyses give insight into utilizing RMT to analyze the neural network theory.**

***Index Terms*—Loss surfaces, Hessian matrix, random matrix theory, standard condition number, neural networks.**

I. Introduction

Communications technology is developing rapidly to satisfy the demands of a wide variety of application scenarios, such as the explosive growth of mobile data services and massive terminal connectivity to mobile networks. The fifth generation (5G) has proposed a variety of new technologies to improve the performance of communication systems. Millimeter wave, massive MIMO, and ultra-dense network are the three key technologies [1]. Above technologies need to satisfy the capability of handling large wireless data. The development of intelligent communication emerges as the times require. Artificial intelligence (AI) is used in various fields of communication, such as cognitive radio, channel estimation and detection, coding and decoding technology [2]–[5]. In order to meet the communication requirements of high reliability and low latency, the optimization performance of neural network needs to be further improved.

Deep learning is flourishing in various aspects of life based on the development of new hardware such as graphics processing unit (GPU), network architectures (RNN, CNN, ResNet, etc.), and improved optimization algorithms. The academic researches on neural network have been very active in the past decade, and hundreds of neural network models have been proposed for pattern recognition, signal processing, fault diagnosis, and computer vision [6]–[8]. However, theoretical researches on neural network are still in the initial stage, which will seriously affect the further development of the neural network. One of the most interesting questions is the characteristics of loss surfaces which is defined as the landscape of loss function relevant to the parameters in a neural network. The loss surfaces have high dimensions (massive data in communication with high dimensions) which often lead to the curse of dimensionality. Furthermore, loss surfaces are usually non-convex so that there are many local minimum points. Therefore, understanding the loss surfaces is one of the most confusing and incomprehensible part at present and it is an indispensable step in studying the optimization performance of neural networks.

The loss surfaces of neural networks have previously been studied in several ways. Q. Nguyen et al. illustrated that the critical point of each empirical loss is the global minimum of zero training error and a sufficiently wide network has a well behaved loss surfaces [9]. Y. Lecun et al. explored the non-convex loss function of a simple model with the spherical spin-glass model, and showed that the minimum critical value of the random loss function is located under a well-defined narrow band and is constrained by the global minimum [10]. J. Pennington et al. used the RMT to study the distribution of eigenvalue of Hessian matrix at critical point to characterize the change of the loss surfaces [11]. Sagun et al. conducted many training and data processing, and turned out that the Hessian is degenerate at any points, moreover, deduced that the characteristics of the loss surfaces are determined jointly by the structure of the network and the input dataset [12]. Louart et al. explored nonlinear problems in neural networks, studied the Gram random matrix model in neural networks, and found the resolvent of it [13].

However, there are still too many unknowns in exploring loss surfaces of neural networks. The dimension of neural network for communication is huge, every parameter in the network is a random variable, and the training process of neural network can be regarded as the operation process of large dimension matrices. RMT is suitable for solving the problems caused by the above characteristics [14]–[16]. Under this background, the random matrix becomes a good choice. More and more disciplines of science and engineering such as wireless communication, quantum mechanics, signal processing, and big data have found RMT valuable [17], [18].

In this paper, a multilayer feed-forward neural network is

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

built for multi-classification problem in order to explore the loss surfaces in the network, and a second order optimization method is adopted. The Hessian matrix which is the second derivatives of the loss function with respect to the weight reflects the convexity of loss surfaces. Then, exact Hessian matrix of the model is calculated and proved to be a Wishart matrix. The geometry of error surfaces is analyzed by using limit spectrum theory (Marcenko-Pastur distribution), extreme eigenvalue distribution (Tracy-Widom distribution) and SCN (Tracy-Widom-Curtiss distribution) of infinite random matrices theory (IRMT). The asymptotic characteristics of loss surfaces are analyzed to obtain the convergence performance of the neural network. Based on the above research, RMT and neural network have a high correlation. To the best of our knowledge, the analyses of extreme eigenvalue and SCN of Hessian are still missing in the literature. The major contributions and novelties of our work are as follows:

- Based on RMT analytical framework, the asymptotic performance of the high-dimensional loss surfaces is analyzed. It is proved that RMT is a powerful tool for analyzing neural network theory.
- The extreme eigenvalue characteristics and SCN of IRMT are firstly applied to the analysis of neural networks.
- The relationships among SCN, extreme eigenvalue and the convergence of neural network are given. It provides a theoretical basis for further enhancing the optimization performance of neural network, and further ensures the high reliability of communication.

These provide several insights into applications and shed light into understanding the loss surfaces of neural network.

The remainder of this paper is organized as follows. Section II describes the neural network model and calculates the exact Hessian matrix. The analysis and derivation process of RMT are presented in Section III. The numerical results are discussed in Section IV. Finally, the conclusions are illustrated in Section V.

II. System model

A. The Multi-classification Model

In this paper, a multilayer feed-forward neural network for multi-classification problem is considered, it is widely used in communication, such as modulation detection, signal processing, and cognitive radio [19], [20]. The system model is introduced in Fig. 1. The input of the neural network is a set of data $\boldsymbol{x} \in \mathbb{R}^{M \times D}$, which is trained to map the target vector $\boldsymbol{y} \in \mathbb{R}^{N \times D}$. The iterative formula $a^k = b^k + w^k h^{k-1}$ is used to calculate the pre-activations for each layer, and $h^0 = x$, where a_k^L denotes the kth pre-activation of L layer, b is the bias of the network, w denotes the weight of the network and h_j^L denotes the jth activation of L layer. For multiple classification problems, the softmax activation function is used at the last layer and can be calculated as

$$h_j^L = \frac{e^{a_j^L}}{\sum_k e^{a_k^L}}. \tag{1}$$

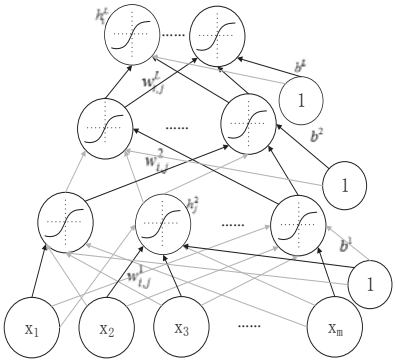


Fig. 1. A multilayer feed-forward neural network.

It can be found that all a_k^L are located in $(0, 1)$, by using the normalization factor, $\sum_k e^{a_k^L}$, the sum of all a_k^L is normalized to one. The hidden layer activation can be expressed as $h(x) = g(a(x))$, where $g(\cdot)$ is an entry-wise rectified linear unit (ReLU) function. The matching of log-likelihood loss function and softmax activation function will achieve very good training results in multi-classification problem [21], the loss L can be calculated as

$$L = -\sum\nolimits_k y_k \log h_k^L. \tag{2}$$

B. Calculate Exact Hessian Matrix

In order to obtain better convergence performance, we make full use of second-order optimization method. The second-order properties are crucial to the neural network, and loss surfaces can be studied through it. For classical back-propagation, the Jacobian matrix, which is the first derivatives of loss function with respect to weight matrix is considered. However, in this network, we are more interested in the properties of the second derivatives of loss function, which are expressed in the form of Hessian matrix. The gradient of the loss is given by

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial w} = \sum\nolimits_k (h_k^L - y_k)(h_k^{L-1})^{\mathrm{T}}. \tag{3}$$

The Hessian matrix can be calculated as

$$\boldsymbol{H} = \frac{\partial^2 L}{\partial w^2} = \frac{\partial^2 L}{\partial a^2}\frac{\partial a}{\partial w}\frac{\partial a}{\partial w}^{\mathrm{T}} + \frac{\partial L}{\partial a}\frac{\partial^2 a}{\partial w^2} \tag{4}$$

where $\frac{\partial^2 a}{\partial w^2}$ reflects the linearity of the network, and tends to 0 in that a is a linear function about the parameter w. Therefore, a simplifying approximation of the Hessian can be calculated as

$$\boldsymbol{H} \sim \frac{\partial^2 L}{\partial a^2}\frac{\partial a}{\partial w}\frac{\partial a}{\partial w}^{\mathrm{T}} = \sum\nolimits_k h_k^L(1 - h_k^L)h_k^{L-1}h_k^{L-1^{\mathrm{T}}}. \tag{5}$$

Sagun in [12] proved that above approximation will become more empirical as the training process progresses, especially in the case of global minimum. Then, the following definitions are obtained

$$\boldsymbol{G} = \sqrt{\frac{\partial^2 L}{\partial a^2}} \frac{\partial a}{\partial w} \tag{6}$$

$$g = \sqrt{h^L(1-h^L)} h^{L-1}. \tag{7}$$

With the above information, the Hessian matrix can be redefined as

$$\boldsymbol{H} \sim \frac{1}{N} \sum\nolimits_k g_k {g_k}^{\mathrm{T}} = \frac{1}{N} \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}. \tag{8}$$

So the Hessian of loss is a positive semi-definite matrix, and the form of Hessian can be taken as a covariance matrix computed from $\boldsymbol{G}$.

III. Random matrix theory analytical framework

In this section, several results of Hessian matrix are illustrated based on the theoretical support of RMT, which analyzes the asymptotic statistical properties of matrices. In the large dimensional matrix, such distribution is like the law of large numbers in statistics. For a multilayer neural network, the dimensions of input and parameters are very large, and so is the dimension of Hessian. Moreover, every element of Hessian is a random variable. With the fact, according to dimension boundary in [22], IRMT can be a sharp tool to deal with the problem of second-order optimization in neural network. In this paper, we mainly focus on the analysis of asymptotic spectrum theory.

A. Marcenko-Pastur Distribution

The Hessian matrix in (8) is in the form of Wishart ensemble for we assume that $\boldsymbol{G}$ is independent and identically distributed (i.i.d). Thus, the Hessian matrix can be written as $\boldsymbol{H} = \frac{1}{N}\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} \sim W_N(D, \sum)$, where N denotes the dimension of Hessian, D is the degree of freedom, and $\sum$ denotes the covariance matrix. The eigenvalue is the steepness of the loss along the direction of its corresponding eigenvector. Therefore, studying the distribution of eigenvalues is particularly important in exploring the geometric characteristics of loss surfaces. With the help of RMT analytical framework, the limiting spectral distribution of $\boldsymbol{H}$ follows Marcenko-Pastur distribution. For over-parametered neural network, we can make the following assumptions. As M, N, and D are very large, it is assumed that $M, N, D \to \infty$, but M is comparable to D, and when $\frac{N}{D} \to c$, where c denotes aspect ratio of matrix, the empirical spectral distribution of matrix converges to Marcenko-Pastur distribution with probability of one [23].

The probability density function (PDF) and cumulative distribution function (CDF) of eigenvalue can be expressed as [24]

$$f_\lambda(x) = \begin{cases} \frac{1}{2\pi c x}\sqrt{(b-x)(x-a)} & a \le x \le b \\ 0 & otherwise \end{cases} \tag{9}$$

$$\begin{aligned} F_\lambda(x) =& \frac{1}{2} + f_\lambda(x) + \frac{1-c}{2\pi}\ \arcsin\ (\frac{(1+c)x-(1-c)^2}{2x\sqrt{c}}) \\ &+ \frac{1+c}{2\pi}\arcsin(\frac{1+c-x}{2\sqrt{c}}) \end{aligned} \tag{10}$$

where λ denotes the eigenvalue of Hessian, and $\lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots \le \lambda_N$, $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$ denote upper bound and lower bound of λ, respectively. Marcenko-Pastur distributions under different aspect ratio are shown in Fig. 2 (a).

B. Extreme Eigenvalue Distribution

For the nonlinear system such as neural network, the extreme eigenvalue is very important. In neural network, the largest and smallest eigenvalue of Hessian matrix determine the steepness of the loss surfaces and affect the convergence performance. The PDF and CDF of extreme eigenvalue can use Tracy-Widom distribution to calculate [25]

$$f_{\lambda_E}(x) = \frac{dF_{\lambda_E}(x)}{dx} \tag{11}$$

$$F_{\lambda_E}(x) = \exp(-\int_x^\infty (x-r)q^2(x)dx) \tag{12}$$

where the extreme eigenvalues are defined as $\lambda_{EN} = \frac{\lambda_N - b}{\mu}$ and $\lambda_{E1} = \frac{\lambda_1 - a}{\nu}$, in which μ and ν are normalized factors and defined as $\mu = (\sqrt{N}+\sqrt{D})(\frac{1}{\sqrt{N}} - \frac{1}{\sqrt{D}})^{\frac{1}{3}}$ and $\nu = (\sqrt{N}-\sqrt{D})(\frac{1}{\sqrt{N}} - \frac{1}{\sqrt{D}})^{\frac{1}{3}}$, and $q(x)$ is the unique solution of Painlevé equation of type $q'' = xq + 2q^3$ and satisfied the following boundary condition

$$q(x) \sim \frac{1}{2}\pi^{-\frac{1}{2}} x^{-\frac{1}{4}} \exp(-\frac{2}{3}x^{\frac{3}{2}}), \qquad x \to \infty. \tag{13}$$

In this analytical framework, the largest eigenvalue is analyzed by the above Tracy-Widom distribution and shown in Fig. 2 (b) with different dimensions.

C. Standard Condition Number

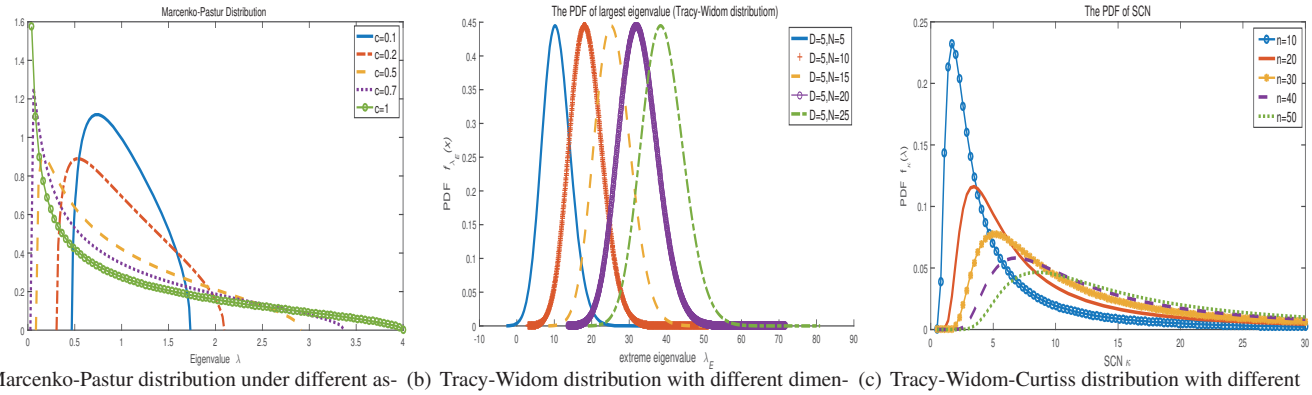
SCN is an important tool to study the distribution of eigenvalues. SCN reflects the stability of systems, and shows how fast the function changes for tiny fluctuation in input. This is the characteristic of the matrix itself. However, it has an unusual impact on network convergence. SCN is the ratio of the largest eigenvalue λ_N to the smallest eigenvalue λ_1, and can be represented by

$$\kappa = \frac{\lambda_N}{\lambda_1}. \tag{14}$$

The PDF and CDF of κ follow Tracy-Widom-Curtiss distribution [22] as

$$\begin{aligned} f_\kappa(x) &= -\tfrac{1}{\mu\nu}\textstyle\int_0^\infty r f_{\lambda_E}(\frac{rx-b}{\nu}) f_{\lambda_E}(\frac{r-a}{\mu})dr, \\ F_\kappa(x) &= \textstyle\int_{\frac{\alpha}{\mu}}^\infty f_{\lambda_E}(-r) F_{\lambda_E}(\frac{x(b-\mu r)-a}{\nu})dr. \end{aligned} \tag{15}$$

The PDF of SCN is shown in Fig. 2 (c).



(a) Marcenko-Pastur distribution under different aspect ratio. (b) Tracy-Widom distribution with different dimensions. (c) Tracy-Widom-Curtiss distribution with different dimensions.

Fig. 2. IRMT theoretical distributions

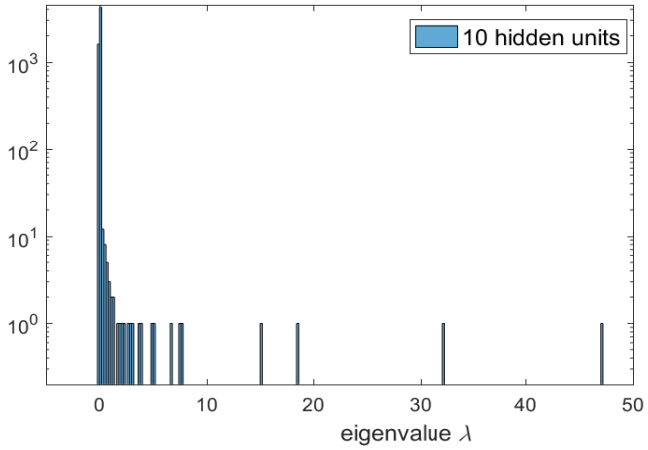


Fig. 3. The spectrum of Hessian matrix after convergence.

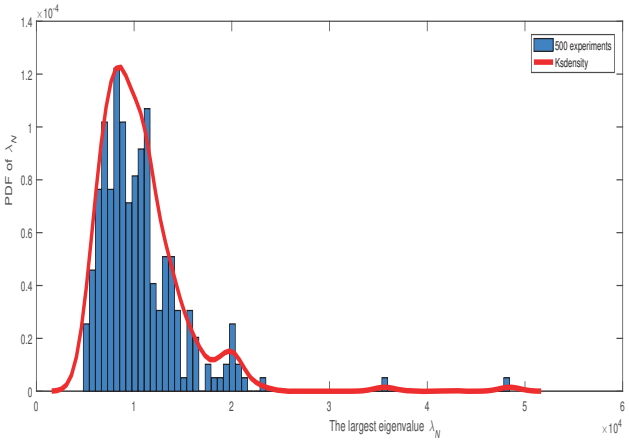


Fig. 4. The largest eigenvalue distribution of 500 repeated tests.

IV. Experiment Results

In this section, we have conducted many experiments to calculate the exact Hessian matrix on the multi-classification neural network in Section II. The training data are all samples from MNIST database for handwritten digit recognition, which is a well known datasets in multi-classification [26]. The exact Hessian matrix can be calculated after convergence. For a neural network with 10 hidden units, it is obversed that the Hessian matrix which contains 7840 $\times$ 7840 elements is very sparse, and the eigenvalue distribution is shown in Fig. 3.

It can be observed that most of the eigenvalues are concentrated near zero. However, there are still a small number of larger eigenvalues. Although it seems that the empirical distribution is quite different from Marcenko-Pastur distribution, there is still much in common between them. From Fig. 2 (a), it can be seen that the eigenvalue gap and tail are determined by aspect ratio. The characteristics of zero eigenvalue gap and long tails are similar to the Marcenko-Pastur distribution with an aspect rate of 1. Just the Hessian matrix is a square matrix that satisfies the aspect ratio of 1. It is observed that there are still some negative quantities in the eigenvalues. However, these values have a minimal order of magnitude (less than 10^{-14} magnitude). In this case, although the point of convergence is not a global optimum, the local optimum is abysmally close to the global optimum. Accordingly, the gradient descent can help escaping from the worse local optimum in the process of training and achieving good optimization performance. Under such conditions, largest eigenvalue attracts more attention. It is observed that the largest eigenvalue is far away from the high frequency center of the distribution, that is, concentrated near zero. At the same time, the distance difference between them exceeded three times the standard deviation. Furthermore, the frequency of the large eigenvalue is very small. Based on the research hereinbefore, such a large eigenvalue is called outlier [27].

We trained a 784-5-10 neural network 500 times and the distribution of the largest eigenvalues is shown in Fig. 4. The PDF of largest eigenvalue is also obtained by using ksdensity function in Matlab. The Tracy-Widom distribution which is shown in Fig. 2 (b) is a statistical bump with a steeper asymmetry on the left side than on the right side. It is observed that the experimental distribution of the largest eigenvalue also

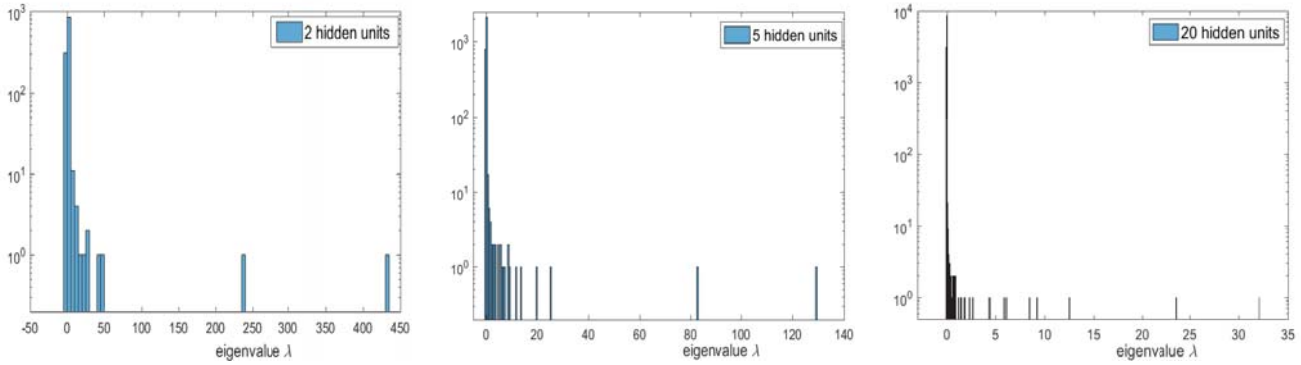


Fig. 5. The distribution of eigenvalues of different network structures.

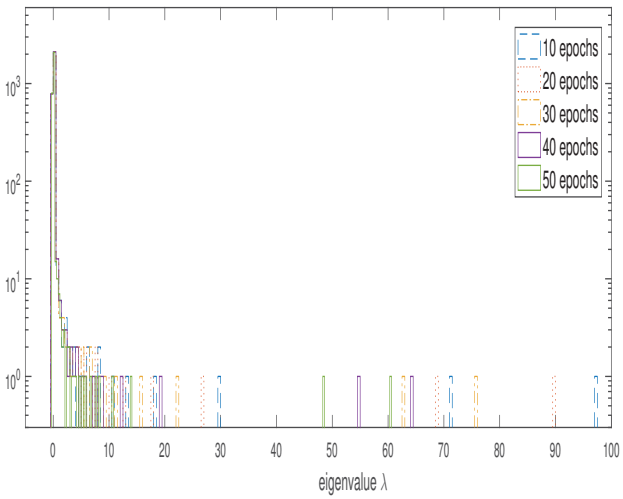


Fig. 6. The eigenvalue distribution of Hessian in training process.

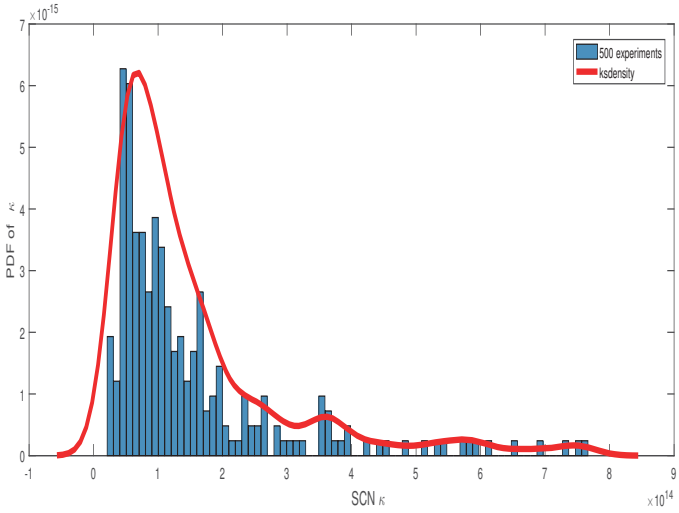


Fig. 7. The SCN distribution of 500 repeated tests.

satisfies the above characteristics. Fig. 5 and Fig. 3 show the distribution of eigenvalues of different network structures. It can be observed that with the number of hidden layer units increases, the largest eigenvalue is decreasing. Fig. 6 depicts the change of eigenvalue distribution during training process. We calculate Hessian matrix every 10 epochs and find that the largest eigenvalue decreases gradually during the convergence process.

The relationship between the largest and the smallest eigenvalue is also a key factor determining the loss surfaces. It is known that for a quadratic loss function, the loss surface is similar to the shape of an ellipse. The smallest eigenvector determines the direction of the long axis, and the size of the long axis is inversely proportional to the square root of the smallest eigenvalue. So are the short axis and the largest eigenvalue [28].

There are two forms of the relationship between the largest eigenvalue and the smallest eigenvalue. The first is the difference between the two, in this neural network, the difference is very close to the largest eigenvalue. The larger the difference, the short axis of the ellipse will be shorter, and the flatter the ellipsoid. Such the loss surface is very steep and the optimization path needs a lot of detours, and the computation efficiency will be very low. The second is the ratio of the two that is SCN. In practice, the ratio of the two is very large, so we use a cardinal number n which is equal to 10^{14} to represent. In Fig. 7, we calculate the distribution of SCN of 500 repetition test and plot the PDF of the SCN. It is observed that the density distribution of SCN has long tail which is very similar to the Tracy-Widom-Curtiss distribution in Fig. 2 (c). Therefore, the experimental results coincide with the theoretical distribution well. In Fig. 8, the convergence characteristics with different SCN are illustrated. It can be observed that the larger the SCN, the worse the network convergence performance. Morever, the speed of convergence is also related to SCN, and the convergence rate slows down with the increase of SCN. Thus, in the intelligent communication network, we will design neural network with small SCN to ensure timely and reliable communication.

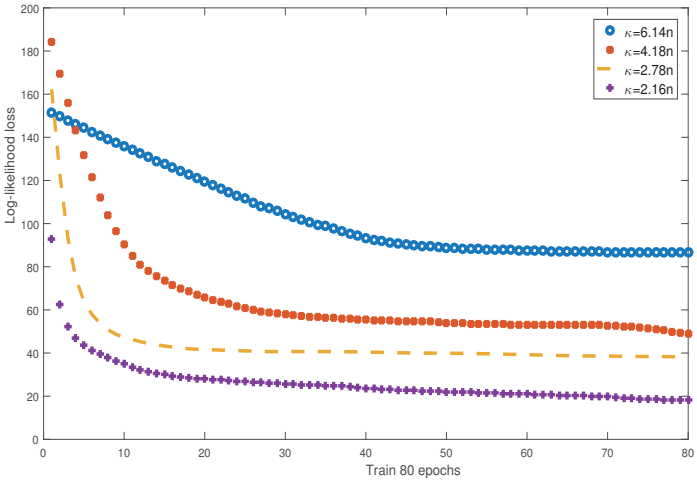


Fig. 8. Convergence properties of different SCN.

V. Conclusions

In order to guarantee the low latency and high reliability of intelligent communication, this paper tries to improve the optimization performance of neural networks. Understanding the loss surfaces in neural networks is particularly important in this problem. An analytical framework has been built for studying the Hessian matrix of second order optimization performance. The limiting spectral distribution, extreme eigenvalue distribution, and SCN have been analyzed using IRMT, and some asymptotic properties of loss surfaces have also been obtained. RMT can become a powerful tool in deep learning theory analysis. Moreover, the relationship between SCN and convergence has been investigated. This work can shed light into understanding the loss surfaces and improving the optimization performance in neural networks. In the future, we will conduct depth analysis of what causes SCN differences, and give better suggestions in practical applications based on SCN. Intelligent communication will further achieve better performance based on this theory.

Acknowledgment

This work was supported by the Natural Science Foundation of China (No. 61771293), Shandong Provincial Natural Science Foundation (ZR2017MF012), Fundamental Research Funds of Shandong University (2017JC029), Taishan Scholar Program of Shandong Province, Science and Technology Project of Guangzhou (201704030105), and EU H2020 RISE TESTBED Project (734325).

References

[1] C.-X. Wang, F. Haider, X. Gao, X. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[2] T. Wang, C. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Commun.*, vol. 14, no. 11, pp. 92–111, Nov. 2017.

[3] L. Bai, C.-X. Wang, J. Huang, Q. Xu, Y. Yang, G. Goussetis, J. Sun, and W. Zhang, "Predicting wireless mmWave massive MIMO channel characteristics using machine learning algorithms," *WirelessCommun. Mob. Com.*, vol. 2018, Aug. 2018.

[4] J. Huang, C.-X. Wang, B. L, J. Sun, Y. Yang, J. Li, O. Tirkkonen, and M. Zhou, "A big data enabled channel model for 5G wireless communication systems," *IEEE Trans. Big Data*, 2019, in press.

[5] Y. Fu, S. Wang, C. Wang, X. Hong, and S. McLaughlin, "Artificial intelligence to manage network traffic of 5G wireless networks," *IEEE Netw.*, vol. 32, no. 6, pp. 58–64, Nov. 2018.

[6] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[7] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging: Beyond analytical methods," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 20–36, Jan. 2018.

[8] E. Nishani and B. io, "Computer vision approaches based on deep learning and neural networks: Deep neural networks for video analysis of human pose estimation," in *2017 6th MECO*, June 2017, pp. 1–4.

[9] Q. Nguyen and M. Hein, "The loss surface and expressivity of deep convolutional neural networks," in *arXiv*, 2017.

[10] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. Lecun, "The loss surfaces of multilayer networks," *Eprint Arxiv*, pp. 192–204, 2014.

[11] J. Pennington and Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory," in *International Conference on Machine Learning*, 2017, pp. 2789–2806.

[12] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, "Empirical analysis of the hessian of over-parametrized neural networks," 2017.

[13] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, Feb. 2017.

[14] Z. Liao and R. Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Trans. on Signal Process.*, vol. PP, Jan. 2017.

[15] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Process. Systems*. Curran Associates, Inc., 2017, pp. 2637–2646.

[16] Z. Ling and R. C. Qiu, "Spectrum concentration in deep residual learning: a free probability appproach," *CoRR*, vol. 1807.11694, 2018.

[17] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 674–686, Mar. 2017.

[18] J. W. Silverstein and P. L. Combettes, "Large dimensional random matrix theory for signal detection and estimation in array processing," in *IEEE 6th SP Workshop on Statistical Signal and Array Processing*, Oct. 1992, pp. 276–279.

[19] D. Wang, M. Zhang, Z. Li, J. Li, M. Fu, Y. Cui, and X. Chen, "Modulation format recognition and OSNR estimation using CNN-based deep learning," *IEEE Photon. Technol. Lett.*, vol. 29, no. 19, pp. 1667–1670, Oct. 2017.

[20] M. Masoumi and A. B. Hamza, "Spectral shape classification: A deep learning approach," *Journal of Visual Commun. Image Representation*, vol. 43, pp. 198–211, 2017.

[21] T. Kanamori, "Deformation of log-likelihood loss function for multiclass boosting," *Neural Networks*, vol. 23, no. 7, pp. 843–864, 2010.

[22] W. Zhang, C. X. Wang, X. Tao, and P. Patcharamaneepakorn, "Exact distributions of finite random matrices and their applications to spectrum sensing," *Sensors*, vol. 16, no. 8, pp. 1–22, 2016.

[23] V. A. Marchenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathemat ics of the USSR-Sbornik*, vol. 1, no. 1, pp. 457–483, 1967.

[24] W. Zhang, G. Abreu, M. Inamori, and Y. Sanada, "Spectrum sensing algorithms via finite random matrices," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 164–175, Jan. 2012.

[25] B. Widom, "Some topics in the theory of fluids," *Journal of Chemical Physics*, vol. 39, no. 11, pp. 2808–2812, 1963.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[27] R.H, "Outlier treatment in data merging," *Journal of Applied Crystallography*, vol. 30, no. 4, pp. 421–426, 2010.

[28] Y. Lecun, L. Bottou, G. B. Orr, and K. R. Mller, "Efficient backprop," *Lecture Notes in Computer Science*, vol. 1524, no. 1, pp. 9–50, 1998.