Benign Loss Landscape of Large Size Deep Nonlinear Neural Networks

Shuai Li
lishuai918@gmail.com

Abstract

We study the optimization problem of deep neural networks (NNs) and push forward its theoretical understanding. More specifically, we study the loss landscape of NNs, and prove that under a set of *practical* boundedness and diversity conditions, for *large size nonlinear deep* NNs with a class of losses, including the hinge loss, *all local minima are global minima with zero loss errors*, We emphasize that the assumptions made are sufficient practice-guiding preconditions instead of unrealistic assumptions made to make the proof work. Intuitively, the conditions ask the neurons in a NN to be cooperative yet stay autonomous to the majority of the neuron population. The conditions are directly related to the centering (Glorot & Bengio, 2010), and normalization (Ioffe, Sergey and Szegedy, 2015) techniques widely used in practice to make NNs optimizable.

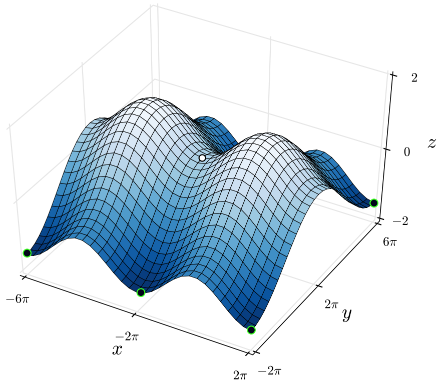

Figure 1. Illustration of loss landscape of NNs under our boundedness and diversity conditions. All stationary points that are not global minima are saddle points, i.e., the white dot in the middle of the graph, at which a descent direction that minimizes the loss can be found. All local minima are global minima, i.e., the black dots at the corners of the graph.

1. Introduction

Deep neural networks (NNs) are biologically inspired families of algorithms that model the perception part of intelligence, e.g., images recognition, in an astoundingly successful way, e.g, empowering human-level image classification system (He et al., 2015). It is tempting to hypothesize that it may hold the key to understanding at least perception intelligence (Kelly, 2015). However, the progress in the field has been largely driven by application-oriented interests, and a mathematical characterization has lagged behind. Among the theoretical aspects of NNs, we focus on the optimization behaviors in this work.[1]

The optimization problem of NNs is a long standing non-convex problem for decades. To introduce the idea would be present in this work, we look at the development of optimization in the past. Convexity analysis was a milestone in the optimization theory that gives an incisive categorization of optimization problems. For a time, the elegance of the formulation has tempted researchers to relax non-convex problems to convex ones. Yet, an "inelegant" approach that "only" converges to a local minimum often works better than the "elegant" convex relaxation ones (Jain & Kar, 2017). Meanwhile, the problems nature faces are often non-convex and accept no relaxation, e.g., the fitness function in biological evolution, the energy function of protein structure. Instead of seeing convex problems as the ideal form of optimization problems, we may think convexity as a component of a bigger theme. The idea has been in operation for quite some time in molecular science (Wales et al., 2003; Ballard et al., 2017), where a local minimum in the energy landscape could correspond to a metastable state of glasses, and an understanding of the landscape helps understand the states of Silicon dioxide, and transition between states, e.g., being crystals, or glasses.

We advocate that to study the optimization problems of NNs, *we should study how a benign real-world case of a system could be designed so that it helps minimize the objective functions, instead of finding artificial adversarial cases and*

[1] The technical content of this work is present in (Li, 2018). We add introduction to relevant domains involved in this manuscript. (Li, 2018) contains theoretical works on a formal measure-theoretical definition, the information geometrical structure of intermediate representations, a novel learning framework, and the optimization landscape of NNs. This work contains the optimization part.

intimidating ourselves by proving that it is hard, e.g., NP-hard. Evolution launches the game of the survival of the fittest to explore the fitness landscape to design organisms by working out tentative solutions and identifying improper designs. Fortunately, a NN is not constrained as much as an organism is constrained by its environment, and its loss landscape can be molded to our needs.

Contribution. Specifically in this work, we prove under a set of *practical* boundedness and diversity conditions, for *large size nonlinear deep* NNs with a class of losses, including the hinge loss, *all local minima are global minima with zero loss errors*, and regions around the minima are flat basins where all eigenvalues of Hessians are concentrated around zero. The loss landscape obtained is illustrated in fig. 1. We emphasize that the assumptions made are sufficient practice-guiding preconditions instead of unrealistic assumptions made to make the proof work. Intuitively, the conditions ask the neurons in a NN to be cooperative yet stay autonomous to the majority of the neuron population — illustrations of the conditions are shown at fig. 4 and fig. 5. The conditions are directly related to the centering (Glorot & Bengio, 2010), and normalization (Ioffe, Sergey and Szegedy, 2015) techniques widely used in practice to make NNs optimizable (discussed in section 4.1.2).

2. Related Works

The optimization problem of NNs is a long standing issue for decades that attracts many attempts to understand it. Due to page limitation, we focus on the works that attack the problem in its full complexity; that is, NNs with arbitrary input data distributions, nonlinear activation functions and number of layers. Interesting readers may refer to related works in (Soltanolkotabi et al., 2018; Chizat & Bach, 2018; Dauphin et al., 2014; Nguyen & Hein, 2017; Liang et al., 2018) for works before the deep learning era, on shallow networks and NP-hardness of NN optimization.

Roughly, two approaches have been taken in analyzing the optimization of NNs, one from the linear algebra perspective, the other from mean field theory using random matrix theory. Our work falls in the latter approach. The linear algebra approach, as the name suggests, shies away from the nonlinear nature of the problem. (Kawaguchi, 2016) proves all local minima of a deep linear NN are global minima when some rank conditions of the weight matrices are held. (Nguyen & Hein, 2017; 2018) prove that if in a certain layer of a NN, it has more neurons than training samples, which makes it possible that the feature maps of all samples are linearly independent, then the network can reach zero training errors. A few works following in the linear-algebraic direction (Laurent & von Brecht, 2018; Liang et al., 2018; Yun et al., 2018) improve upon the two previous results, but using essentially the same approach. As the conditions in (Kawaguchi, 2016) indicate, the rank related linear algebraic condition does not transport to nonlinear NNs. While for (Nguyen & Hein, 2017), it characterizes a phenomenon that if in a layer of a NN, it can allocate a neuron to memorize each training sample, then based on the memorization, it can reach zero errors. It is likely we do not need so many linearly independent intermediate features, which would lead to poor generalization. Thus, to truly understand the optimization behavior of NNs, we need to step out of the comfort zone of linearity. But we note that the linear algebraic approach is important, for finer grained analysis that may be combined with the mean field approach, in that NNs are cascaded generalized linear models.

The mean field theory approach using the tools of random matrix theory can attack the optimization of NNs in its full complexity, though existing works tend to be confused on the source of randomness. Due to an inadequate understanding of the randomness induced by activation function, (Choromanska et al., 2015a) tries to get rid of the activation function by assuming that its value is independent of its input, which is unrealistic (Choromanska et al., 2015b). Nevertheless, it is an important attempt, and the first paper to attack deep NNs in its full complexity. After (Choromanska et al., 2015a) which approaches by analogizing with spin glass systems — it is a complex system, as NNs are — some researchers start to study NNs from mean field theory from the first principle instead of by analog. Again, confused with the source of randomness in activation in the intermediate layers of NNs, (Jeffrey Pennington, 2017) just assumes data, weights and errors are of i.i.d. Gaussian distribution, which are mean field approach assumptions and unrealistic, and proceeds to analyze the Hessian of the loss function of NNs, though due to limitations of their assumptions, they can only analyze a NN with one hidden layer.

Compared with existing works, our results apply to NNs used in practice. Verbosely, theorem 4.1 holds for deep large size nonlinear NNs on any input data distributions, and its assumptions 4.1 4.2 are sufficient practical preconditions that identify a class of optimizable NNs used in practice.

3. Preliminaries

The proof is built on the key observation that NNs are inherently stochastic, thus the Hessian of the loss function can be analyzed with random matrix techniques. In this section, we introduce the preliminaries to understand theorem 4.1, and hopefully its proof. For the readers who are not familiar with measure theory, matrix calculus, or random matrix, the content is written with key concepts illustrated; you can substitute probability measure with probability distribution and pay more attention to illustrations.

We note the notation used. All scalar functions are denoted as normal letters, e.g., f; bold, lowercase letters denote vectors, e.g., $\boldsymbol{x}$; bold, uppercase letters denote matrices, e.g., $\boldsymbol{W}$; normal, uppercase letters denotes random elements/variables, all the remaining symbols are defined when needed, and should be self-clear in the context. r.v. and r.v.s are short for random variable, random variables respectively. To index entries of a matrix, following (Erdos et al., 2017), we denote $\mathbb{J} = [N] := \{1, \ldots, N\}$, the ordered pairs of indices by $\mathbb{I} = \mathbb{J} \times \mathbb{J}$. For $\alpha \in \mathbb{I}, A \subseteq \mathbb{I}, i \in \mathbb{J}$, given a matrix $\boldsymbol{W}$, $w_\alpha, \boldsymbol{W}_A, \boldsymbol{W}_{i:}, \boldsymbol{W}_{:i}$ denotes the entry at α, the vector consists of entries at A, the ith row, ith column of $\boldsymbol{W}$ respectively. Given two matrix $\boldsymbol{A}, \boldsymbol{B}$, the curly inequality $\preceq$ between matrices, i.e., $\boldsymbol{A} \preceq \boldsymbol{B}$, means $\boldsymbol{B} - \boldsymbol{A}$ is a positive definite matrix. Similar statements apply between a matrix and a vector, and a matrix and a scalar. $\succeq$ is defined similarly. $:=$ is the "define" symbol, where $x := y$ defines a new symbol x by equating it with y. tr denotes matrix trace. $\mathrm{dg}(\boldsymbol{h})$ denotes the diagonal matrix whose diagonal is the vector $\boldsymbol{h}$.

3.1. Measure Transport

NNs are widely taken as universal function approximators, and since a probability distribution is a function, it also can approximate probabilities. However, underlying the widely perceived picture described, a richer structure exists: the probability distribution a NN approximates is the probability measure transport from the data to the hidden neurons. To describe the structure, we introduce measure transport.

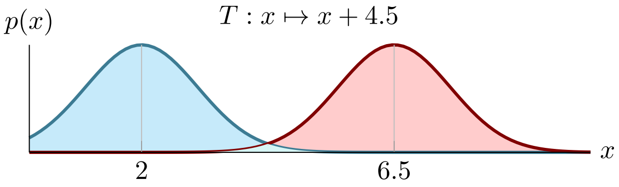

Figure 2. Measure Transport Illustration. The blue shaded area on the left is the probability measure of a random variable X taking value in $[-\infty, +\infty]$. It is a normal distribution centering at 2. A transport map $T : x \mapsto x + 4.5$ transports the measure of X to the measure of a new random variable $Y := X + 4.5$. The pushforward measure is the red shaded area on the right, centering at 6.5. That is, the measure transported from X to Y by T.

Given measurable spaces $(E, \mathcal{E})$, $(F, \mathcal{F})$, a measurable mapping $T : E \to F$, and a measure $\mu : \mathcal{E} \to [0, +\infty]$, the *pushforward probability measure* of μ by T is defined to be the measure $\nu : \mathcal{E} \to [0, +\infty]$ given by

$$\nu(B) = \mu(T^{-1}(B)), B \in \mathcal{E}$$

The mapping T is often called *transport map*. Without further delving into the formalism of probability measure theory, given a random variable (r.v.) X with a probability measure μ, a measurable mapping T, and the r.v. $Y := T(X)$ with probability measure ν obtained by applying the measurable mapping T on X, we say the *measure transported* from X to Y is the pushforward probability measure ν of μ by T. **An example of measure transport is shown in fig. 2**.

3.2. Statistical Learning

With measure transport at our disposal, we are ready to introduce the learning problem of supervised NNs.

Assume a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the space of input data, and $\mathcal{Y}$ is the label space. $Z := (X, Y)$ are the r.v.s with an unknown distribution μ, from which we draw samples. We use $S_m = \{s_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m$ to denote the training set of size m whose samples are drawn independently and identically distributed (i.i.d.) by sampling Z. Given a loss function l, the goal of learning is to identify a function $T : \mathcal{X} \mapsto \mathcal{Y}$ in a hypothesis space (a class $\mathcal{F}$ of functions) that minimizes the expected risk

$$R(T) = \mathbb{E}_{Z\sim\mu}\left[l\left(T(X), Y\right)\right],$$

Since μ is unknown, the observable quantity serving as the proxy to the true risk $R(f)$ is the empirical risk

$$R_m(T) = \frac{1}{m}\sum_{i=1}^{m} l\left(T(\boldsymbol{x}_i), y_i\right) \tag{1}$$

When the function T to learn is a NN, **the risk function R_m measures the empirical discrepancy between the probability measure $T(X)$ transported from X and the probability measure of Y.** Thus, the learning of NNs is to learn parametric approximation of measure of Y through minimizing a surrogate risk.

3.3. Neural Networks

Not only the ultimate output of a NN, $T(X)$, is a r.v., so are the activation in the intermediate layers of a NN. We characterize the finer stochastic structure in NNs using Multiple Layer Perceptron (MLP) in this section.

Denote g as the activation function, $\boldsymbol{W}$ the linear transform, X the input, $\hat{H}$ the output activation, of a layer of a NN. We have

$$\begin{aligned} \hat{H} &:= g(\boldsymbol{W}X) = \mathrm{ReLU}(\boldsymbol{W}X) \\ \mathrm{ReLU}(\boldsymbol{W}X) &:= \boldsymbol{W}X \odot H = \mathrm{dg}(H)\boldsymbol{W}X \\ H &:= \mathbf{1}^{\boldsymbol{W}X}, \mathbf{1}_i^{\boldsymbol{W}X} = \begin{cases} 1, \text{ if } (\boldsymbol{W}X)_i > 0 \\ 0, \text{ otherwise} \end{cases} \end{aligned} \tag{2}$$

where $\odot$ denotes Hadamard product (element-wise multiplication); ReLU denotes the Rectified Linear Unit; $\mathrm{dg}(H)$ is the diagonal matrix whose diagonal is H. We use the widely used activation function ReLU as an example. The derivation in this paper can be trivially generalized to activation functions like Swish (Ramachandran et al., 2017), or variants of ReLU (He et al., 2015; Clevert et al., 2015) etc. Generalizing to classic activation functions like Sigmoid or Tanh may need some more involved matrix calculus, but is fundamentally similar.

$H, \hat{H}$ are both r.v.s created by transporting measure of X through map $g(\boldsymbol{W}X), \mathbf{1}^{\boldsymbol{W}X}$ respectively, thus they are *stochastic* (which is a tautology). Repeat the process, and **we can write a MLP with a single output neuron as the following transport map**

$$T(X;\boldsymbol{\theta}) = X^T \prod_{i=1}^{L-1} \boldsymbol{W}_i \mathrm{dg}(H_i)\boldsymbol{\alpha} \tag{3}$$

where $\boldsymbol{\alpha}$ is a vector; i is the layer index; $\mathrm{dg}(H_i)$ is the diagonal matrix whose diagonal is H_i; $\boldsymbol{\theta}$ is the parameters of T, a.k.a. $\{\boldsymbol{W_i}\}_{i=1,\ldots,L-1}, \boldsymbol{\alpha}$; L is the layer number.

3.4. Hessian and Loss Landscape

Establishing NNs are an inherently stochastic model, we are ready to study the loss landscape. **As well known, the landscape of a function is characterized by the eigenspectrum of its Hessian**: when all eigenvalues are positive, we are at a local minimum (black dots in fig. 1); when all eigenvalues are negative, we are at a local maximum; otherwise, we are at a saddle point (the white dot in fig. 1).

Our goal is to investigate the fundamental principle that makes $R_m(T)$ in eq. (1) tractable when T is a NN. To do this, we study the risk landscape by studying the Hessian of $R(T)$ of a particular class of loss functions. To motivate the class of losses, we calculate the Hessian of eq. (1) when T is given by eq. (3). The choice of a NN with a single output neuron simplifies the problem considerably: it reduces a weighted summation of a great number of matrices to a single matrix, though we do not want to digress too much to elaborate the simplification. The calculation gives

$$\frac{d^2}{d\boldsymbol{\theta}^2} l(T(\boldsymbol{x};\boldsymbol{\theta}), y) = \tag{4}$$

$$l''(T(\boldsymbol{x};\boldsymbol{\theta}), y)\frac{d}{d\boldsymbol{x}}T(\boldsymbol{x};\boldsymbol{\theta})\frac{d}{d\boldsymbol{x}}T(\boldsymbol{x};\boldsymbol{\theta})^T \tag{5}$$

$$+ l'(T(\boldsymbol{x};\boldsymbol{\theta}), y)\frac{d^2}{d\boldsymbol{x}}T(\boldsymbol{x};\boldsymbol{\theta}) \tag{6}$$

Note that $\boldsymbol{x}$ is a realization sampled from X now. To further simplify the problem, we restrict the class of loss functions to study as following. We study the class $\mathcal{L}_0$ of functions l and the class of NNs T, such that for $l \in \mathcal{L}_0$, it satisfies:

1. $l : \mathbb{R} \to \mathbb{R}^+$, when y is taken as a constant;
2. l is convex;
3. the second order derivatives $\frac{d^2}{d\boldsymbol{x}^2} l$ is zero;
4. $\min_x l(\boldsymbol{x}, y) = 0$, while for T it satisfies: $\dim(T(\boldsymbol{x};\boldsymbol{\theta})) = 1$.

The restriction allows us to study the most critical aspect of the risk function of supervised NNs by making eq. (5) zero, while the eq. (6) a single matrix. The class of l includes important loss functions like the **hinge loss** $\max(0, 1 - \hat{H}_L Y)$, and the absolute loss $|\hat{H}_L - Y|$, which were studied in (Choromanska et al., 2015a) under unrealistic assumptions.

Equation (6) is of the form as following. The calculation is deferred to appendix A.1. Denote

$$\begin{aligned}\boldsymbol{H}_{pq} =& l'(T\boldsymbol{x}, y)\mathrm{dg}(\tilde{\boldsymbol{h}}''_q) \prod_{k=q+1}^{L-1} (\boldsymbol{W}_k \mathrm{dg}(\tilde{\boldsymbol{h}}''_k))\boldsymbol{\alpha} \\ &\otimes \prod_{j=p+1}^{q-1} (\mathrm{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}_j^T)\mathrm{dg}(\tilde{\boldsymbol{h}}'_p) \\ &\otimes \boldsymbol{x}^T \prod_{i=1}^{p-1} (\boldsymbol{W}_i \mathrm{dg}(\tilde{\boldsymbol{h}}_i))\end{aligned} \tag{7}$$

We have the Hessian of l as

$$\boldsymbol{H} = \begin{bmatrix} \mathbf{0} & \boldsymbol{H}_{12}^T & \ldots & \boldsymbol{H}_{1L}^T \\ \boldsymbol{H}_{12} & \mathbf{0} & \ldots & \boldsymbol{H}_{2L}^T \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{H}_{1L} & \boldsymbol{H}_{2L} & \ldots & \mathbf{0} \end{bmatrix} \tag{8}$$

For now, we just need to focus on the form of eq. (8), and ignore the complicated details, e.g., Kronecker products. Notice that $\boldsymbol{x}$ and $\tilde{\boldsymbol{h}}$ (subscripts omitted) are both realizations sampled from X and H (derivative symbols like $''$ can be ignored for now, and can be taken as a function of H). Thus, **Hessian $\boldsymbol{H}$ is a matrix obtained by applying a complicated transport map on X. The matrix is a symmetric random matrix.** It implies the Hessian $\boldsymbol{H}$ of $R(T)$ is a random matrix created by summing random matrices, each of which is a gradient l' multiplies the Hessian of T. **Its eigen-spectrum can be studied through random matrix theory** (Tao, 2012). Thus, we can have a complete characterization of the landscape of $R_m(T)$ of eq. (1).

3.5. Random Matrices with Correlations

We introduce the core idea in random matrix theory, and how it leads to the study of the eigen-spectrum of $\boldsymbol{H}$ in eq. (8). To begin with, we introduce the phenomenon of *concentration of measure*.

The core idea of concentration of measure has all been contained in the classic normal distribution. Given an ensemble of i.i.d. r.v.s centering at 0, since the r.v.s are independent with each other, and have an equal probability being positive, or negative, they tend to cancel each other out when averaing their values, and approach a normal distribution. Thus, sampling many samples from the average of the ensemble, 95% of them would fall in the range between $[+2\sigma, 2\sigma]$, where σ is the standard deviation of the ensemble.

What we need is the concentration of measure in the space of matrices. For $1 \leq i < j < \infty$, let a_{ij} be i.i.d. r.v.s. with mean 0 and variance 1, and $a_{ij} = a_{ji}$. Let a_{ii} be i.i.d. real r.v.s. with mean 0 and variance 1. Then $\boldsymbol{A}_N = [a_{ij}]_{i,j=1}^N$ is a matrix with random entries. It is the classic random matrix named *Wigner matrix*. Given an eigenvector $\boldsymbol{t}$ of $\boldsymbol{A}_N$, we know that when we multiply it with $\boldsymbol{A}$, we have $(\boldsymbol{At})_i = \sum_k a_{ik}t_k$. Note that when $\{a_{ik}t_k\}_{k=1,\ldots,N}$ are i.i.d. r.v.s, $a_{ik}t_k$ also tend to cancel each other out, resulting in a close-to-zero eigenvalue. Thus, when the dimension of the matrix N is large enough, we would have the eigenspectrum of the distribution shown in fig. 3. It is termed *semi-circle law*. The eigen-spectrum is obtained by solving a *self-consistent* equation as following

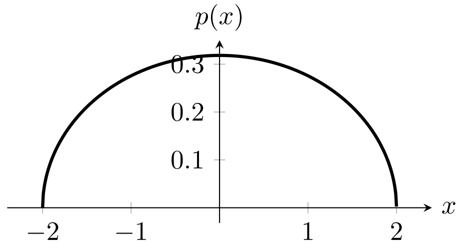

Figure 3. Eigen-spectrum of Wigner matrix. The x-axis is the numerical value of eigenvalues, and the y-axis is the probability of obtaining a eigenvalue of a specific numeric value if we randomly sample a eigenvalue of all the eigenvalues of $\boldsymbol{A}_N$.

$$1 + (z + m)m = 0, \Im m > 0, \Im z > 0$$

where z is a constant complex number, m is the unknown complex number to solve, and $\Im$ means taking the imaginary part of its operand. Then, we apply *inverse stieltjes transform* to get the spectrum.

Note that Wigner matrix has very strong requirements on its entries: independent identical distributed. However, notice that the key phenomenon we are needed is just the concentration of measure, a.k.a., $\{a_{ik}t_k\}_{k=1,\ldots,N}$ tend to cancel each other out. Thus, the conditions we need are just they are of low correlation with each other to obtain a semi-circle-law-like eigenspectrum. Indeed, this can be done. Under correlation characterized by conditions 4.1 4.2, **we can obtain the eigenspectrum of a symmetric random matrix by solving a matrix version of the above scalar equation**, termed *Matrix Dyson Equation*

$$\boldsymbol{I} + (z - \boldsymbol{A} + \mathcal{S}[\boldsymbol{G}])\boldsymbol{G} = \boldsymbol{0}, \Im \boldsymbol{G} \succ \boldsymbol{0}, \Im z > 0 \qquad (9)$$

, where $\Im \boldsymbol{G} \succ 0$ means $\Im \boldsymbol{G}$ is positive definite, and the rest of the symbols are defined at eq. (10). The details on how the equation is derived, and how it is related to eigenspectrum are explained in appendix A.2. We only aim to debrief its intuition in this section.

4. Main Results

Now we are ready to introduce our main results. Essentially, we calculate the Hessian $\boldsymbol{H}$ (eq. (8)) of the risk function $R(T)$ of NNs (eq. (1)), and analyze the eigen-spetrum of $\boldsymbol{H}$ with eq. (9), and obtain the following theorem. We present the theorem first, which are rather dense math, then the proof sketch; then elaborate on the assumptions in section 4.1, and the implications of the theorem in section 4.2; the full proof is put in appendix A, considering it may tax a high level of mathematical maturity to read.

With the following assumptions on $\boldsymbol{H}$, we show that l has a surprising benign landscape. Suppose $\boldsymbol{H}$ is a $N \times N$ matrix, and let

$$\boldsymbol{A} := \mathbb{E}[\boldsymbol{H}], \frac{1}{\sqrt{N}}\boldsymbol{W} := \boldsymbol{H} - \boldsymbol{A}, \mathcal{S}[\boldsymbol{R}] := \frac{1}{N}\mathbb{E}_{\boldsymbol{W}}[\boldsymbol{WRW}] \qquad (10)$$

where the expectation in $\mathcal{S}$ is taken w.r.t. $\boldsymbol{W}$ while keeping $\boldsymbol{R}$ fixed — it is a linear operator on the space of matrices. κ denotes cumulant, whose definition and norms, e.g., $|||\kappa|||$, are reviewed at appendix A.2. **Note that $\boldsymbol{W}$ is reused.**

Assumption 4.1 (Boundedness). *1) $\exists C \in \mathbb{R}, \forall N \in \mathbb{N}, ||\boldsymbol{A}|| \leq C$, where $||\boldsymbol{A}||$ denotes the operator norm. 2) $\exists \mu_q \in \mathbb{R}, \forall q \in \mathbb{N}, \forall \alpha \in \mathbb{J}, \mathbb{E}[|\boldsymbol{W}_\alpha|^q] \leq \mu_q$, where $\mathbb{J} = \mathbb{I} \times \mathbb{I}$, and $\mathbb{I} = \{1, \ldots, N\}$. 3) $\exists C_1, C_2 \in \mathbb{R}, \forall R \in \mathbb{N}, \epsilon > 0, |||\kappa|||_2^{iso} \leq C_1, |||\kappa||| \leq C_2 N^\epsilon$; 4) $\exists 0 < c < C, \forall \boldsymbol{T} \succ \boldsymbol{0}, c\, N^{-1} tr\, \boldsymbol{T} \preceq \mathcal{S}[\boldsymbol{T}] \preceq C N^{-1} tr\, \boldsymbol{T}$.*

Assumption 4.2 (Diversity). *There exists $\mu > 0$ such that the following holds: for every $\alpha \in \mathbb{I}$ and $q, R \in \mathbb{N}$, there exists a sequence of nested sets $\mathcal{N}_k = \mathcal{N}_k(\alpha)$ such that $\alpha \in \mathcal{N}_1 \subset \mathcal{N}_2 \subset \cdots \subset \mathcal{N}_R = \mathcal{N} \subset \mathbb{I}, |\mathcal{N}| \leq N^{1/2-\mu}$ and*

$$\kappa\Bigg(f(\boldsymbol{W}_{\mathbb{I}\setminus\cup_j \mathcal{N}_{n_j+1}(\alpha_j)}), g_1(\boldsymbol{W}_{\mathcal{N}_{n_1}(\alpha_1)\setminus\cup_{j\neq 1}\mathcal{N}(\alpha_j)}), \ldots,$$

$$g_q(\boldsymbol{W}_{\mathcal{N}_{n_q}(\alpha_q)\setminus\cup_{j\neq q}\mathcal{N}(\alpha_j)})\Bigg) \leq N^{-3q}||f||_{q+1}\prod_{j=1}^{q}||g_j||_{q+1}$$

, for any $n_1, \ldots, n_q < R$, $\alpha_1, \ldots, \alpha_q \in \mathbb{I}$ and real analytic functions $f, g_1, \ldots, g_q$, where $||||_p$ is the L^p norm on function space. We call the set $\mathcal{N}$ of α the **coupling set** *of α.*

Theorem 4.1. *Let $R(T)$ be the risk function defined at eq. (1), where the loss function l is of class $\mathcal{L}_0$ defined in* *section 3.4, and the transport map T is a neural network defined at eq. (3). If the Hessian $\boldsymbol{H}$ of $R(T)$ satisfies assumptions 4.1 4.2, $\mathbb{E}(\boldsymbol{H}) = \mathbf{0}$, and $N \to \infty$, then*

1. *all local minima are global minima with zero risk*
2. *A constant $\lambda_0 \in \mathbb{R}$ exists, such that the operator norm $||\boldsymbol{H}||$ of $\boldsymbol{H}$ is upper bounded by $\mathbb{E}_m[l''(T(X), Y)]\lambda_0$, where $\mathbb{E}_m[l'(T(X), Y)]$ is the empirical expectation of l'. It implies the regions around the minima are flat basins, where the eigen-spectrum of $\boldsymbol{H}$ is increasingly concentrated around zero.*

A note on the generality of the theorem. First, though the network T is restricted to be of a MLP defined in eq. (3), it is chosen this way to communicate the core idea better. The results can trivially generalize to arbitrary feed forward type network architecture, e.g., Convolutional Neural Network (Krizhevsky et al., 2012), Residual Network (He et al., 2016). Second, the requirement $N \to \infty$ is the result of a probably-approximately-correct type result, which means the result holds for finite N as well though with an error bound; the error for finite N is discussed at the remarks of theorem A.1.

Proof Sketch. First, we calculate the Hessian of eq. (6), done in appendix A.1. Then, as discussed before in section 3.4, the Hessian is a real symmetric random matrix. We introduce Matrix Dyson Equation eq. (9) (MDE) in details in appendix A.2, explaining how it is derived. Third, though the eigen-spectrum cannot be obtained in closed-form as semi-circle law in the case of Wigner matrix, we still can analyze it behaviors through MDE. In theorem A.2, we prove the solutions to MDE is symmetry w.r.t. y-axis under the assumptions of theorem 4.1: that is, given any positive eigenvalue λ, $-\lambda$ also is an eigenvalue. Since there are always non-zero eigenvalues by theorem A.1, we always have negative eigenvalues. Fourth, we break down the analysis of Hessian $\boldsymbol{H}$ into two cases: 1) for all training samples, at least one sample (x, y) has non-zero loss value; 2) and all training samples are classified properly with zero loss values. In case 1), by theorem A.2, $\boldsymbol{H}$ is always indefinite, thus corresponds to a saddle point. In case 2), we are at global optima, where all losses are zero. Lastly, theorem 4.1.2 is proved, which is straightforward. □

4.1. Boundedness and Diversity Assumptions are Practical Preconditions to the Power of NNs

We show that assumptions 4.1 4.2 are practical, and directly related to the normalization techniques in modern practices that make NNs optimizable. First, we explain the assumptions, mostly the diversity assumption in general in section 4.1.1. Then, we explain it in the context of NNs in section 4.1.2.

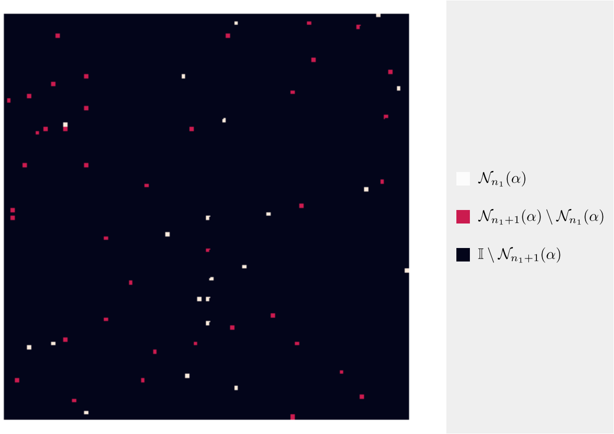
$\mathcal{N}_{n_1}(\alpha)$
$\mathcal{N}_{n_1+1}(\alpha) \setminus \mathcal{N}_{n_1}(\alpha)$
$\mathbb{I} \setminus \mathcal{N}_{n_1+1}(\alpha)$

Figure 4. Illustration of the boundedness and diversity assumptions. The image represents the square matrix $\boldsymbol{W}$ defined at eq. (10) with its index set $\mathbb{I}$ color coded. Using the second order cumulants $\kappa(f(\boldsymbol{W}_{\mathbb{I}\setminus\mathcal{N}_{n_1+1}(\alpha)}), g_1(\boldsymbol{W}_{\mathcal{N}_{n_1}(\alpha)}))$ as an example, the diversity assumption roughly states that for the two set $\mathcal{N}_1$ (color coded by white pixels) and $\mathbb{I} \setminus \mathcal{N}_2$ (color coded by black pixels) of entries of $\boldsymbol{H}$, the correlation between $\boldsymbol{W}_{\mathcal{N}_1}$ and $\boldsymbol{W}_{\mathbb{I}\setminus\mathcal{N}_2}$ is small. While notice that the boundedness assumption 3) is a bound that bounds the strength of the entries of $\boldsymbol{H}$ that do correlate, and it only requires they are somehow bounded. It implies that in $\mathcal{N}_2$, the correlation between entries could be strong. Note $n_1 = 1$ here. More specifically, the diversity assumption states that for any $\alpha \in \mathbb{I}$, nested sets $\mathcal{N}_1 \subset \mathcal{N}_2 = \mathcal{N}, |\mathcal{N}| \leq N^{1/2-\mu}$ exist (when only cumulants up to the second order exist, it suffices to let R be 2 instead of every $R \in \mathbb{N}$ (Erdos et al., 2017)), such that $\kappa(f(\boldsymbol{W}_{\mathbb{I}\setminus\mathcal{N}}), g_1(\boldsymbol{W}_{\mathcal{N}_1})) = \sum_{\beta \in \mathcal{N}_1' \subset \mathcal{N}_1 \cup \mathbb{I}\setminus\mathcal{N}} a_\beta \kappa(\beta) + \sum_{\beta,\gamma \in \mathcal{N}' \subset \mathcal{N}_1 \cup \mathbb{I}\setminus\mathcal{N}} a_{\beta,\gamma}\kappa(\beta,\gamma) \leq N^{-3}||f||_2||g_1||_2$, where $\kappa(\beta)$ is the mean of w_β, $\kappa(\beta,\gamma)$ is the covariance of w_β, w_γ, $\mathcal{N}_1'$ is a subset that depends on f, g_1, and $a_\beta, a_{\beta,\gamma}$ are real number coefficients depending on f, g_1. We do not repeat the quantitative formula of boundedness assumptions here.

4.1.1. Assumptions Explained

We defer the boundedness assumptions in the end since they are more obvious. The diversity assumption 4.2 roughly states that if we randomly pick q entries from the random matrix $\boldsymbol{H}$, the correlation of their coupling set, characterized by cumulants κ, is less than a certain value. We illustrate it qualitatively in a special yet practical case using concepts that are used more widely to characterize correlations, i.e., mean and covariance, though we emphasize that the assumption made is more general than the special case present.

Recall that the objective function is the empirical risk function $R(T)$ at eq. (1). Given a set of i.i.d. training samples $(X_i, Y_i)_{i=1,\ldots,m}$, $R(T)$ is a summation of the i.i.d. random matrices. Formally, reusing the notation to denote $\boldsymbol{H}$ the Hessian of $R(T)$ and $\boldsymbol{H}_i$ the Hessian of $l(T(X_i; \boldsymbol{\theta}), Y_i)$, we have

$$\boldsymbol{H} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{H}_i \tag{11}$$

Acute reader may realize that by the multivariate central limiting theorem ((Klenke, 2012) theorem 15.57), $\boldsymbol{H}$ will converge to a Gaussian ensemble, i.e., a random matrix of which the distribution of entries is a Gaussian process (GP), asymptotically as $m \to \infty$.

When $\boldsymbol{H}$ is a Gaussian ensemble, all higher order cumulants vanishes, thus the diversity assumption is solely about the second order cumulants, and the case when $q = 1$ and $q = 2$. Since $f, \{g_i\}_{i=1,\dots,q}$ are analytic, $\kappa(f(\cdot), g_1(\cdot))$ is a generalized cumulant ((McCullagh, 1987) Chapter 3), which can be decomposed into a sum of cumulants of entries of the Hessian. So is $\kappa(f(\cdot), g_1(\cdot), g_2(\cdot))$. Considering that only first and second cumulants exist, which are means and covariance respectively, $\kappa(f(\cdot), g_1(\cdot))$ *thus is a sum of means and covariance of the Hessian's entries*. So is $\kappa(f(\cdot), g_1(\cdot), g_2(\cdot))$. **We illustrate the correlation calculated in the diversity assumption $\kappa(f(\cdot), g_1(\cdot))$ in fig. 4, and it characterizes the weakness of the entries that are not correlated.**

The practicality of boundedness assumptions is obvious, since we do not want values to blow up. We only note for the two outliers. First, the lower bound in 4) in boundedness assumption, $cN^{-1}\text{tr}\,\boldsymbol{G} \preceq \mathcal{S}[\boldsymbol{G}]$, which is not about infinity. It asks the eigenvalues of $\mathcal{S}[\boldsymbol{G}]$ to stay close to its average value, so to let $\mathcal{S}[\boldsymbol{G}]$ stay in the cone of the positive definite matrices to ensure the stability of the MDE. It is essentially a constraint on the interaction of second order cumulants, and is realizable in a NN, though we are not clear on its physical meaning for the time being. Second, **boundedness assumption 3) characterizes the strength of the correlation of the entries in the random matrix that do correlate. It is also illustrated in fig. 4.**

4.1.2. Assumptions in NN Context

To see the practicality of assumptions 4.1 4.2 in the NN context, first we come back to the concrete form of Hessian. We rewrite eq. (7) in the following form (the equation should be read vertically)

$$\begin{aligned}
\boldsymbol{H}_{pq} = l'(T\boldsymbol{x}, y)\tilde{\boldsymbol{W}}_{q\sim(L-1)}\text{dg}(\tilde{\boldsymbol{h}}''_{L-1})\boldsymbol{\alpha} \quad &= \tilde{\boldsymbol{\alpha}}_q && \text{(12a)}\\
\otimes\, \tilde{\boldsymbol{W}}_{p\sim(q-1)} \quad &\otimes\, \tilde{\boldsymbol{W}}_{p\sim(q-1)} && \text{(12b)}\\
\otimes\, \boldsymbol{x}^T\tilde{\boldsymbol{W}}_{1\sim(p-1)} \quad &\otimes\, \tilde{\boldsymbol{x}}^T_{p-1} && \text{(12c)}
\end{aligned}$$

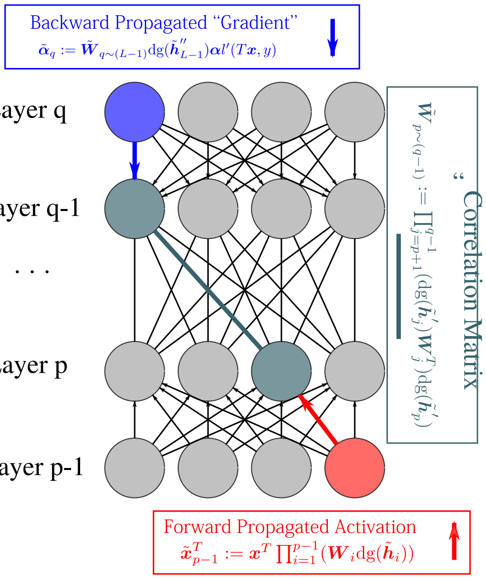
Backward Propagated "Gradient"
$\tilde{\boldsymbol{\alpha}}_q := \tilde{\boldsymbol{W}}_{q\sim(L-1)}\text{dg}(\tilde{\boldsymbol{h}}''_{L-1})\boldsymbol{\alpha}l'(T\boldsymbol{x}, y)$
Layer q
Layer q-1
. . .
Layer p
Layer p-1
"Correlation Matrix
$\tilde{\boldsymbol{W}}_{p\sim(q-1)} := \prod_{j=p+1}^{q-1}(\text{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}^T_j)\text{dg}(\tilde{\boldsymbol{h}}'_p)$
Forward Propagated Activation
$\tilde{\boldsymbol{x}}^T_{p-1} := \boldsymbol{x}^T\prod_{i=1}^{p-1}(\boldsymbol{W}_i\text{dg}(\tilde{\boldsymbol{h}}_i))$

Figure 5. Illustration of the entries of the Hessian of a NN. Informally, it characterizes the strength of the activation path drawn in colors that connects layer $p-1$ and layer q. Drawn in blue, eq. (12a) is a vector $\tilde{\boldsymbol{\alpha}}_q$ created by back propagating the vector $\text{dg}(\tilde{\boldsymbol{h}}''_{L-1})\boldsymbol{\alpha}l'(T\boldsymbol{x}, y)$ to layer q by left multiplying $\tilde{\boldsymbol{W}}_{q\sim(L-1)}$—note that if you replace $\boldsymbol{h}''_k$ with $\boldsymbol{h}'_k$, you get the *back propagated gradient*. Drawn in green, eq. (12b) is the *covariance matrix without removing the mean* between neurons at layer p and layer $q-1$, when taking expectation w.r.t. samples, i.e., $\mathbb{E}_{z\sim\mu^{\mathcal{Z}}}[\tilde{\boldsymbol{W}}_{p\sim(q-1)}]$ (it could be understood as the strength of the paths that connect the neurons at layer $p, q-1$). Drawn in red, eq. (12c) is the *forward propagated activation* at layer $p-1$.

where

$$\begin{aligned}
\tilde{\boldsymbol{W}}_{q\sim(L-1)} &:= \text{dg}(\tilde{\boldsymbol{h}}''_q)\prod_{k=q+1}^{L-2}(\boldsymbol{W}_k\text{dg}(\tilde{\boldsymbol{h}}''_k))\boldsymbol{W}_{L-1}\\
\tilde{\boldsymbol{\alpha}}_q &:= \tilde{\boldsymbol{W}}_{q\sim(L-1)}\text{dg}(\tilde{\boldsymbol{h}}''_{L-1})\boldsymbol{\alpha}l'(T\boldsymbol{x}, y)\\
\tilde{\boldsymbol{W}}_{p\sim(q-1)} &:= \prod_{j=p+1}^{q-1}(\text{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}^T_j)\text{dg}(\tilde{\boldsymbol{h}}'_p)
\end{aligned}$$

$$\begin{aligned}
\tilde{\boldsymbol{W}}_{1\sim(p-1)} &:= \prod_{i=1}^{p-1}(\boldsymbol{W}_i\text{dg}(\tilde{\boldsymbol{h}}_i))\\
\tilde{\boldsymbol{x}}^T_{p-1} &:= \boldsymbol{x}^T\tilde{\boldsymbol{W}}_{1\sim(p-1)}
\end{aligned}$$

As illustrated in fig. 5, the correlation between entries of the Hessian is the correlation between the products of forward propagated neuron activation $\tilde{\boldsymbol{x}}^T_{p-1}$ at layer $p-1$, the back propagated "gradient" $\tilde{\boldsymbol{\alpha}}_q$ at layer q, and the strength of activation paths $\tilde{\boldsymbol{W}}_{p\sim(q-1)}$ that connects the neurons at layer p and $q-1$. It can see that **the diversity assumption is**

on the weakness of the mean and covariance of $\tilde{\alpha}_i \tilde{w}_{jk} \tilde{x}_l$, the product of "gradient", activation path strength, and forward activation, while the boundedness assumption is about its strength. Additionally, to prove theorem 4.1, we need a further assumption that $\mathbb{E}[\boldsymbol{H}] = \boldsymbol{0}$. It clearly connects to the experiment tricks used in the community, such as the early initialization schemes that try to keep mean of gradient and activation zero, and standard deviation (std) small (Glorot & Bengio, 2010; He et al., 2015), the normalization schemes that keep the mean of activation zero and std small (Ioffe, Sergey and Szegedy, 2015; Salimans & Kingma, 2016). It is an interesting to work out a similar normalization technique derived from the assumptions rigorously in the future works.

The assumptions state that a diversity should exist in the neuron population, so that for any neurons, it does not strongly correlate with the majority of the neuron population. It very much corresponds to the hierarchical nature of the problem it aims to solve: a high level feature of an object, e.g, the leg of a person, does not need a high level feature of another object, e.g., the leg of a chair. Those features only correlate at very low level features, e.g., textures, and their neuron activation are of no, or low correlation. Qualitatively, we estimate how many neurons could be correlated in the following. Suppose we have a 10-layer NN with 100 neurons in each layer, then the Hessian $\boldsymbol{H}$ would be of dimension $N = 10^8$ (due to Kronecker product), then $N^{1/2}$ is 10^4. That is that the coupling set could have almost 10^4 $\tilde{\alpha}_i \tilde{w}_{jk} \tilde{x}_l$ tuples, i.e., the product of activation, activation path strength, and gradients. Note that we can incorporate the neuron that contributes gradient, i.e., neuron at layer q in fig. 5, and the neuron that contributes activation, i.e., neuron at layer $p-1$ in fig. 5, into the activation path that connects layer $p, q-1$. **That means for each activation path, it could have almost 10^4 number of correlated activation paths among only 10^3 neurons!** Each activation path probably codes some hierarchical patterns, e.g, the patterns of a person. It is reasonable to suggest that the capacity is quite significant. It is interesting to formulate testable quantities to see how much correlation exists in real world NNs. It also excitingly connects to the sparsity of neuron activation, and the long held hypothesis that information is coded in the neuron connection patterns in biological NNs.

Back to a larger context, NN is a fabulous mechanism that can indefinitely increase the number of parameters, thus its learning capacity, in a meaningful way, i.e., creating higher level features yet maintaining the diversity of the features created. Such mechanism does not normally hold in other systems or algorithms. Taking linear NNs for example, though with the potential to infinitely increase its parameters, matrices that multiply together still have a highly correlation structure within, thus cannot create a population of diverse neurons that are of low correlation with a majority of the other neurons. **Accompanying theorem 4.1, which states $R(T)$ can be optimized to zero, we can see that assumptions 4.1 4.2 actually characterizes sufficient preconditions to the optimization power of NNs.**

4.2. Implications

The the phenomenon characterized by theorem 4.1.1 is rather remarkable, if not marvelous. It shows that instead of seeing non-convex optimization as something to avoid, a class of non-convex objective functions can be that powerful to the point of "solving" — minimizing the error to the point of vanishing — complex problems that nature is dealing with in a rather reliable fashion. We feel like this is how a brain is doing optimization. We envision that a much larger class of functions possess such benign loss landscapes than the one here we have studied. Actually, we have isolated a function class that represents some of the most essential characteristics of a more general class of function as shown in eq. (4), so that we can show the principle underlying. That is, diverse yet cooperative subunits aggregating together to form a system can optimize an objective consistently. This larger class of function could be as important as the concept of convexity, and would play an important role in optimization. The goal of the paper is to lay one backbone of the theory of the NNs that make the principles underlying clear, instead of presenting the theory in its complete form in one go. Thus, essential properties of the function class are yet to be identified, and will be part of our future works.

The theorem also contributes to explaining why depth is crucial. The large N limits of the Hessian can be achieved by adding more layers, even though the number of neurons in each of the layers may be quite small compared with the overall number of neurons. The diversity of neurons is possible thanks to activation functions .

The phenomenon characterized by theorem 4.1.2 explains why there are two phases in the training dynamics of NNs, i.e., the rapid error decreasing phase when loss value is high, and the slow error decaying phase when the loss value is close to minima. As the error decreases, the expectation of the derivative of loss values in $R(T)$ will increasingly approach zero, thus the eigen-spectrum will concentrate around zero increasingly, making the landscape increasingly flat and the training process slowly. It probably also explains why we need to gradually decrease the step size in the gradient descent algorithms in practice. Very likely the flat regions are of a small volume compared with the overall parameter space. Thus, if the training goes conservatively, and inches towards the global minima, the risk will gradually decrease. But if we give a powerful kick to the training that induces a large shift in the parameter space, it may kick the current parameter out of the flat region that can inch toward the global minima, like kicking a ball from a valley to another mountain in the hyperspace, thus making the training starts all over again to find a valley to decrease the risk. A further characterization of the landscape goes beyond infinitesimal local regions may rigorously prove the conjecture. It even poses the possibility to move across the flat region rather swiftly, as long as we figure out how to stay in the valley as we stride big.

References

Alt, J., Erdos, L., and Krüger, T. The Dyson equation with linear self-energy: spectral bands, edges and cusps. Technical report, 2018. URL http://arxiv.org/abs/1804.07752.

Ballard, A. J., Das, R., Martiniani, S., Mehta, D., Sagun, L., Stevenson, J. D., and Wales, D. J. Perspective: Energy Landscapes for Machine Learning. *Phys. Chem. Chem. Phys.*, 19(20):12585–12603, 2017.

Chizat, L. and Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *NIPS*, 2018.

Choromanska, A., Henaff, M., and Mathieu, M. The Loss Surfaces of Multilayer Networks. In *AISTATS*, 2015a.

Choromanska, A., LeCun, Y., and Ben Arous, G. Open Problem: The landscape of the loss surfaces of multilayer networks. In *COLT*, 2015b.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In *ICLR*, 2015.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, 2014.

Dunford, N. and Schwartz, J. *Linear operators. Part 1: General theory*. Pure and Applied Mathematics. Interscience Publishers, 1957.

Erdos, L., Kruger, T., and Schroder, D. Random Matrices with Slow Correlation Decay. Technical report, 2017. URL http://arxiv.org/abs/1705.10661.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Haagerup, U. and Thorbjørnsen, S. A new application of random matrices:. *Annals of Mathematics*, 162(2):711–775, 2005. ISSN 0003-486X. doi: 10.4007/annals.2005.162.711.

He, K., Zhang, X., Ren, S., and Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *ICCV*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Helton, J. W., Far, R. R., and Speicher, R. Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints. *International Mathematics Research Notices*, 2007:1–14, 2007. ISSN 10737928. doi: 10.1093/imrn/rnm086.

Ioffe, Sergey and Szegedy, C. Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015.

Jain, P. and Kar, P. Non-convex Optimization for Machine Learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017. ISSN 1935-8237. doi: 10.1561/2200000058.

Jeffrey Pennington, Y. B. Geometry of Neural Network Loss Surfaces via Random Matrix Theory. In *ICML*, 2017.

Kadanoff, L. *Statistical Physics: Statics, Dynamics and Renormalization*. Statistical Physics: Statics, Dynamics and Renormalization. World Scientific, 2000. ISBN 9789810237646.

Kawaguchi, K. Deep Learning without Poor Local Minima. In *NIPS*, 2016.

Kelly, J. E. Computing, cognition and the future of knowing. Technical report, 2015.

Klenke, A. *Probability Theory: A Comprehensive Course*. World Publishing Corporation, 2012. ISBN 9787510044113.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

Laurent, T. and von Brecht, J. Deep linear neural networks with arbitrary loss: All local minima are global. In *ICML*, 2018.

Li, S. Measure, Manifold, Learning, and Optimization: A Theory Of Neural Networks. Technical report, nov 2018. URL http://arxiv.org/abs/1811.12783.

Liang, S., Sun, R., Li, Y., and Srikant, R. Understanding the Loss Surface of Neural Networks for Binary Classification. In *ICML*, 2018.

Magnus, J. and Neudecker, H. *Matrix differential calculus with applications in statistics and econometrics: 3rd. ed.* John Wiley & Sons, Limited, 2007.

McCullagh, P. *Tensor methods in statistics*. Monographs on statistics and applied probability. Chapman and Hall, 1987. ISBN 9780412274800.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. In *ICML*, 2017.

Nguyen, Q. and Hein, M. Optimization Landscape and Expressivity of Deep CNNs. In *ICML*, 2018.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for Activation Functions. Technical report, oct 2017. URL http://arxiv.org/abs/1710.05941.

Salimans, T. and Kingma, D. P. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *NIPS*, 2016.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of overparameterized shallow neural networks. *IEEE Transactions on Information Theory*, 4(8), 2018. doi: 10.1109/TIT.2018.2854560.

Tao, T. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc., 2012. ISBN 9780821885079.

Wales, D., Saykally, R., of Cambridge, U., Zewail, A., and King, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge Molecular Science. Cambridge University Press, 2003. ISBN 9780521814157.

Wigner, E. P. Statistical properties of real symmetric matrices with many dimensions.pdf. In *Canadian Mathematical Congress Proceedings*, 1957.

Yun, C., Sra, S., and Jadbabaie, A. Global optimality conditions for deep neural networks. In *ICLR*, 2018.

A. Appendices

A.1. Hessian of NN Is Inherently A Huge Random Matrix

As explained, to study the landscape of the loss function, we study the eigenvalue distribution of its Hessian $\boldsymbol{H}$ at the critical points. First, we derive the Hessian $\boldsymbol{H}$ of loss function of class $\mathcal{L}_0$ composed upon NNs with a single output. For a review of matrix calculus, the reader may refer to (Magnus & Neudecker, 2007).

The first partial differential of l defined at eq. (6) w.r.t. $\boldsymbol{W}_p$ is

$$\begin{aligned}\partial l(T\boldsymbol{x}, y) = l'(T\boldsymbol{x}, y)\boldsymbol{\alpha}^T &\prod_{j=p+1}^{L-1} (\mathrm{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}_j^T)\mathrm{dg}(\tilde{\boldsymbol{h}}'_p) \\ &\otimes \boldsymbol{x}^T \prod_{i=1}^{p-1} (\boldsymbol{W}_i \mathrm{dg}(\tilde{\boldsymbol{h}}_i))\partial \mathrm{vec}\boldsymbol{W}_p\end{aligned}$$

where $\otimes$ denotes Kronecker product. Note that for clarity of presentation, we use the partial differential the same way as differential is defined and used in (Magnus & Neudecker, 2007), i.e., ∂l is a number instead of an infinitely small quantities, though in the book, partial differential is not defined explicitly. $\tilde{\boldsymbol{h}}'$ is an abuse of notation for clarity and needs some explanation. $\{\tilde{\boldsymbol{h}}_i\}_{i=1,\dots,L-1}$ are realization of r.v.s defined at eq. (2). $\tilde{\boldsymbol{h}}_i$ is a scalar function, and denote it as $\tilde{\boldsymbol{h}}_i(a)$, where a is the computed input. When computing the partial differential w.r.t. $\boldsymbol{W}_p$, by the chain rule, the differential of $\tilde{\boldsymbol{h}}_i(a)$ w.r.t. $\boldsymbol{W}_p$ is $\partial \tilde{\boldsymbol{h}}_i(a)/\partial a$. To avoid introducing too much clutter, we denote $\tilde{\boldsymbol{h}}'_i$ as $\partial \tilde{\boldsymbol{h}}_i(a)/\partial a$.

Since $\boldsymbol{H}$ is symmetric, we only need to compute the block matrices by taking partial differential w.r.t. $\boldsymbol{W}_q$, where $q > p$ — taking partial differential w.r.t. $\boldsymbol{W}_p$ again gives zero matrix.

$$\begin{aligned}&\partial^2 l(T\boldsymbol{x}, y) \\ =&l'(T\boldsymbol{x}, y)[(\boldsymbol{\alpha}^T \prod_{k=q+1}^{L-1} (\mathrm{dg}(\tilde{\boldsymbol{h}}''_k)\boldsymbol{W}_k^T)\mathrm{dg}(\tilde{\boldsymbol{h}}''_q) \\ &\otimes \mathrm{dg}(\tilde{\boldsymbol{h}}'_p) \prod_{j=p+1}^{q-1} (\boldsymbol{W}_j \mathrm{dg}(\tilde{\boldsymbol{h}}'_j))\partial \mathrm{vec}\boldsymbol{W}_q)^T \\ &\otimes \boldsymbol{x}^T \prod_{i=1}^{p-1} (\boldsymbol{W}_i \mathrm{dg}(\tilde{\boldsymbol{h}}_i))]\partial \mathrm{vec}\boldsymbol{W}_p \\ =&l'(T\boldsymbol{x}, y) \\ &(\partial \mathrm{vec}\boldsymbol{W}_q)^T[\mathrm{dg}(\tilde{\boldsymbol{h}}''_q) \prod_{k=q+1}^{L-1} (\boldsymbol{W}_k \mathrm{dg}(\tilde{\boldsymbol{h}}''_k))\boldsymbol{\alpha} \\ &\otimes \prod_{j=p+1}^{q-1} (\mathrm{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}_j^T)\mathrm{dg}(\tilde{\boldsymbol{h}}'_p) \\ &\otimes \boldsymbol{x}^T \prod_{i=1}^{p-1} (\boldsymbol{W}_i \mathrm{dg}(\tilde{\boldsymbol{h}}_i))]\partial \mathrm{vec}\boldsymbol{W}_p\end{aligned}$$

where again $\tilde{\boldsymbol{h}}''$ is an abuse of notation, it is actually $\tilde{\boldsymbol{h}}' \odot \tilde{\boldsymbol{h}}'$, the hamadard product of partial differentials obtained by taking partial differential w.r.t. $\boldsymbol{W}_p$, and w.r.t. $\boldsymbol{W}_q$. The two partial differentials are the same because $\boldsymbol{a}$, the input the $\boldsymbol{h}$, is the same throughout. Thus, $\boldsymbol{H}$ is an ensemble of real symmetric random matrix with correlated entries. Denote

$$\begin{aligned}\boldsymbol{H}_{pq} =&l'(T\boldsymbol{x}, y)\mathrm{dg}(\tilde{\boldsymbol{h}}''_q) \prod_{k=q+1}^{L-1} (\boldsymbol{W}_k \mathrm{dg}(\tilde{\boldsymbol{h}}''_k))\boldsymbol{\alpha} \\ &\otimes \prod_{j=p+1}^{q-1} (\mathrm{dg}(\tilde{\boldsymbol{h}}'_j)\boldsymbol{W}_j^T)\mathrm{dg}(\tilde{\boldsymbol{h}}'_p) \\ &\otimes \boldsymbol{x}^T \prod_{i=1}^{p-1} (\boldsymbol{W}_i \mathrm{dg}(\tilde{\boldsymbol{h}}_i))\end{aligned} \tag{13}$$

We have the Hessian of l as

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{H}_{12}^T & \dots & \boldsymbol{H}_{1L}^T \\ \boldsymbol{H}_{12} & \boldsymbol{0} & \dots & \boldsymbol{H}_{2L}^T \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{H}_{1L} & \boldsymbol{H}_{2L} & \dots & \boldsymbol{0} \end{bmatrix} \tag{14}$$

A.2. Eigenvalue Distribution of Symmetry Random Matrix with Slow Correlation Decay

In this section, we show how to obtain the eigen-spectrum of $\boldsymbol{H}$ through random matrix theory (RMT) (Tao, 2012). RMT has been born out of the study on the nuclei of heavy atoms, where the spacings between lines in the spectrum of a heavy atom nucleus is postulated the same with spacings between eigenvalues of a random matrix (Wigner, 1957). In a certain way, it seems to be the backbone math of complex systems, where the collective behaviors of sophisticated subunits can be analyzed stochastically when deterministic or analytic analysis is intractable.

The following definitions can be found in (Tao, 2012) unless otherwise noted.

The eigen-spectrum is studied as empirical spectral distribution (ESD) in RMT, define as

Definition A.1 (Empirical Spectral Distribution)**.** *Given a $N \times N$ random matrix $\boldsymbol{H}$, its* **empirical spectral distribution** $\mu_{\boldsymbol{H}}$ *is*

$$\mu_{\boldsymbol{H}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}$$

where $\{\lambda_i\}_{i=1,\dots,N}$ are all eigenvalues of $\boldsymbol{H}$ and δ is the delta function.

Given a hermitian matrix $\boldsymbol{H}$, its ESD $\mu_{\boldsymbol{H}}(\lambda)$ can be studied via its resolvent $\boldsymbol{G}$.

Definition A.2 (Resolvent)**.** *Let $\boldsymbol{H}$ be a normal matrix, and $z \in \mathbb{H}$ a spectral parameter. The* **resolvent** $\boldsymbol{G}$ *of $\boldsymbol{H}$ at z is defined as*

$$\boldsymbol{G} = \boldsymbol{G}(z) = \frac{1}{\boldsymbol{H} - z}$$

where

$$\mathbb{H} := \{z \in \mathbb{C} : \Im z > 0\}$$

$\mathbb{C}$ denotes the complex field, and $\Im$ is the function gets imaginary part of a complex number z.

$\boldsymbol{G}$ compactly summarizes the spectral information of $\boldsymbol{H}$ around z, which is normally analyzed by *functional calculus* on operators, and defined through Cauchy integral formula on operators

Definition A.3 (Functions of Operators)**.**

$$f(T) = \frac{1}{2\pi i} \int_C \frac{f(\lambda)}{\lambda - T} d\lambda \qquad (15)$$

where f is an analytic scalar function and C is an appropriately chosen contour in $\mathbb{C}$.

The formula can be defined on a range of linear operators (Dunford & Schwartz, 1957) (Recall that a linear operator is a mapping whose domain and codomain are defined on the same field). Since the most complex case involved here will be a normal matrix, we stop at stating that the formula holds true when T is a normal matrix.

Resolvent $\boldsymbol{G}$ is related to eigen-spectrum of $\boldsymbol{H}$ through stieltjes transform of $\mu_{\boldsymbol{H}}(\lambda)$.

Definition A.4 (Stieltjes Transform)**.** *Let μ be a Borel probability measure on $\mathbb{R}$. Its* **Stieltjes transform** *at a spectral parameter $z \in \mathbb{H}$ is defined as*

$$m_\mu(z) = \int_{\mathbb{R}} \frac{d\mu(x)}{x - z}$$

With some efforts, it can be seen that the normalized trace of $\boldsymbol{G}$ is stieltjes transform of eigen-spectrum of $\boldsymbol{H}$

$$m_{\mu_{\boldsymbol{H}}}(z) = \frac{1}{N} \mathrm{tr}\, G$$

For a proof, the reader may refer to proposition 2.1 in (Alt et al., 2018). $\mu_{\boldsymbol{H}}$ can be recovered from $m_{\mu_{\boldsymbol{H}}}$ through the inverse formula of Stieltjes-Perron.

Lemma A.1 (Inverse Stieljies Transform)**.** *Suppose that μ is a probability measure on $\mathbb{R}$ and let m_μ be its Stieltjes transform. Then for any $a < b$, we have*

$$\mu((a,b)) + \frac{1}{2}[\mu(\{a\}) + \mu(\{b\})] = \lim_{\Im z \to 0} \frac{1}{\pi} \int_b^a \Im m_\mu(z) d\Re z$$

where $\Re$ is the function that gets the real part of z. The proof can be found at (Tao, 2012) p. 144.

Consequently, the problem converts to obtain $\boldsymbol{G}$ if we want to obtain $\mu_{\boldsymbol{H}}$. A recent advance in the RMT community has enabled the analysis of ESD of symmetric random matrix with correlation (Erdos et al., 2017) from the perspective of mean field theory (Kadanoff, 2000), which we debrief in the following.

The resolvent $\boldsymbol{G}$ holds an identity by definition

$$\boldsymbol{H}\boldsymbol{G} = \boldsymbol{I} + z\boldsymbol{G} \qquad (16)$$

Note that in the above equation $\boldsymbol{G}$ is a function $\boldsymbol{G}(\boldsymbol{H})$ of $\boldsymbol{H}$. When the average fluctuation of entries of $\boldsymbol{G}$ w.r.t. to its mean is small as N grows large, eq. (16) can be turned into a solvable equation regarding $\boldsymbol{G}$ instead of merely a definition. Formally, it is achieved by taking the expectation of eq. (16)

$$\mathbb{E}[\boldsymbol{H}\boldsymbol{G}] = \boldsymbol{I} + z\mathbb{E}[\boldsymbol{G}] \qquad (17)$$

When fluctuation of moments beyond the second order are negligible, we can obtain a class of random matrices whose ESD can be obtained by solving a truncated cumulant expansion of eq. (17). With the above approach, using sophisticated multivariate cumulant expansion, (Erdos et al., 2017) proves $\boldsymbol{G}$ can be obtained as the unique solution to *Matrix Dyson Equation (MDE)* below

$$\boldsymbol{I} + (z - \boldsymbol{A} + \mathcal{S}[\boldsymbol{G}])\boldsymbol{G} = \boldsymbol{0}, \Im\boldsymbol{G} \succ \boldsymbol{0}, \Im z > 0 \qquad (18)$$

where $\Im\boldsymbol{G} \succ 0$ means $\Im\boldsymbol{G}$ is positive definite,

$$\boldsymbol{A} := \mathbb{E}[\boldsymbol{H}], \frac{1}{\sqrt{N}}\boldsymbol{W} := \boldsymbol{H} - \boldsymbol{A}, \mathcal{S}[\boldsymbol{R}] := \frac{1}{N}\mathbb{E}_{\boldsymbol{W}}[\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}] \qquad (19)$$

$\mathcal{S}$ is a linear map on the space of $N \times N$ matrices and $\boldsymbol{W}$ is a random matrix with zero expectation. The expectation is taken w.r.t. $\boldsymbol{W}$, while taking $\boldsymbol{R}$ as a deterministic matrix.

We describe their results formally in the following, which begins with some more definitions, adopted from (Erdos et al., 2017).

Definition A.5 (Cumulant)**.** **Cumulants** *of $\kappa_{\boldsymbol{m}}$ of a random vector $\boldsymbol{w} = (w_1, \dots, w_n)$ are defined as the coefficients of log-characteristic function*

$$\log \mathbb{E}e^{i\boldsymbol{t}^T\boldsymbol{w}} = \sum_{\boldsymbol{m}} \kappa_{\boldsymbol{m}} \frac{i\boldsymbol{t}^{\boldsymbol{m}}}{\boldsymbol{m}!}$$

where $\sum_{\boldsymbol{m}}$ is the sum over all n-dimensional multi-indices $\boldsymbol{m} = (m_1, \dots, m_n)$.

To recall, a multi-indices is

Definition A.6 (Multi-index)**.** *a n-dimensional multi-index is an n-tuple*

$$\boldsymbol{m} = (m_1, \dots, m_n)$$

of non-negative integers. Note that $|\boldsymbol{m}| = \sum_{i=1}^{n} m_i$, and $\boldsymbol{m}! = \prod_{i=1}^{n} m_i!$.

Similar with the more familiar concept *moment*, cumulant is also a measure of statistic properties of r.v.s. Particularly, the k-order cumulant κ characterizes the k-way correlation of a set of r.v.s. The key of insight of the paper is to identify the condition where a matrix entry $w_\alpha, \alpha \in \mathbb{I}$ is only strongly correlated with a minority of $\boldsymbol{W}_{\mathbb{I}\setminus\{\alpha\}}$, and higher order cumulants tend to be weak and not influential in large N limit. Thus, a proper formulation of the correlation strength is needed, and is defined as the cumulant norms on entries of $\boldsymbol{W}$ in the following. Given k entries $\boldsymbol{W}_{\boldsymbol{\alpha}}$ at $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1,\dots,k}, \alpha_i \in \mathbb{I}$ of matrix $\boldsymbol{W}$, where duplication is allowed, denote $\kappa(\alpha_1, \dots, \alpha_k) = \kappa(w_{\alpha_1}, \dots, w_{\alpha_k})$.

Definition A.7 (Cumulant Norms)**.**

$$\begin{aligned}
|||\kappa||| &:= |||\kappa|||_{\leq R} := \max_{2\leq k\leq R} |||\kappa|||_k, \\
|||\kappa|||_k &:= |||\kappa|||_k^{av} + |||\kappa|||_k^{iso} \\
|||\kappa|||_2^{av} &:= ||\,|\kappa(*,*)|\,||, \qquad (20a)\\
|||\kappa|||_k^{av} &:= N^{-2} \sum_{\alpha_1,\dots,\alpha_k} |\kappa(\alpha_1,\dots,\alpha_k)|, k \geq 4 \\
|||\kappa|||_3^{av} &:= ||\sum_{\alpha_1} |\kappa(\alpha_1,*,*)|\,||+ \\
&\quad \inf_{\kappa=\kappa_{dd}+\kappa_{dc}+\kappa_{cd}+\kappa_{cc}} (|||\kappa_{dd}|||_{dd} + |||\kappa_{dc}|||_{dc} \\
&\quad + |||\kappa_{cd}|||_{cd} + |||\kappa_{cc}|||_{cc}) \qquad (20b)\\
|||\kappa|||_{cc} &= |||\kappa|||_{dd} \\
&:= N^{-1} \sqrt{\sum_{b_2,a_3} (\sum_{a_2,b_3} \sum_{\alpha_1} |\kappa(\alpha_1, a_2b_2, a_3b_3)|)^2} \\
|||\kappa|||_{cd} &:= N^{-1} \sqrt{\sum_{b_3,a_1} (\sum_{a_3,b_1} \sum_{\alpha_2} |\kappa(a_1b_1, \alpha_2, a_3b_3)|)^2}, \\
|||\kappa|||_{dc} &:= N^{-1} \sqrt{\sum_{b_1,a_2} (\sum_{a_1,b_2} \sum_{\alpha_3} |\kappa(a_1b_1, a_2b_2, \alpha_3)|)^2} \\
|||\kappa|||_2^{iso} &:= \inf_{\kappa=\kappa_d+\kappa_c} (|||\kappa_d|||_d + |||\kappa_c|||_c), \\
|||\kappa_d|||_d &:= \sup_{||\boldsymbol{x}||\leq \mathbf{1}} ||||\,\kappa(\boldsymbol{x}*, \cdot *)||\,||, \\
|||\kappa|||_c &:= \sup_{||\boldsymbol{x}||\leq \mathbf{1}} ||\,||\kappa(\boldsymbol{x}*, *\cdot)||\,|| \qquad (20c)\\
|||\kappa|||_k^{iso} &:= ||\sum_{\alpha_1,\dots,\alpha_{k-2}} |\kappa(\alpha_1,\dots,\alpha_{k-2},*,*)|\,||, k \geq 3
\end{aligned}$$

where in eq. (20b), the infimum is taken over all decomposition of κ in four symmetric functions $\kappa_{dd}, \kappa_{dc}, \kappa_{cd}, \kappa_{cc}$; in eq. (20c) the infimum is taken over all decomposition of κ into the sum of symmetric κ_c and κ_d. The norms defined in eq. (20a) and eq. (20c) need some explanation on the notation. If, in place of an index $\alpha \in \mathbb{J}$, we write a dot $(\cdot)$ in a scalar quantity then we consider the quantity as a vector indexed by the coordinate at the place of the dot. For example, $\kappa(a_1\cdot, a_2b_2)$ is a vector, the i-th entry of which is $\kappa(a_1 i, a_2b_2)$ and therefore the inner norms in eq. (20a) indicate vector norms. In contrast, the outer norms indicate the operator norm of the matrix indexed by star $(*)$. More specifically, $||A(*,*)||$ refers to the operator norm of the matrix with matrix elements A_{ij}. Thus $||\,||\kappa(\boldsymbol{x}*, *\cdot)||\,||$ is the operator norm $||A||$ of the matrix A with matrix elements $A_{ij} = ||\kappa(\boldsymbol{x}i, j\cdot)||$. $\kappa(\boldsymbol{x}b_1, a_2b_2)$ denotes $\sum_{a_1} \kappa(a_1b_1, a_2b_2)x_{a_1}$, where $\boldsymbol{x}$ is a vector.

We do not want to explain the cumulant norms beyond what has been said, considering it is too technically involved and rather a distraction. For interested readers, we suggest reading the paper (Erdos et al., 2017). Equipped with the cumulant norms, we would have the assumptions stated in assumption 4.1 4.2 that make MDE valid.

Remark. *In (Erdos et al., 2017), the functions $f, g_1, \dots, g_q$ in assumption 4.2 are assumed to be functions without any qualifiers. We change it to analytic functions for further usage. In the proof of theorem A.1, the functions are only required to be analytic, thus even if the assumptions are changed, the conclusion still holds.*

The diversity assumption requires that a matrix entry w_α only couples with a minority of the overall entries, and for the rest of the entries, the coupling strength does not exceed a certain value $N^{-3q}||f||_{q+1} \prod_{j=1}^{q} ||g_j||_{q+1}$ characterized by cumulants. For example, suppose $q = 1$, given a entry α_1, the assumption essentially states that the entries in the coupling set $\mathcal{N}(\alpha_1)$ is not strongly coupled with the resting of the population $\boldsymbol{W}_{\mathbb{I}\setminus\mathcal{N}_{n_1+1}(\alpha_1)}$. The explanation goes similar as q grows, of which the coupling strengh is characterized by higher order cumulants. While boundedness assumptions 1)2)3)4) states the expectation of $\boldsymbol{H}$ is bounded, moments are finite, cumulants are bounded for the entries that do strongly couples, and $\mathcal{S}[\boldsymbol{W}]$ is bounded in the sense of eigenvalues.

When the assumptions satisfies, we have the resolvent $\boldsymbol{G}$ of a random matrix $\boldsymbol{H}$ close to the solution to the MDE probabilistically with some regular properties as the following, adopted in an informal style to ease reading from (Erdos et al., 2017) theorem 2.2, (Helton et al., 2007) theorem 2.1, and (Alt et al., 2018) theorem 2.5.

Theorem A.1. *Let $\boldsymbol{M}$ be the solution to the Matrix Dyson Equation eq. (18), and ρ the density function (measure) recovered from normalized trace $\frac{1}{N}tr\,\boldsymbol{M}$ through Stieljies* *inverse lemma A.1. We have*

1. *The MDE has a unique solution $\boldsymbol{M} = \boldsymbol{M}(z)$ for all $z \in \mathbb{H}$.*
2. *suppρ is a finite union of closed intervals with nonempty interior. Moreover, the nonempty interiors are called the* **bulk** *of the eigenvalue density function ρ.*
3. *The resolvent $\boldsymbol{G}$ of $\boldsymbol{H}$ converges to $\boldsymbol{M}$ as $N \to \infty$.*

Remark. *theorem A.1.3 is a probably-approximately-correct type result, where the error depends on N. We do not present the exact error bound here, for that it is rather complicated, and does not help understanding — since we are not working on finer behaviors of NNs with a particular size, and do not need such a fine granularity characterization yet. We refer interested readers to (Erdos et al., 2017) theorem 2.1, 2.2, where the exact error bounds are present.*

A.3. NN Loss Landscape: All Local Minima Are Global Minima with Zero Losses

We have obtained the operator equation to describe the eigen-spectrum of the Hessian of NNs and explained its assumptions. With one further assumption, we show in this section that for NNs with objective function belonging to the function class $\mathcal{L}_0$, all local minima are global minima with zero loss values.

We outline the strategy first. Since MDE is a nonlinear operator equation, it is not possible to obtain a close form analytic solution. The only way to get its solution is an iterative algorithm (Helton et al., 2007), which is not an easy task given the millions of parameters of a NN — remembering that we are dealing with large N limit — though it can serve as an exploratory tool. However, we do are able to get qualitative results by directly analyzing the equation. Our goal is to show all critical points are saddle points, except for the ones has zero loss values, which are global minima. To prove it, we prove that at the points where $R(T) \neq 0$, the eigen-spectrum $\mu_{\boldsymbol{H}}$ of the Hessian $\boldsymbol{H}$ is symmetric w.r.t. the y-axis, which implies that as long as non-zero eigenvalues exist, half of them will be negative. To prove it, we prove the stieltjes transform $m_{\mu_{\boldsymbol{H}}}(z)$ of $\mu_{\boldsymbol{H}}$ satisfies $\Im m_{\mu_{\boldsymbol{H}}}(-z^*) = \Im m_{\mu_{\boldsymbol{H}}}(z)$, where z^* denote the complex conjugate of z. In the following, we present the proof formally.

Lemma A.2. *Let $\boldsymbol{M}(z), \boldsymbol{M}'(-z^*)$ be the unique solution to the MDE at spectral parameter $z, -z^*$ defined at eq. (18) respectively, and $\boldsymbol{A} = \boldsymbol{0}$. We have*

$$\boldsymbol{M}' = -\boldsymbol{M}^*$$

where $*$ means taking conjugate transpose.

Proof. First, we rewrite the MDE. Note that $\mathcal{S}[\boldsymbol{G}]$ is positivity preserving, i.e., $\forall \boldsymbol{G} \succ \boldsymbol{0}, \mathcal{S}[\boldsymbol{G}] \succ \boldsymbol{0}$ by assumption 4.1 4). In addition, we have $\Im z > 0$, thus $\Im(z + \mathcal{S}[\boldsymbol{G}]) \succ \boldsymbol{0}$. Then, by (Haagerup & Thorbjørnsen, 2005) lemma 3.2, we have $z + \mathcal{S}[\boldsymbol{G}] \succ \boldsymbol{0}$, so it is invertable. Thus, we can rewrite the MDE into the following form

$$\boldsymbol{G} = -(z + \mathcal{S}[\boldsymbol{G}])^{-1} \tag{21}$$

Suppose $\boldsymbol{M}$ is a solution to the MDE at spectral parameter z. The key to the proof is the fact that $\mathcal{S}[\boldsymbol{G}]$ is linear and commutes with taking conjugate, thus by replacing $\boldsymbol{M}$ with $-\boldsymbol{M}^*$, and z with $-z^*$, we would get the same equation. We show it formally in the following.

First, note that $\mathcal{S}[\boldsymbol{M}]$ is a linear map of $\boldsymbol{M}$, so the we have

$$\mathcal{S}[-\boldsymbol{M}] = -\mathcal{S}[\boldsymbol{M}]$$

Also, $\mathcal{S}[\boldsymbol{M}]$ commutes with $*$, for the fact

$$\mathcal{S}[\boldsymbol{M}^*] = \mathbb{E}[\boldsymbol{W}\boldsymbol{M}^*\boldsymbol{W}] = \mathbb{E}[(\boldsymbol{W}\boldsymbol{M}\boldsymbol{W})^*] = \mathbb{E}[\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}]^* = \mathcal{S}[\boldsymbol{M}]^*$$

Furthermore, we $*$ is commute with taking inverse, for the fact

$$\begin{aligned} & \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I} \\ \Longrightarrow \quad & (\boldsymbol{A}\boldsymbol{A}^{-1})^* = \boldsymbol{I} \\ \Longrightarrow \quad & \boldsymbol{A}^{-1*}\boldsymbol{A}^* = \boldsymbol{I} \\ \Longrightarrow \quad & \boldsymbol{A}^{-1*} = \boldsymbol{A}^{*-1} \end{aligned}$$

With the commutativity results, we do the proof. The solution $\boldsymbol{M}$ satisfies the equation

$$\boldsymbol{M} = -(z + \mathcal{S}[\boldsymbol{M}])^{-1}$$

Replacing $\boldsymbol{M}$ with $-\boldsymbol{M}^*$, z with $-z^*$, we have

$$\begin{aligned} & -\boldsymbol{M}^* = -(-z^* + \mathcal{S}[-\boldsymbol{M}^*])^{-1} \\ \Longrightarrow \quad & \boldsymbol{M}^* = -(z^* + \mathcal{S}[\boldsymbol{M}^*])^{-1} \\ \Longrightarrow \quad & \boldsymbol{M}^* = -(z^* + \mathcal{S}[\boldsymbol{M}]^*)^{-1} \\ \Longrightarrow \quad & \boldsymbol{M}^* = -(z + \mathcal{S}[\boldsymbol{M}])^{-1*} \\ \Longrightarrow \quad & \boldsymbol{M} = -(z + \mathcal{S}[\boldsymbol{M}])^{-1} \end{aligned}$$

After the replacement, we actually get the same equation. Thus, $-\boldsymbol{M}^*, -z^*$ also satisfy eq. (21). Since the pair also satisfies the constrains $\Im \boldsymbol{M} \succ 0, \Im z > 0$, and by theorem A.1, the solution is unique, we proved the solution $\boldsymbol{M}'$ at the spectral parameter $-z^*$ is $-\boldsymbol{M}^*$.

□

Theorem A.2. *Let $\boldsymbol{H}$ be a real symmetric random matrix satisfies assumptions 4.1 4.2, in addition to the assumption* *that $\boldsymbol{A} = \boldsymbol{0}$. Let the ESD of $\boldsymbol{H}$ be $\mu_{\boldsymbol{H}}$. Then, $\mu_{\boldsymbol{H}}$ is symmetric w.r.t. to y-axis. In other words, half of the non-zero eigenvalues are negative. Furthermore, non-zero eigenvalues always exist, implying $\boldsymbol{H}$ will always have negative eigenvalues.*

Proof. By theorem A.1, the resolvent $\boldsymbol{G}$ of $\boldsymbol{H}$ is given by the the unique solution to eq. (18) at spectral parameter z. Let the solution to eq. (18) at spectral parameter $z, -z^*$ be $\boldsymbol{M}, \boldsymbol{M}'$, By lemma A.2, we have the solutions satisfies

$$\boldsymbol{M}' = -\boldsymbol{M}^*$$

By A.1, the ESD of $\boldsymbol{H}$ at $\Re z$ is given at

$$\mu_{\boldsymbol{H}}(\Re z) = \lim_{\Im z \to 0} \frac{1}{\pi} \Im m_{\mu_{\boldsymbol{H}}}(z)$$

Since $m_{\mu_{\boldsymbol{H}}}(z) = \frac{1}{N} \text{tr } M$, we have

$$\mu_{\boldsymbol{H}}(\Re z) = \lim_{\Im z \to 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M$$

Similarly,

$$\mu_{\boldsymbol{H}}(\Re(-z^*)) = \lim_{\Im(-z^*) \to 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M'$$

Note that

$$\begin{aligned} & \mu_{\boldsymbol{H}}(\Re(-z^*)) = \lim_{\Im(-z^*) \to 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M' \\ \implies \quad & \mu_{\boldsymbol{H}}(\Re(-z^*)) = \lim_{\Im(-z^*) \to 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } (-M^*) \\ \implies \quad & \mu_{\boldsymbol{H}}(-\Re z) = \lim_{\Im z \to 0} \frac{1}{\pi} \frac{1}{N} \Im \text{tr } M \end{aligned}$$

Thus, $\mu_{\boldsymbol{H}}(\lambda), \lambda \in \mathbb{R}$ is symmetric w.r.t. y-axis. It follows that for all non-zero eigenvalues, half of them are negative.

By theorem A.1 2, there are always bulks in supp$\mu_{\boldsymbol{H}}$, thus there are always non-zero eigenvalues. Since half of the non-zero eigenvalues are negative, it follows $\boldsymbol{H}$ always has negative eigenvalues. □

Proof of theorem 4.1. First, we prove part 1 of the theorem. The majority of the proof of part 1 have been dispersed earlier in the paper. What the proof here does mostly is to collect them into one piece.

The Hessian $\boldsymbol{H}$ of the risk function eq. (1), can be decomposed into a summation of Hessians of loss functions of each training sample, which is described in eq. (11). For each Hessian in the decomposition, it is computed in eq. (14), and it has been shown that $\boldsymbol{H}$ is a random matrix in appendix A.1.

The analysis of the random matrix $\boldsymbol{H}$ needs to break down into two cases: 1) for all training samples, at least one sample (x, y) has non-zero loss value; 2) and all training samples are classified properly with zero loss values.

We first analyze case 1), since the loss l belongs to function class $\mathcal{L}_0$, l is convex and is valued zero at its minimum. When $l(x, y) \neq 0$, we have $l'(x, y) \neq 0$, thus $\boldsymbol{H}$ is a random matrix — not a zero matrix. The analysis of this type of random matrix is undertaken in appendix A.2. For a NN, the assumptions 4.1 4.2 can be satisfied, and the eigenspectrum $\mu_{\boldsymbol{H}}$ of $\boldsymbol{H}$ is given by the MDE defined at eq. (18). The practicality and its potential to guide real world NN optimization is discussed in section 4.1.

By theorem A.2, $\mu_{\boldsymbol{H}}$ is symmetric w.r.t. y-axis, and half of its non-zero eigenvalues are negative. Thus, for all critical points of $R(T)$, its will have half of its non-zero eigenvalues negative. It implies all critical points are saddle points.

Now we turn to the case 2). In this case, all training samples are properly classified with zero loss value. Considering the lower bound of l is zero, we have reached the global minima. Also, since all critical points in case 1) are saddle points, local minima can only be reached in case 2), implying all local minima are global minima. Thus, the first part of the theorem is proved.

Now we prove part 2 of the theorem.

Note that the minima is reached for the fact that we have reached the situation where the Hessian $\boldsymbol{H}$ has degenerated into a zero matrix. Thus, each local minimum is not a critical point, but an infimum, where in a local region around the infimum in the parameter space, all the eigenvalues are increasingly close to zero as the parameters of the NN approach the parameters at the infimum. We show it formally in the following.

Writing a block $\boldsymbol{H}_{pq}$ (defined at eq. (13)) in the Hessian $\boldsymbol{H}_i$ of one sample (defined at eq. (14)) in the form of

$$\boldsymbol{H}_{pq} = l'(T\boldsymbol{x}, y)\tilde{\boldsymbol{H}}_{pq}$$

where i is the index of the training samples, defined at eq. (1). Then, putting together $\tilde{\boldsymbol{H}}_{pq}$ together to form $\tilde{\boldsymbol{H}}_i$, $\boldsymbol{H}_i$ is rewritten in the form of

$$\boldsymbol{H}_i = l'(T\boldsymbol{x}_i, y_i)\tilde{\boldsymbol{H}}_i$$

Then the Hessian $\boldsymbol{H}$ (defined at eq. (11)) of the risk function defined eq. (1) can be rewritten in the form of

$$\boldsymbol{H} = \frac{1}{m} \sum_{i=1}^{m} l'(T\boldsymbol{x}_i, y_i)\tilde{\boldsymbol{H}}_i$$

Taking the operator norm on the both sides

$$||\boldsymbol{H}|| = ||\frac{1}{m} \sum_{i=1}^{m} l'(T\boldsymbol{x}_i, y_i)\tilde{\boldsymbol{H}}_i|| \leq \frac{1}{m} \sum_{i=1}^{m} |l'(T\boldsymbol{x}_i, y_i)| \, ||\tilde{\boldsymbol{H}}_i||$$

Denote $\max_i\{||\tilde{\boldsymbol{H}}_i||\}$ as λ_0, we have

$$\begin{aligned}||\boldsymbol{H}|| &\leq \frac{1}{m}\sum_{i=1}^{m}|l'(T\boldsymbol{x}_i, y_i)|\lambda_0 \\ &= \mathbb{E}_m[l'(TX, Y)]\lambda_0\end{aligned}$$

The above inequality shows that, as the risk decreases, more and more samples will have zero loss value, consequently $l' = 0$, thus $\mathbb{E}_m[l']$ will be increasingly small, thus the operator norm of $\boldsymbol{H}$. At the minima where all $l' = 0$, the Hessian degenerates to a zero matrix. □

Remark. *It is not necessary for the assumption $\boldsymbol{A} = \boldsymbol{0}$ to be held for the theorem to hold, or more specifically, for $\boldsymbol{H}$ to have negative eigenvalues at its critical points. By proposition 2.1 in (Alt et al., 2018)*

$$\mathit{supp}\mu_{\boldsymbol{H}} \subset \mathit{Spec}\boldsymbol{A} + [-2||\mathcal{S}||^{1/2}, 2||\mathcal{S}||^{1/2}]$$

where Spec$\boldsymbol{A}$ denotes the support of the spectrum of the expectation matrix $\boldsymbol{A}$ and $||\mathcal{S}||$ denotes the norm induced by the operator norm. Thus, it is possible for the spectrum $\mu_{\boldsymbol{H}}$ of $\boldsymbol{H}$ to lie at the left side of the y-axis, as long as the spectrum of $\boldsymbol{A}$ is not too way off from the origin. However, existing characterizations on supp$\mu_{\boldsymbol{H}}$ based on bound are too inexact to make sure the existence of support on the left of the y-axis. To get rid of the zero expectation assumption, more works are needed to obtain a better characterization, and could be a direction for future work.