On the Flatness of Loss Surface for Two-layered ReLU Networks

Jiezhang Cao† SECAOJIEZHANG@MAIL.SCUT.EDU.CN
School of Software Eigineering, South China University of Technology

Qingyao Wu† QYW@SCUT.EDU.CN
School of Software Eigineering, South China University of Technology

Yuguang Yan YAN.YUGUANG@MAIL.SCUT.EDU.CN
School of Software Eigineering, South China University of Technology

Li Wang LI.WANG@UTA.EDU
Department of Mathematics, University of Texas at Arlington

Mingkui Tan* MINGKUITAN@SCUT.EDU.CN
School of Software Eigineering, South China University of Technology

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

Deep learning has achieved unprecedented practical success in many applications. Despite its empirical success, however, the theoretical understanding of deep neural networks still remains a major open problem. In this paper, we explore properties of two-layered ReLU networks. For simplicity, we assume that the optimal model parameters (also called ground-truth parameters) are known. We then assume that a network receives Gaussian input and is trained by minimizing the expected squared loss between the prediction function of the network and a target function. To conduct the analysis, we propose a normal equation for critical points, and study the invariances under three kinds of transformations, namely, scale transformation, rotation transformation and perturbation transformation. We prove that these transformations can keep the loss of a critical point invariant, thus can incur flat regions. Consequently, how to escape from flat regions is vital in training neural networks.
Keywords: Two-layered ReLU network, flatness, critical points, loss surface

1. Introduction

Deep learning plays a remarkable role with its state-of-the-art performance in many fields, including computer vision (Krizhevsky et al., 2012; He et al., 2016), natural language processing (Sarikaya et al., 2014), speech recognition (Hinton et al., 2012), reinforcement learning (Mnih et al., 2015; Silver et al., 2016), etc. Despite its empirical success, theoretical understanding of neural networks, however, is still very limited.

The loss function of a neural network is highly non-convex, leading to numerous obstacles for training with gradient-based methods, with fear of proliferation of saddle points and

† The first two authors contributed to this work equally.
* Correspondence should be addressed to M. Tan.

existence of local minima. Recently, Kawaguchi (2016) proves the nonexistence of poor local minima for deep nonlinear neural networks under strong assumptions, that is, the activation is independent of the input. Dauphin et al. (2014) argue that the difficulty in training neural networks mainly comes from saddle points rather than poor local minima, and such saddle points are surrounded by high-error plateaus, which may slow down the learning procedure. Interestingly, Shamir (2016) states that the reason of difficulty in training is not plethora of critical points but flat loss surfaces. Two natural questions thus arise: does the flatness exist and what kind of reasons causes such flatness?

Answering these questions, however, is difficult. In the past few years, analysis on flatness of loss surface for high-dimensional and non-convex problems has attracted much attention. For example, Hochreiter and Schmidhuber (1997) propose a method to find flat minima whose neighbors have similar errors. Dauphin et al. (2014) argue that saddle points are surrounded by plateaus, which slow down the learning procedure. Shamir (2016) discusses a special case where a loss function may exhibit flatness nearly everywhere. Freeman and Bruna (2016) show that the solution space of a two-layered ReLU network is connected, which shapes the flat landscape. Lipton (2016) experimentally demonstrates that the algorithm does not converge to critical points, instead a flat region of weight space. On the other hand, some researchers have noticed that neural networks do have many local minima, while the performances of them are very similar. These results raise a theoretical question that why the loss remains approximately constant.

In this paper, we seek to analyze the flat phenomenon from different perspectives and aim to study how the critical points form flat regions on the loss surface. Specifically, we focus on genreral two-layered ReLU networks with a linear output layer. Moreover, the network is trained by minimizing the expected squared loss between the prediction function and a target function with known optimal parameters (also called ground-truth parameters) over the input distribution. Combination of assumptions on both the input distribution and the target function is sufficient to guarantee computationally tractable learning (Shamir, 2016). Based on this, we consider zero-mean Gaussian inputs and a target function that is related to a two-layered ReLU network with fixed ground-truth weights.

In this paper, we provide theoretical evidences that the flatness of loss surface does indeed exist. Our main contributions are summarized as follows:

- Tian (2017) studies critical points in a special case where the last layer contains fixed weights of value 1, which makes the problem easier to handle. Unlike Tian (2017)'s work, we study the flatness of loss surface for general two-layered ReLU networks without the fixed-weights setting for the last layer.

- We provide a normal equation for the loss function hao which helps to understand the behaviors of critical points and the loss function.

- Based on the normal equation, we consider three kinds of transformations and explore the invariance of the loss function under these transformations.

The rest of the paper is organized as follows. Section 2 introduces the problem definition. Section 3 presents main results under different transformations. In Section 4, we provide the detailed proofs corresponding to the main results. Section 5 concludes this work and opens many future directions.

2. Problem Definition

2.1. Notations

We denote by $[m]$ the set $\{1, \ldots, m\}$. Let $\text{vec}(\mathbf{M}) \in \mathbb{R}^{mn}$ be the vectorization of a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, i.e., $\text{vec}(\mathbf{M}) = \left[\mathbf{M}_{,1}^\mathsf{T}, \mathbf{M}_{,2}^\mathsf{T}, \ldots, \mathbf{M}_{,n}^\mathsf{T}\right]^\mathsf{T}$ where $\mathbf{M}_{,j}$ is the j-th colunm vector of $\mathbf{M}$. Let $\mathbf{M} \otimes \mathbf{M}'$ be the Kronecker product of $\mathbf{M}$ and $\mathbf{M}'$. Let $\mathcal{D}_{\text{vec}(\mathbf{W}_k)} f(\cdot) = \frac{\partial f(\cdot)}{\partial \text{vec}(\mathbf{W}_k)}$ be the partial derivative of f with respect to $\text{vec}(\mathbf{W}_k)$ in the numerator layout.

2.2. Two-layered ReLU Networks

Let $(\mathbf{X}, \mathbf{Y})$ be the training data set, where $\mathbf{X} \in \mathbb{R}^{N \times d_x}$ denotes the input data matrix, and $\mathbf{Y} \in \mathbb{R}^{N \times d_y}$ denotes the ground-truth output. Without loss of generality, we consider a two-layered neural network with ReLU activation function, let $\sigma(\mathbf{A}) = \max(\mathbf{0}, \mathbf{A}) : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ be the element-wise rectified linear unit (ReLU) function, and let $\mathbf{W} = \{\mathbf{W}_1 \in \mathbb{R}^{d_x \times d_1}, \mathbf{W}_2 \in \mathbb{R}^{d_y \times d_1}\}$ be the weights of the network, where $\mathbf{W}_1 = \left[\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_{d_1}^{(1)}\right] \in \mathbb{R}^{d_x \times d_1}$ and $\mathbf{W}_2 = \left[\mathbf{w}_1^{(2)}, \mathbf{w}_2^{(2)}, \ldots, \mathbf{w}_{d_1}^{(2)}\right] \in \mathbb{R}^{d_y \times d_1}$ are the weights of the first and second layers, respectively. For convenience, we exclude the bias terms in our definition, and the output $g(\mathbf{W}, \mathbf{X}) \in \mathbb{R}^{N \times d_y}$ of a two-layered ReLU network can be written as

$$g(\mathbf{W}, \mathbf{X}) = \sigma\left(\mathbf{X}\mathbf{W}_1\right)\mathbf{W}_2^\mathsf{T} = \sum_{i=1}^{d_1} \sigma\left(\mathbf{X}\mathbf{w}_i^{(1)}\right)\mathbf{w}_i^{(2)\mathsf{T}}, \tag{1}$$

where $\mathbf{w}_i^{(1)}$ and $\mathbf{w}_i^{(2)}$ are the i-th column vector of $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively.

We define $\mathbf{D}_i = \mathbf{D}(\mathbf{w}_i) = \sigma\left(\mathbf{X}\mathbf{w}_i\right) = \text{diag}\left(\text{sgn}\left(\mathbf{X}\mathbf{w}_i > 0\right)\right)$ as an $N \times N$ diagonal matrix, where $i \in [d_1]$. Here, the sign function $\text{sgn}(\cdot)$ is defined as $\text{sgn}(a) = 1$ if $a > 0$, otherwise, $\text{sgn}(a) = 0$. The l-th diagonal element of $\mathbf{D}_i$ is 1 or 0, which indicates whether the l-th neuron is activated or not. Based on the above definitions, we can rewrite the prediction function of the two-layered ReLU networks as:

$$g(\mathbf{W}, \mathbf{X}) = \sum_{i=1}^{d_1} \mathbf{D}_i \mathbf{X} \mathbf{w}_i^{(1)} \mathbf{w}_i^{(2)\mathsf{T}}. \tag{2}$$

Now we are ready to define the loss function between learned weights and the ground-truth weights. We assume the optimal weights $\mathbf{W}^* = \{\mathbf{W}_1^*, \mathbf{W}_2^*\}$ are known in advance, where $\mathbf{W}_1^* \in \mathbb{R}^{d_x \times d_1'}, \mathbf{W}_2^* \in \mathbb{R}^{d_y \times d_1'}$ are the weights of two layers respectively, we define the loss function $\widetilde{\mathcal{L}}(\mathbf{W}) = \frac{1}{2}\left\|g(\mathbf{W}, \mathbf{X}) - g(\mathbf{W}^*, \mathbf{X})\right\|_F^2$, and consider the expected squared loss:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\widetilde{\mathcal{L}}(\mathbf{W})\right] = \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\left\|g(\mathbf{W}, \mathbf{X}) - g(\mathbf{W}^*, \mathbf{X})\right\|_F^2\right], \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm. Note that we do not restrict that $d_1 = d_1'$, which means that, compared to the network corresponding to $\mathbf{W}^*$, the network corresponding to $\mathbf{W}$ may have a different number of hidden nodes.

2.3. Other Definitions

Definition 1 [Critical Point] *Given a loss function $\mathcal{L}(\mathbf{W})$ of a two-layered ReLU network, a point $\mathbf{W}$ is a critical point of $\mathcal{L}(\mathbf{W})$ if $\mathcal{D}_{\mathrm{vec}(\mathbf{W}_i)}\mathcal{L}(\mathbf{W}) = 0$, where $i \in [2]$.*

Some previous works provided many different definitions of flat minima in experiment and theory. Intuitively, Hochreiter and Schmidhuber (1997) define flat minima as a large connected region in the weight space where the losses remain approximately constant. Additionally, Dinh et al. (2017) give a definition of ϵ-flatness of minima. For convenience, we define the flatness of critical points as follows:

Definition 2 [Flatness of Critical Points] *Given $\epsilon > 0$, and a critical point $\mathbf{W}$ of a loss $\mathcal{L}(\mathbf{W})$, we define $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ as the largest connected set containing $\mathbf{W}$, such that $\forall\ \mathbf{W}' \in \mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon), |\mathcal{L}(\mathbf{W}) - \mathcal{L}(\mathbf{W}')| < \epsilon$. The ϵ-flatness will be defined as the volume of $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$.*

The above definition is very intuitive: a critical point is called flat if there exists a large region around it in which the absolute loss difference between any different point and this critical point is less than ϵ. Otherwise, the critical point is sharp.

A flat critical point always leads to a flat loss surface. However, A question is worth exploring: if the flatness of a critical point is unknown, and will it also cause a flat loss surface? To answer this question, we are interested in whether a critical point is isolated. We define the non-isolated critical point as follows:

Definition 3 [Non-isolated Critical Point] *A critical point $\mathbf{W}$ is non-isolated if there is a largest connected set $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ containing $\mathbf{W}$, and $\mathbf{W}$ is not the only critical point. Otherwise such critical point is called isolated.*

Relying on Definition 3, we will explore whether a large region that leads to the at loss surface contains many non-isolated critical points. To this end, we need some transformations to form these critical points in the large region.

3. Main Results under Different Transformations

In this section, we first introduce the normal equation, and then analyze the behaviors of critical points under three kinds of transformations.

3.1. Normal Equation

To identify and analyze the critical points of the loss function, we take the expectation of the partial derivative of Eqn. (3) w.r.t. $\mathrm{vec}(\mathbf{W}_1)$ and $\mathrm{vec}(\mathbf{W}_2)$, respectively, and set them to 0. Then, we have $\mathbb{E}\left[\mathcal{D}_{\mathrm{vec}(\mathbf{W}_i)}\widetilde{\mathcal{L}}(\mathbf{W})\right] = 0$, where $i \in [2]$. Ideally, we shall find critical points by solving the first-order equations. However, it is often difficult to achieve closed-form solutions of $\mathbf{W}_1$ and $\mathbf{W}_2$. Nevertheless, we can expand the expectation of the partial derivatives and introduce the following normal equation for critical points.

Lemma 1 (Sufficient and necessary conditions for critical points) $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ *is a critical point of* $\widetilde{\mathcal{L}}(\mathbf{W})$ *if and only if*

$$\mathbf{0} = \mathbb{E}\left[\left(\mathcal{D}_{\mathrm{vec}(\mathbf{W}_1)}\widetilde{\mathcal{L}}(\mathbf{W})\right)^{\mathsf{T}}\right] = \frac{N}{2\pi}\sum_k \left(\mathbf{A}_k - \mathbf{A}_k^*\right)\mathbf{W}_2\mathbf{P}_k, \tag{4}$$

$$\mathbf{0} = \mathbb{E}\left[\left(\mathcal{D}_{\mathrm{vec}(\mathbf{W}_2)}\widetilde{\mathcal{L}}(\mathbf{W})\right)^{\mathsf{T}}\right] = \frac{N}{2\pi}\sum_k \mathbf{P}_k\mathbf{W}_1^{\mathsf{T}}\left(\mathbf{A}_k - \mathbf{A}_k^*\right), \tag{5}$$

where $\mathbf{A}_k = \mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^{\mathsf{T}} + \mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^{\mathsf{T}}$, $\mathbf{A}_k^* = \mathbf{W}_1^*\mathbf{\Omega}_k^*\mathbf{W}_2^{*\mathsf{T}} + \mathbf{B}^*\mathbf{\Lambda}_k^*\mathbf{W}_2^{*\mathsf{T}}$, $\mathbf{B} = \left[\mathbf{e}_1^{(1)}, \mathbf{e}_2^{(1)}, \ldots, \mathbf{e}_{d_1}^{(1)}\right]$ *with* $\mathbf{e}_k^{(1)} = \frac{\mathbf{w}_k^{(1)}}{\left\|\mathbf{w}_k^{(1)}\right\|}$, $\mathbf{\Lambda}_k = \frac{N}{2\pi}diag\left(\left[\left\|\mathbf{w}_1^{(1)}\right\|\sin\theta_{k,1}, \left\|\mathbf{w}_2^{(1)}\right\|\sin\theta_{k,2}, \ldots, \left\|\mathbf{w}_{d_1}^{(1)}\right\|\sin\theta_{k,d_1}\right]\right)$, $\mathbf{\Omega}_k = \frac{N}{2\pi}\left(\pi\mathbf{I}_{d_x} - diag([\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,d_1}])\right)$, *where* $\theta_{k,i} = \angle\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)$, $i \in [d_1]$. *Here, the k-th diagonal element of the matrix* $\mathbf{P}_k$ *is equal to 1 and other diagonal elements are equal to 0.*

Instead of solving the above equation, we use it to analyze the behaviors of critical points. Note that the expectation function is differentiable everywhere except the origin, i.e., $\mathbf{W}_1 = \mathbf{0}$. In other words, $\mathbb{E}\left[\mathcal{D}_{\mathrm{vec}(\mathbf{W}_1)}\widetilde{\mathcal{L}}(\mathbf{W})\right]$ and $\mathbb{E}\left[\mathcal{D}_{\mathrm{vec}(\mathbf{W}_2)}\widetilde{\mathcal{L}}(\mathbf{W})\right]$ are not continuous.

Baldi and Hornik (1989) and Kawaguchi (2016) have analyzed linear networks without ReLU. Tian (2017) have analyzed the special case where the last layer contains fixed weights of value 1. Here, for simplicity, we first analyze the expected loss of a two-layered linear network without ReLU.

Proposition 1 *For a two-layered linear network without ReLU, for any fixed* $\mathbf{W}_2 \in \mathbb{R}^{d_y \times d_1}$, *the expectation function* $\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathbf{W})\right]$ *is convex with respect to* $\mathbf{W}_1$, *and its minimum satisfies*

$$\mathbf{W}_1\mathbf{W}_2^{\mathsf{T}}\mathbf{W}_2 = \mathbf{W}_1^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2.$$

If $\mathbf{W}_2$ *is full rank, then* $\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathbf{W})\right]$ *is strictly convex and has a unique minimum:*

$$\mathbf{W}_1 = \mathbf{W}_1^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2\left(\mathbf{W}_2^{\mathsf{T}}\mathbf{W}_2\right)^{-1}.$$

Similarly, for any fixed matrix $\mathbf{W}_1$, the expected loss $\mathbb{E}\left[\mathcal{L}(\mathbf{W})\right]$ is convex with respect to $\mathbf{W}_2$. In this case, if $\mathbf{W}_1$ is full rank, then $\mathcal{L}(\mathbf{W})$ is also strictly convex and has a unique minimum.

In the following, the newtork without ReLU, we will focus on the expectation of the loss function for a two-layered network with ReLU activations. We aim to analyze the behavior of the critical points and the loss function with three common transformations.

3.2. Invariance under Scale Transformation

We first present the definition of scale transformation provided in (Dinh et al., 2017):

Definition 4 [**α-scale transformation**] *For a two-layered ReLU network, the α-scale transformation is defined as:* $\mathcal{T}_\alpha : \{\mathbf{W}_1, \mathbf{W}_2\} \mapsto \{\alpha\mathbf{W}_1, \alpha^{-1}\mathbf{W}_2\}, \alpha > 0$.

Note that an α-scale transformation $\mathcal{T}_\alpha$ dose not affect the prediction function and the loss function, which leads to the following theorem which indicates that $\mathcal{T}_\alpha$ keeps the loss of a critical point invariant.

Theorem 1 [Scale-invariant] *If $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ is a critical point satisfying Eqns. (4) and (5), then for any $\alpha > 0$, $\hat{\mathbf{W}} = \mathcal{T}_\alpha(\mathbf{W}) = \{\alpha\mathbf{W}_1, \alpha^{-1}\mathbf{W}_2\}$ is also a critical point, and $\mathcal{L}(\mathbf{W}) = \mathcal{L}(\hat{\mathbf{W}})$.*

This theorem depicts the level function of the loss $\mathcal{L}$ w.r.t. given weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. From this theorem, all critical points generated by $\mathcal{T}_\alpha$ form continuous critical points and a flat surface, i.e., the losses of these critical points are the same. Furthermore, this Theorem leads to the following proposition.

Proposition 2 *Given a two-layered ReLU network, it follows that a critical point $\mathbf{W} \neq \mathbf{0}$ is non-isolated, and $\forall\ \epsilon > 0, \mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ has an infinite volume.*

The proof is similar to the method in (Dinh et al., 2017). This proposition means that around every critical point, there exists an infinitely large region in which the losses of points remain approximately constant.

The reason of this invariance of loss is that the prediction function is not changed under the α-scale transformation $\mathcal{T}_\alpha$, which changes $\mathbf{W}_1$ and $\mathbf{W}_2$ simultaneously based on parameter α. Next, we pay attention to more complicated cases where $\mathbf{W}_1$ and $\mathbf{W}_2$ can be changed separately.

3.3. Invariance under Rotation Transformation

Without loss of generality, we assume that the optimal parameters $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are known in advance. The optimal parameters are also called ground-truth parameters.

Definition 5 [Principal Hyperplane] *Define Π_1^* and Π_2^* as Principal Hyperplanes spanned by the ground-truth weight vectors $\mathbf{W}_1^* = [\mathbf{w}_1^{*(1)}, \ldots, \mathbf{w}_{d_1'}^{*(1)}]$ and $\mathbf{W}_2^* = \left[\mathbf{w}_1^{*(2)}, \ldots, \mathbf{w}_{d_1'}^{*(2)}\right]$, respectively.*

Note that Π_1^* and Π_2^* are at most d_1'-dimensional, since Π_1^* and Π_2^* are spanned by d_1' ground-truth weight vectors. $\left\{\mathbf{w}_j^{(i)}\right\}_{j=1}^{d_1}$ is said to be in-plane, if all $\mathbf{w}_j^{(i)} \in \Pi^*$, where $i \in \{1, 2\}$; Otherwise, it is out-of-plane. Recall that $\mathbf{\Omega}_k$ only depends on the angles between the column vectors of $\mathbf{W}_1$, and $\mathbf{\Lambda}_k$ only depends on the magnitudes of column vectors of $\mathbf{W}_1$ and the angles between them. Based on these, the following theorem shows the rotation invariance for critical points.

Theorem 2 [Rotation Invariance] *If $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ is a critical point satisfying Eqns. (4) and (5), for any orthogonal mapping matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ with $\mathbf{R}_1|_{\Pi_1^*} = \mathbf{I}_{d_x}$ and $\mathbf{R}_2|_{\Pi_2^*} = \mathbf{I}_{d_y}$ such that $\mathbf{R}_1\mathbf{W}_1^* = \mathbf{W}_1^*$ and $\mathbf{R}_2\mathbf{W}_2^* = \mathbf{W}_2^*$, respectively, then $\bar{\mathbf{W}} = \{\mathbf{R}_1\mathbf{W}_1, \mathbf{W}_2\mathbf{R}_2\}$ is also a critical point.*

When $d_x \geq d_1' + 2$ or $d_y \geq d_1' + 2$, there always exist $\mathbf{R}_1 \neq \mathbf{I}_{d_x}$ and $\mathbf{R}_2 \neq \mathbf{I}_{d_x}$ that yield continuous critical points that lie on a loss surface. Specifically, the rotation invariance in Theorem 2 leads to the following theorem that characterizes the structure of out-of-plane critical points and shows that these critical points lie on a manifold:

Theorem 3 *Given $d_x \geq d_1' + 2$ or $d_y \geq d_1' + 2$, if a critical point $\mathbf{W}$, which satisfies Eqns. (4) and (5), is out-of-plane, then it is non-isolated and lies in a manifold. For $\forall\ \epsilon > 0$, $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ has a large volume ϵ-flatness in this manifold.*

For any out-of-plane critical point, there exists a set of matrices that are not identity matrices but keep Π_i^* invariant. These matrices form a Lie group $SO(d_x)$, in which each element can transform one critical point to another critical point. This theorem shows that such critical point would be adequately flat according to the arbitrary volume ϵ-flatness. As a result, there exists a citical point with a large and flat region where the loss of each point is arbitrarily close to a constant.

Corollary 1 *Suppose $d_1 = d_1' = 1$, $d_x > 1$ is an odd number, if $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ is a critical point satisfying Eqns. (4) and (5), then there exist rotation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ satisfying $\mathbf{R}_1\mathbf{W}_1^* = \mathbf{W}_1^*$ and $\mathbf{R}_2\mathbf{W}_2^* = \mathbf{W}_2^*$, such that $\bar{\mathbf{W}} = \{\mathbf{R}_1\mathbf{W}_1, \mathbf{R}_2\mathbf{W}_2\}$ is also a critical point, and $\mathcal{L}(\mathbf{W}) = \mathcal{L}(\bar{\mathbf{W}})$.*

This corollary also shows that there exists a set of rotaion matrices yielding continuous critical points which form a flat surface.

3.4. Invariance under Perturbation Transformation

The parameter space of a two-layered ReLU network is complicated and high-dimensional. Therefore it is difficult to characterize the behaviors of the critical points and losses between two critical points. Consequently, we do not know whether a training algorithm will be stuck in a critical point, a flat region or a large obstacle.

In fact, Goodfellow et al. (2014) introduce a simple technique for qualitatively analyzing objective functions experientially. In this paper, we introduce a transformation to conduct perturbation on a line between two points theoretically. The definition of perturbation transformation is given as follows:

Definition 6 [Perturbation Transformation] *Given two points $\mathbf{W}$ and $\bar{\mathbf{W}}$, we define a perturbation transformation on a straight line as: $\mathcal{P}_\mu(\mathbf{W}, \bar{\mathbf{W}}) = \mathbf{W} + \mu(\bar{\mathbf{W}} - \mathbf{W}), \mu > 0$.*

Here, $\mu = 0$ or $\mu = 1$ means that there is no perturbation, namely, $\mathcal{P}_0(\mathbf{W}, \bar{\mathbf{W}}) = \mathbf{W}$ and $\mathcal{P}_1(\mathbf{W}, \bar{\mathbf{W}}) = \bar{\mathbf{W}}$. This definition of perturbation transformation is consistent with a straight path in which we can perturb a point to another.

We measure the expected loss for a series of points generated by perturbation transformation with varying values of $\mu > 0$, and employ this simple technique to analyze the behavior of critical points and loss function theoretically.

Theorem 4 [Perturbation Invariance] *Given a fixed weight matrix $\mathbf{W}_1$, if $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ is a critical point satisfying Eqn. (5), then there exists a perturbation of $\mathbf{W}$, such that $\tilde{\mathbf{W}} = \{\mathbf{W}_1, \tilde{\mathbf{W}}_2\}$ is also a critical point and is non-isolated.*

This theorem shows that in the straight path, there exists a flat region where the losses of points remain approximately constant. This raises a question that if weight matrix is not fixed, will it break this perturbation invariance?

Moreover, it is worth mentioning that, in (Goodfellow et al., 2014), they shows that there exists a certain linear space in which the algorithm does not meet another critical point. To understand this phenomenon theoretically, we utilize another critical point $\bar{\mathbf{W}}$ given in Theorem 2 to construct a linear parameter space, and have the following theorem.

Theorem 5 *Let $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ be a critical point satisfying Eqns. (4) and (5), for any orthogonal mapping pair $\mathbf{R}_1$ and $\mathbf{R}_2$ with $\mathbf{R}_1|_{\Pi_1^*} = \mathbf{I}_{d_x}$ and $\mathbf{R}_2|_{\Pi_2^*} = \mathbf{I}_{d_y}$, if $\bar{\mathbf{W}}_1 = \mathbf{R}_1\mathbf{W}_1$ and $\bar{\mathbf{W}}_2 = \mathbf{R}_2\mathbf{W}_2$, then $\tilde{\mathbf{W}} = \mathcal{P}_\mu(\mathbf{W}, \bar{\mathbf{W}}) = \{\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2\}$ cannot be a critical point.*

This theorem indicates that given a critical point, there does not exist perturbation invariance for another critical point that is generated by an orthogonal mapping in Theorem 2. This result explains why there is no critical point in the straight line between two given critical points.

4. Detailed Proofs

4.1. Preliminaries

To facilitate the proof, we need to introduce Population Gating (PG) function $F(\mathbf{e}, \mathbf{w}) \equiv \mathbf{X}^\mathsf{T}\mathbf{D}(\mathbf{e})\mathbf{D}(\mathbf{w})\mathbf{X}\mathbf{w}$ in (Tian, 2017). Moreover, its expectation is given as follows.

Theorem 6 (Tian, 2017) *Given $F(\mathbf{e}, \mathbf{w}) = \mathbf{X}^\mathsf{T}\mathbf{D}(\mathbf{e})\mathbf{D}(\mathbf{w})\mathbf{X}\mathbf{w}$, where $\mathbf{e}$ is a unit vector, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\mathsf{T}$ is an N-by-d data matrix and $\mathbf{D}(\mathbf{w}) = diag(sgn(\mathbf{X}\mathbf{w} > 0))$ is a diagonal matrix with binary elements. If $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}), i \in [d_1]$, then*

$$\mathbb{E}\left[F(\mathbf{e}, \mathbf{w})\right] = \frac{N}{2\pi}\left[(\pi - \theta)\mathbf{w} + \|\mathbf{w}\| \sin\theta \mathbf{e}\right], \tag{6}$$

where $\theta = \angle(\mathbf{e}, \mathbf{w}) \in [0, \pi]$ is the angle between $\mathbf{e}$ and $\mathbf{w}$.

Based on the PG function, we have the following lemma study the normal equation of a two-layered ReLU network.

Lemma 2 *Given any k, let $F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right) = \mathbf{X}^\mathsf{T}\mathbf{D}\left(\mathbf{e}_k^{(1)}\right)\mathbf{D}\left(\mathbf{w}_i^{(1)}\right)\mathbf{X}\mathbf{w}_i^{(1)}, i \in [d_1]$, where $\mathbf{e}_k^{(1)} = \frac{\mathbf{w}_k^{(1)}}{\left\|\mathbf{w}_k^{(1)}\right\|}$, $\mathbf{D}(\mathbf{w}) = diag(sgn(\mathbf{X}\mathbf{w} > 0))$. If $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}), i \in [d_1]$, then*

$$\sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)\right] {\mathbf{w}_i^{(2)}}^\mathsf{T} = \mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^\mathsf{T} + \mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^\mathsf{T}, \tag{7}$$

where $\mathbf{B} = \left[\mathbf{e}_1^{(1)}, \mathbf{e}_2^{(1)}, \ldots, \mathbf{e}_{d_1}^{(1)}\right]$ with $\mathbf{e}_k^{(1)} = \frac{\mathbf{w}_k^{(1)}}{\left\|\mathbf{w}_k^{(1)}\right\|}$, $\mathbf{\Omega}_k = \frac{N}{2\pi}\left(\pi\mathbf{I}_{d_x} - diag([\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,d_1}])\right)$, $\mathbf{\Lambda}_k = \frac{N}{2\pi} diag\left(\left[\left\|\mathbf{w}_1^{(1)}\right\| \sin\theta_{k,1}, \left\|\mathbf{w}_2^{(1)}\right\| \sin\theta_{k,2}, \ldots, \left\|\mathbf{w}_{d_1}^{(1)}\right\| \sin\theta_{k,d_1}\right]\right)$, and $\theta_{k,i} = \angle\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)$, $i \in [d_1]$.

Proof With Eqn. (6) in Theorem 6, we have

$$\begin{aligned}\sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)\right] \mathbf{w}_i^{(2)\mathsf{T}} =& \frac{N}{2\pi} \sum_i \left[(\pi - \theta_{k,i})\mathbf{w}_i^{(1)} + \left\|\mathbf{w}_i^{(1)}\right\| \sin\theta_{k,i}\mathbf{e}_k\right] \mathbf{w}_i^{(2)\mathsf{T}} \\ =& \frac{N}{2\pi} \sum_i \left[(\pi - \theta_{k,i})\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)\mathsf{T}} + \left\|\mathbf{w}_i^{(1)}\right\| \sin\theta_{k,i}\mathbf{e}_k\mathbf{w}_i^{(2)\mathsf{T}}\right] \\ =& \frac{N}{2\pi}\left(\mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^{\mathsf{T}} + \mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^{\mathsf{T}}\right),\end{aligned}$$

where $\mathbf{W}_1 = \left[\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_{d_1}^{(1)}\right] \in \mathbb{R}^{d_x \times d_1}$ and $\mathbf{W}_2 = \left[\mathbf{w}_1^{(2)}, \mathbf{w}_2^{(2)}, \ldots, \mathbf{w}_{d_1}^{(2)}\right] \in \mathbb{R}^{d_y \times d_1}$. ■

Note that the left side of Eqn. (7) is linear to $\mathbf{W}_2^{\mathsf{T}}$, and is also dependent on the weight matrix $\mathbf{W}_1$, including the magnitudes of its column vectors and the angles between them. This lemma will be used to analyze the normal equation of the loss function $\mathcal{L}(\mathbf{W})$ in the following part.

4.2. Proof of Lemma 1

Proof Recall that $\widetilde{\mathcal{L}}(\mathbf{W}) = \frac{1}{2}\left\|g(\mathbf{W}, \mathbf{X}) - g(\mathbf{W}^*, \mathbf{X})\right\|_F^2 = \frac{1}{2}\operatorname{vec}(\mathbf{E})^{\mathsf{T}}\operatorname{vec}(\mathbf{E})$. By taking the partial derivative of $\widetilde{\mathcal{L}}(\mathbf{W})$ w.r.t. $\mathbf{w}_k^{(1)}$ as follows

$$\begin{aligned}\mathcal{D}_{\mathbf{w}_k^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W}) &= \left(\mathcal{D}_{\operatorname{vec}(\mathbf{E})}\widetilde{\mathcal{L}}(\mathbf{W})\right)\left(\mathcal{D}_{\mathbf{w}_k^{(1)}}\operatorname{vec}(\mathbf{E})\right) \\ &= \operatorname{vec}(\mathbf{E})^{\mathsf{T}}\mathcal{D}_{\mathbf{w}_k^{(1)}}\left(\sum_i \operatorname{vec}(g(\mathbf{W}, \mathbf{X})) - \operatorname{vec}(g(\mathbf{W}^*, \mathbf{X}))\right) \\ &= \operatorname{vec}(\mathbf{E})^{\mathsf{T}}\mathcal{D}_{\mathbf{w}_k^{(1)}}\left(\sum_i \operatorname{vec}\left(\mathbf{D}_i\mathbf{X}\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)\mathsf{T}}\right) - \operatorname{vec}(g(\mathbf{W}^*, \mathbf{X}))\right) \\ &= \operatorname{vec}(\mathbf{E})^{\mathsf{T}}\mathcal{D}_{\mathbf{w}_k^{(1)}}\left(\sum_i \left(\mathbf{w}_i^{(2)} \otimes \mathbf{D}_i\mathbf{X}\right)\mathbf{w}_i^{(1)}\right) \\ &= \operatorname{vec}(\mathbf{E})^{\mathsf{T}}\left(\mathbf{w}_k^{(2)} \otimes \mathbf{D}_k\mathbf{X}\right), k \in [d_1],\end{aligned}$$

where $\mathbf{E} = g(\mathbf{W}, \mathbf{X}) - g(\mathbf{W}^*, \mathbf{X})$. By expanding $\mathbf{E}$, we obtain

$$\begin{aligned}\left(\mathcal{D}_{\mathbf{w}_k^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^{\mathsf{T}} &= \left(\mathbf{w}_k^{(2)\mathsf{T}} \otimes \mathbf{X}^{\mathsf{T}}\mathbf{D}_k\right)\operatorname{vec}(\mathbf{E}) = \operatorname{vec}\left(\mathbf{X}^{\mathsf{T}}\mathbf{D}_k\mathbf{E}\mathbf{w}_k^{(2)}\right) = \mathbf{X}^{\mathsf{T}}\mathbf{D}_k\mathbf{E}\mathbf{w}_k^{(2)} \\ &= \mathbf{X}^{\mathsf{T}}\mathbf{D}\left(\mathbf{w}_k^{(1)}\right)\left(\sum_i \mathbf{D}_i\mathbf{X}\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)\mathsf{T}} - g(\mathbf{W}^*, \mathbf{X})\right)\mathbf{w}_k^{(2)} \\ &= \sum_i \mathbf{X}^{\mathsf{T}}\mathbf{D}\left(\mathbf{w}_k^{(1)}\right)\mathbf{D}\left(\mathbf{w}_i^{(1)}\right)\mathbf{X}\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)\mathsf{T}}\mathbf{w}_k^{(2)} - \mathbf{X}^{\mathsf{T}}\mathbf{D}\left(\mathbf{w}_k^{(1)}\right)g(\mathbf{W}^*, \mathbf{X})\mathbf{w}_k^{(2)} \\ &= \sum_i \sum_{j:\mathbf{x}_j^{\mathsf{T}}\mathbf{w}_k^{(1)}>0, \mathbf{x}_j^{\mathsf{T}}\mathbf{w}_i^{(1)}>0} \mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)\mathsf{T}}\mathbf{w}_k^{(2)} - \mathbf{X}^{\mathsf{T}}\mathbf{D}\left(\mathbf{w}_k^{(1)}\right)g(\mathbf{W}^*, \mathbf{X})\mathbf{w}_k^{(2)} \\ &= \sum_i F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)\mathbf{w}_i^{(2)\mathsf{T}}\mathbf{w}_k^{(2)} - \sum_i F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{*(1)}\right)\mathbf{w}_i^{*(2)\mathsf{T}}\mathbf{w}_k^{(2)},\end{aligned}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\mathsf{T}$. The last equation holds since $F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)$ can be written as

$$F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right) = \sum_{j:\mathbf{x}_j^\mathsf{T}\mathbf{w}_k^{(1)}>0, \mathbf{x}_j^\mathsf{T}\mathbf{w}_i^{(1)}>0} \mathbf{x}_j\mathbf{x}_j^\mathsf{T}\mathbf{w}_i^{(1)}.$$

By taking the expectation, we get

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_k^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right] &= \sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)\right] \mathbf{w}_i^{(2)^\mathsf{T}}\mathbf{w}_k^{(2)} - \sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{*(1)}, \mathbf{w}_i^{(1)}\right)\right] \mathbf{w}_i^{*(2)^\mathsf{T}}\mathbf{w}_k^{(2)} \\ &= \frac{N}{2\pi}\left[\left(\mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^\mathsf{T} + \mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^\mathsf{T}\right) - \left(\mathbf{W}_1^*\mathbf{\Omega}_k^*\mathbf{W}_2^{*\mathsf{T}} + \mathbf{B}^*\mathbf{\Lambda}_k^*\mathbf{W}_2^{*\mathsf{T}}\right)\right]\mathbf{w}_k^{(2)}, k \in [d_1],\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{D}_{\mathrm{vec}(\mathbf{W}_1)}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right] &= \left[\mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_1^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right], \mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_2^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right], \ldots, \mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_{d_1}^{(1)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right]\right] \\ &= \frac{N}{2\pi}\left[\left(\mathbf{W}_1\sum_k \mathbf{\Omega}_k\mathbf{W}_2^\mathsf{T}\mathbf{W}_2\mathbf{P}_k + \mathbf{B}\sum_k \mathbf{\Lambda}_k\mathbf{W}_2^\mathsf{T}\mathbf{W}_2\mathbf{P}_k\right)\right. \\ &\quad \left. - \left(\mathbf{W}_1^*\sum_k \mathbf{\Omega}_k^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2\mathbf{P}_k + \mathbf{B}^*\sum_k \mathbf{\Lambda}_k^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2\mathbf{P}_k\right)\right] \\ &= \frac{N}{2\pi}\sum_k \left(\mathbf{A}_k - \mathbf{A}_k^*\right)\mathbf{W}_2\mathbf{P}_k.\end{aligned}$$

Similarly, by taking the partial derivative of $\widetilde{\mathcal{L}}(\mathbf{W})$ with respect to $\mathbf{w}_k^{(2)}$, we get

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_k^{(2)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right] &= \mathbb{E}\left[\mathbf{w}_k^{(1)^\mathsf{T}}\sum_i \mathbf{X}^\mathsf{T}\mathbf{D}_k\mathbf{D}_i\mathbf{X}\mathbf{w}_i^{(1)}\mathbf{w}_i^{(2)^\mathsf{T}} - \mathbf{w}_k^{(1)^\mathsf{T}}\mathbf{X}^\mathsf{T}\mathbf{D}_k g(\mathbf{W}^*, \mathbf{X})\right] \\ &= \mathbf{w}_k^{(1)^\mathsf{T}}\left[\sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{(1)}, \mathbf{w}_i^{(1)}\right)\right]\mathbf{w}_i^{(2)^\mathsf{T}} - \sum_i \mathbb{E}\left[F\left(\mathbf{e}_k^{*(1)}, \mathbf{w}_i^{*(1)}\right)\right]\mathbf{w}_i^{*(2)^\mathsf{T}}\right] \\ &= \frac{N}{2\pi}\mathbf{w}_k^{(1)^\mathsf{T}}\left[\left(\mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^\mathsf{T} + \mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^\mathsf{T}\right) - \left(\mathbf{W}_1^*\mathbf{\Omega}_k^*\mathbf{W}_2^{*\mathsf{T}} + \mathbf{B}^*\mathbf{\Lambda}_k^*\mathbf{W}_2^{*\mathsf{T}}\right)\right], k \in [d_1].\end{aligned}$$

We then have

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{D}_{\mathrm{vec}(\mathbf{W}_2)}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right] &= \left[\mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_1^{(2)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right], \mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_2^{(2)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right], \ldots, \mathbb{E}\left[\left(\mathcal{D}_{\mathbf{w}_{d_1}^{(2)}}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right]\right] \\ &= \frac{N}{2\pi}\left[\sum_k \mathbf{P}_k\mathbf{W}_1^\mathsf{T}\mathbf{W}_1\mathbf{\Omega}_k\mathbf{W}_2^\mathsf{T} + \sum_k \mathbf{P}_k\mathbf{W}_1^\mathsf{T}\mathbf{B}\mathbf{\Lambda}_k\mathbf{W}_2^\mathsf{T}\right] \\ &\quad - \frac{N}{2\pi}\left[\sum_k \mathbf{P}_k\mathbf{W}_1^\mathsf{T}\mathbf{W}_1^*\mathbf{\Omega}_k^*\mathbf{W}_2^{*\mathsf{T}} + \sum_k \mathbf{P}_k\mathbf{W}_1^\mathsf{T}\mathbf{B}^*\mathbf{\Lambda}_k^*\mathbf{W}_2^{*\mathsf{T}}\right] \\ &= \frac{N}{2\pi}\sum_k \mathbf{P}_k\mathbf{W}_1^\mathsf{T}\left(\mathbf{A}_k - \mathbf{A}_k^*\right).\end{aligned}$$

By setting $\mathbb{E}\left[\left(\mathcal{D}_{\mathrm{vec}(\mathbf{W}_i)}\widetilde{\mathcal{L}}(\mathbf{W})\right)^\mathsf{T}\right] = \mathbf{0}$ for all $i \in \{1, 2\}$, we achieve the statement of Lemma 1. ■

4.3. Proof of Proposition 1

Proof For any fixed $\mathbf{W}_2$, we can rewrite $\widetilde{\mathcal{L}}(\mathbf{W})$ as

$$\begin{aligned}\widetilde{\mathcal{L}}(\mathbf{W}) &= \|g(\mathbf{W},\mathbf{X}) - g(\mathbf{W}^*,\mathbf{X})\|_F^2 \\ &= \left\|\mathrm{vec}\left(\mathbf{X}\mathbf{W}_1\mathbf{W}_2^\mathsf{T}\right) - \mathrm{vec}\left(\mathbf{X}\mathbf{W}_1\mathbf{W}_2^\mathsf{T}\right)\right\|_2^2 \\ &= \|(\mathbf{W}_2\otimes\mathbf{X})\,\mathrm{vec}\,(\mathbf{W}_1) - (\mathbf{W}_2^*\otimes\mathbf{X})\,\mathrm{vec}\,(\mathbf{W}_1^*)\|_2^2\,.\end{aligned}$$

By setting $\nabla\widetilde{\mathcal{L}}(\mathbf{W}) = \mathbf{0}$, we obtain

$$\begin{aligned}&\left(\mathbf{W}_2^\mathsf{T}\otimes\mathbf{X}^\mathsf{T}\right)(\mathbf{W}_2\otimes\mathbf{X})\,\mathrm{vec}(\mathbf{W}_1) = \left(\mathbf{W}_2^\mathsf{T}\otimes\mathbf{X}^\mathsf{T}\right)(\mathbf{W}_2^*\otimes\mathbf{X})\,\mathrm{vec}\,(\mathbf{W}_1^*) \\ \Rightarrow &\left(\mathbf{W}_2^\mathsf{T}\mathbf{W}_2\otimes\mathbf{X}^\mathsf{T}\mathbf{X}\right)\mathrm{vec}(\mathbf{W}_1) = \left(\mathbf{W}_2^\mathsf{T}\mathbf{W}_2^*\otimes\mathbf{X}^\mathsf{T}\mathbf{X}\right)\mathrm{vec}\,(\mathbf{W}_1^*) \\ \Rightarrow &\,\mathrm{vec}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{W}_1\mathbf{W}_2^\mathsf{T}\mathbf{W}_2\right) = \mathrm{vec}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{W}_1^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2\right) \\ \Rightarrow &\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{W}_1\mathbf{W}_2^\mathsf{T}\mathbf{W}_2 = \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{W}_1^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2.\end{aligned}$$

For the standard Gaussian input $\mathbf{X}$, we have $\mathbb{E}_\mathbf{X}[\mathbf{X}^\mathsf{T}\mathbf{X}] = \mathbf{I}$. Considering that $\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathbf{W})\right]$ is convex with respect to $\mathbf{W}_1$, $\mathbf{W}_1$ is a global minimum if and only if

$$\mathbf{W}_1\mathbf{W}_2^\mathsf{T}\mathbf{W}_2 = \mathbf{W}_1^*\mathbf{W}_2^{*\mathsf{T}}\mathbf{W}_2.$$

If $\mathbf{W}_2$ is full rank, $\mathbf{W}_2^\mathsf{T}\mathbf{W}_2$ is symmetric and positive definite, then $\mathbb{E}\left[\mathcal{L}(\mathbf{W})\right]$ is strictly convex w.r.t. $\mathbf{W}_1$ and has a unique minimum. ■

4.4. Proof of Theorem 1

Proof Since $\alpha > 0$, for a rectified linear function $\sigma(\cdot)$, we have $\sigma(\mathbf{X}(\alpha\mathbf{W}_1)) = \alpha\sigma(\mathbf{X}\mathbf{W}_1)$. For the two-layered ReLU network, we have

$$\sigma(\mathbf{X}(\alpha\mathbf{W}_1))(\alpha^{-1}\mathbf{W}_2^\mathsf{T}) = \sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2^\mathsf{T}.$$

Therefore, the loss is invariant. If $\mathbf{W} = \{\mathbf{W}_1,\mathbf{W}_2\}$ is a critical point, then it is easy to verify that $\hat{\mathbf{W}} = \{\alpha\mathbf{W}_1,\alpha^{-1}\mathbf{W}_2^\mathsf{T}\}$ is also a critical point satisfying Eqns. (4) and (5). ■

4.5. Proof of Theorem 2

Proof Since $\mathbf{R}_1$ and $\mathbf{R}_2$ are orthogonal transformations, they do not change all the magnitudes of the column vectors of $\mathbf{W}_1$ and the angles between them. i.e., $\angle\left(\mathbf{w}_k^{(1)},\mathbf{w}_{k'}^{(1)}\right)$ is not changed, and $\left\|\mathbf{R}_i\mathbf{w}_k^{(i)}\right\| = \left\|\mathbf{w}_k^{(i)}\right\|$, $\left\|\mathbf{R}_i\mathbf{w}^{*(i)}_k\right\| = \left\|\mathbf{w}^{*(i)}_k\right\|$, where $k\in[d_1], i\in[2]$. Therefore, $\mathbf{\Omega}_k$ and $\mathbf{\Lambda}_k$ are not changed. Since $\mathbf{R}_1|_{\Pi_1^*} = \mathbf{I}_{d_x}$ and $\mathbf{R}_2|_{\Pi_2^*} = \mathbf{I}_{d_y}$, then $\mathbf{\Omega}_k^*$ and $\mathbf{\Lambda}_k^*$ are not changed, we thus have $\mathbf{R}_i\mathbf{W}_i^* = \mathbf{W}_i^*$ but $\mathbf{R}_i\mathbf{W}_i \neq \mathbf{W}_i$, where $i\in[2]$.

We firstly discuss the left side of Eqn. (4) with the orthogonal transformation $\mathbf{R}_1$,

$$\begin{aligned}
&\bar{\mathbf{W}}_1 \sum_k \mathbf{\Omega}_k \bar{\mathbf{W}}_2^\top \bar{\mathbf{W}}_2 \mathbf{P}_k + \bar{\mathbf{B}} \sum_k \mathbf{\Lambda}_k \bar{\mathbf{W}}_2^\top \bar{\mathbf{W}}_2 \mathbf{P}_k \\
=&\mathbf{R}_1 \mathbf{W}_1 \sum_k \mathbf{\Omega}_k \mathbf{W}_2^\top \mathbf{R}_2^\top \mathbf{R}_2 \mathbf{W}_2 \mathbf{P}_k + \mathbf{R}_1 \mathbf{B} \sum_k \mathbf{\Lambda}_k \mathbf{W}_2^\top \mathbf{R}_2^\top \mathbf{R}_2 \mathbf{W}_2 \mathbf{P}_k \\
=&\mathbf{R}_1 \sum_k \mathbf{A}_k \mathbf{W}_2 \mathbf{P}_k. \qquad (8)
\end{aligned}$$

For the right side of Eqn. (4), we have

$$\begin{aligned}
&\mathbf{W}_1^* \sum_k \mathbf{\Omega}_k^* \mathbf{W}_2^{*\top} \mathbf{W}_2 \mathbf{P}_k + \mathbf{B}^* \sum_k \mathbf{\Lambda}_k^* \mathbf{W}_2^{*\top} \mathbf{W}_2 \mathbf{P}_k \\
=&\mathbf{R}_1 \mathbf{W}_1^* \sum_k \mathbf{\Omega}_k^* \mathbf{W}_2^{*\top} \mathbf{R}_2^\top \mathbf{R}_2 \mathbf{W}_2 \mathbf{P}_k + \mathbf{R}_1 \mathbf{B}^* \sum_k \mathbf{\Lambda}_k^* \mathbf{W}_2^{*\top} \mathbf{R}_2^\top \mathbf{R}_2 \mathbf{W}_2 \mathbf{P}_k \\
=&\mathbf{R}_1 \sum_k \mathbf{A}_k^* \mathbf{W}_2 \mathbf{P}_k. \qquad (9)
\end{aligned}$$

Combining Eqns. (8) and (9), we get $\mathbf{R}_1 \sum_k (\mathbf{A}_k - \mathbf{A}_k^*) \mathbf{W}_2 \mathbf{P}_k = \mathbf{0}$, which satisfies the condition of a critical point in Eqn. (4). Now, we discuss Eqn. (5) under the orthogonal transformation $\mathbf{R}_2$. For the left side of Eqn. (5), we have

$$\begin{aligned}
&\sum_k \mathbf{P}_k \bar{\mathbf{W}}_1^\top \bar{\mathbf{W}}_1 \mathbf{\Omega}_k \bar{\mathbf{W}}_2^\top + \sum_k \mathbf{P}_k \mathbf{W}_1^\top \bar{\mathbf{B}} \mathbf{\Lambda}_k \bar{\mathbf{W}}_2^\top \\
=&\sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{R}_1^\top \mathbf{R}_1 \mathbf{W}_1 \mathbf{\Omega}_k \mathbf{W}_2^\top \mathbf{R}_2^\top + \sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{R}_1^\top \mathbf{R}_1 \mathbf{B} \mathbf{\Lambda}_k \mathbf{W}_2^\top \mathbf{R}_2^\top \\
=&\sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{A}_k \mathbf{R}_2^\top. \qquad (10)
\end{aligned}$$

For the right side of Eqn. (5), we have

$$\begin{aligned}
&\sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{W}_1^* \mathbf{\Omega}_k^* \mathbf{W}_2^{*\top} + \sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{B}^* \mathbf{\Lambda}_k^* \mathbf{W}_2^{*\top} \\
=&\sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{R}_1^\top \mathbf{R}_1 \mathbf{W}_1^* \mathbf{\Omega}_k^* \mathbf{W}_2^{*\top} \mathbf{R}_2^\top + \sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{R}_1^\top \mathbf{R}_1 \mathbf{B}^* \mathbf{\Lambda}_k^* \mathbf{W}_2^{*\top} \mathbf{R}_2^\top \\
=&\sum_k \mathbf{P}_k \mathbf{W}_1^\top \mathbf{A}_k^* \mathbf{R}_2^\top. \qquad (11)
\end{aligned}$$

Combining Eqns. (10) and (11), we get $\sum_k \mathbf{P}_k \mathbf{W}_1^\top (\mathbf{A}_k - \mathbf{A}_k^*) \mathbf{R}_2^\top = \mathbf{0}$, which also satisfies the condition of a critical point in Eqn. (5). Thus, $\bar{\mathbf{W}} = \{\mathbf{R}_1 \mathbf{W}_1, \mathbf{R}_2 \mathbf{W}_2\}$ is also a critical point satisfying Eqns. (4) and (5). ■

4.6. Proof of Theorem 3

Proof Without loss of generality, we only consider $\mathbf{W}_1$, and the discussions for $\mathbf{W}_2$ are similar. If $d_x \geq d_1' + 2$, since Π_1^* is spanned by d_1' ground-truth weight vectors, then the dimension of Π_1^* is at most d_1'. Therefore, Π_1^* is embedded in a d_x-dimensional space, and any $\mathbf{W}_1$ satisfying Eqns. (4) and (5) is outside Π_1^* .

We will first show that there exists a rotation matrix that is not an identity matrix but keeps Π_1^* invariant. In fact, there always exist such matrices since for a $(d_x - d_1')$-dimensional subspace, (i) if $d_x - d_1'$ is an odd number, then we can always choose a rotation matrix whose fixed axis is not aligned with all d_1' weights; (ii) if $d_x - d_1'$ is an even number, then there exists a rotation matrix without a fixed point. Therefore, given $\epsilon > 0$, such matrices will form a Lie group $SO(d_x)$ that transforms a critical point $\mathbf{W}$ to a different yet infinitely close critical point $\mathbf{W}'$, s.t., $\mathbf{W}' \in \mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon), \epsilon > 0$. Since $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ contains $\mathbf{W}$ and $\mathbf{W}'$, then the critical point $\mathbf{W}$ is non-isolated. In addition, such metrices yield continuous critical points with an equal loss.

Now we introduce a small region $\mathcal{B}_\infty(r, \mathbf{W}) \subseteq \mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ with approximately constant loss around by $\mathbf{W}$ with non-zero volume. Here, $\mathcal{B}_\infty(r, \mathbf{W})$ is an ℓ_∞ ball of radius r centered $\mathbf{W}$, $r \leq \min(\|\operatorname{vec}(\mathbf{W}_1)\|_\infty, \|\operatorname{vec}(\mathbf{W}_2)\|_\infty)$. The volume of the ℓ_∞ ball is $v = (2r)^{n_1+n_2}$, where $n_1 = \dim(\operatorname{vec}(\mathbf{W}_1))$ and $n_2 = \dim(\operatorname{vec}(\mathbf{W}_2))$.

We construct a subset $\mathcal{R}(\mathbf{W}) \subseteq SO(d_x)$ with m rotation matrices, s.t., $\mathbf{R}^{(i)}(\mathcal{B}_\infty(r, \mathbf{W})) \cap \mathcal{B}_\infty(r, \mathbf{W}) = \emptyset$, where $\mathbf{R}^{(i)} \in \mathcal{R}(\mathbf{W}), i \in [m]$, that is, $\mathbf{R}^{(i)}(\mathcal{B}_\infty(r, \mathbf{W}))$ is disjoint of $\mathcal{B}_\infty(r, \mathbf{W})$.

We define a connected region

$$\mathcal{C}' = \left\{ \mathbf{R}^{(i)}\mathbf{W}' = \left\{ \mathbf{R}_1^{(i)}\mathbf{W}_1', \mathbf{R}_2^{(i)}\mathbf{W}_2' \right\} \middle| \mathbf{W}' = \{\mathbf{W}_1', \mathbf{W}_2'\} \in \mathcal{B}(r, \mathbf{W}), \mathbf{R}^{(i)} \in \mathcal{R}(\mathbf{W}), i \in [n] \right\}$$

in which the loss is approximately equivalent.

Let the volume of $\mathcal{B}_\infty(r, \mathbf{W})$ be v_0, and the volume of $\mathbf{R}^{(i)}(\mathcal{B}_\infty(r, \mathbf{W}))$ be v_i. We sum their volume to lower bound the volume of $\mathcal{C}'$ by $\sum_{i=0}^{m} v_i$. Therefore, $\mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ has a finite volume, leading to the large volume ϵ-flatness. ■

4.7. Proof of Corollary 1

Proof Since d_x is an odd number, there exists at least one eigenvalue equal to 1, and at least one axis being unaffected by the rotation. Given a non-zero ground-truth weight matrix $\mathbf{W}^* = \{\mathbf{W}_1^*, \mathbf{W}_2^*\}$, it implies that there exist rotation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ satisfying $\mathbf{R}_1\mathbf{W}_1^* = \mathbf{W}_1^*$ and $\mathbf{W}_2^{*\mathsf{T}}\mathbf{R}_2^\mathsf{T} = \mathbf{W}_2^{*\mathsf{T}}$ if $\det(\mathbf{R}_i - \mathbf{I}) = 0, i \in [2]$. That is to say, the rotations $\mathbf{R}_1$ and $\mathbf{R}_2$ around the rotation axes $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ do not affect $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$, respectively. Note that $\mathbf{R}_1$ and $\mathbf{R}_2$ can be non-identity matrices.

The proof is similar to that of Theorem 2, $\bar{\mathbf{W}} = \{\mathbf{R}_1\mathbf{W}_1, \mathbf{R}_2\mathbf{W}_2\}$ is also a critical point. Now, we will show the orthogonal matrix $\mathbf{R}_2$ remains the loss invariant. Since $\mathbf{R}_2^\mathsf{T}\mathbf{R}_2 = \mathbf{R}_2\mathbf{R}_2^\mathsf{T} = \mathbf{I}$, we have

$$\begin{aligned}\widetilde{\mathcal{L}}(\mathbf{W}) &= \frac{1}{2}\left\|\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2^{\mathsf{T}} - \sigma(\mathbf{X}\mathbf{W}_1^*)\mathbf{W}_2^{*\mathsf{T}}\right\|_F^2 \\ &= \frac{1}{2}\left\|\sigma(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2{}^{\mathsf{T}}\mathbf{R}_2^{\mathsf{T}} - \sigma(\mathbf{X}\mathbf{W}_1^*)\mathbf{W}_2^{*\mathsf{T}}\mathbf{R}_2^{\mathsf{T}}\right\|_F^2 \\ &= \frac{1}{2}\left\|\sigma(\mathbf{X}\mathbf{W}_1)\bar{\mathbf{W}}_2^{\mathsf{T}} - \sigma(\mathbf{X}\mathbf{W}_1^*)\mathbf{W}_2^{*\mathsf{T}}\right\|_F^2.\end{aligned}$$

Therefore, the orthogonal matrix $\mathbf{R}_2$ does not change $\widetilde{\mathcal{L}}(\mathbf{W})$ and its expectation. For an orthogonal matrix $\mathbf{R}_1$, we expand the loss function $\mathcal{L}(\mathbf{W})$ and get

$$\begin{aligned}&\mathcal{L}(\mathbf{W}) = \frac{1}{2}\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathbf{W})\right] = \frac{1}{2}\mathrm{tr}\left(\mathbb{E}\left[\mathbf{E}^{\mathsf{T}}\mathbf{E}\right]\right) \\ =&\frac{1}{2}\mathrm{tr}\left(\sum_i\sum_j \mathbf{w}_i^{(2)}\mathbf{w}_i^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}_i^{(1)},\mathbf{w}_j^{(1)}\right)\right]\mathbf{w}_j^{(2)\mathsf{T}} - \mathbf{w}_i^{(2)}\mathbf{w}_i^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}_i^{(1)},\mathbf{w}_j^{(1)*}\right)\right]\mathbf{w}_j^{(2)*\mathsf{T}}\right. \\ &\left.-\mathbf{w}_i^{(2)*}\mathbf{w}_i^{(1)*\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}_i^{(1)*},\mathbf{w}_j^{(1)}\right)\right]\mathbf{w}_j^{(2)\mathsf{T}} + \mathbf{w}_i^{(2)*}\mathbf{w}_i^{(1)*\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}_i^{(1)*},\mathbf{w}_j^{(1)*}\right)\right]\mathbf{w}_j^{(2)*\mathsf{T}}\right) \\ =&\frac{1}{2}\mathrm{tr}\left(\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)},\mathbf{w}^{(1)}\right)\right]\mathbf{w}^{(2)\mathsf{T}} - \mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)},\mathbf{w}^{(1)*}\right)\right]\mathbf{w}^{(2)*\mathsf{T}}\right. \\ &\left.-\mathbf{w}^{(2)*}\mathbf{w}^{(1)*\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)*},\mathbf{w}^{(1)}\right)\right]\mathbf{w}^{(2)\mathsf{T}} + \mathbf{w}^{(2)*}\mathbf{w}^{(1)*\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)*},\mathbf{w}^{(1)*}\right)\right]\mathbf{w}^{(2)*\mathsf{T}}\right),\end{aligned} \tag{12}$$

where $\mathbf{E} = g(\mathbf{W},\mathbf{X}) - g(\mathbf{W}^*,\mathbf{X})$, $\mathbf{W}_i = \mathbf{w}^{(i)}$ and $\mathbf{W}_i^* = \mathbf{w}^{(i)*}, i \in [2]$.

Now we discuss the first two terms on the right side of Eqn. (12),

$$\begin{aligned}&\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)},\mathbf{w}^{(1)}\right)\right]\mathbf{w}^{(2)\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\left[(\pi-\theta)\mathbf{w}^{(1)} + \left\|\mathbf{w}^{(1)}\right\|\sin\theta\mathbf{e}^{(1)}\right]\mathbf{w}^{(2)\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbf{R}_1^{\mathsf{T}}\mathbf{R}_1\left[(\pi-\theta)\mathbf{w}^{(1)} + \left\|\mathbf{w}^{(1)}\right\|\sin\theta\mathbf{e}^{(1)}\right]\mathbf{w}^{(2)\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\bar{\mathbf{w}}^{(1)\mathsf{T}}\left[(\pi-\theta)\bar{\mathbf{w}}^{(1)} + \left\|\bar{\mathbf{w}}^{(1)}\right\|\sin\theta\bar{\mathbf{e}}^{(1)}\right]\mathbf{w}^{(2)\mathsf{T}}.\end{aligned}$$

Since the rotation $\mathbf{R}_1$ around the rotation axis $\mathbf{W}_1^*$ keeps the weight vector $\mathbf{W}_1^*$ invariant, then $\mathbf{R}_1$ does not change $\theta^* = \angle\left(\mathbf{w}^{(1)},\mathbf{w}^{(1)*}\right)$. As a result, we have the following property:

$$\begin{aligned}&\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbb{E}\left[F\left(\mathbf{e}^{(1)},\mathbf{w}^{(1)*}\right)\right]\mathbf{w}^{(2)*\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\left[(\pi-\theta^*)\mathbf{w}^{(1)*} + \left\|\mathbf{w}^{(1)*}\right\|\sin\theta^*\mathbf{e}^{(1)}\right]\mathbf{w}^{(2)*\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\mathbf{w}^{(1)\mathsf{T}}\mathbf{R}_1^{\mathsf{T}}\mathbf{R}_1\left[(\pi-\theta^*)\mathbf{w}^{(1)*} + \left\|\mathbf{w}^{(1)*}\right\|\sin\theta^*\mathbf{e}^{(1)}\right]\mathbf{w}^{(2)*\mathsf{T}} \\ =&\frac{N}{2\pi}\mathbf{w}^{(2)}\bar{\mathbf{w}}^{(1)\mathsf{T}}\left[(\pi-\theta^*)\mathbf{w}^{(1)*} + \left\|\mathbf{w}^{(1)*}\right\|\sin\theta^*\bar{\mathbf{e}}^{(1)}\right]\mathbf{w}^{(2)*\mathsf{T}}\end{aligned}$$

The remaining two items can be discussed in a similar way. Therefore, the expected loss function $\mathbb{E}\left[\mathcal{L}(\widetilde{\mathbf{W}})\right]$ is invariant. ■

4.8. Proof of Theorem 5

Proof Recall that $\mathbf{\Omega}_k$ only depends on the angles between the column vectors of $\mathbf{W}_1$, and $\mathbf{\Lambda}_k$ only depends on the magnitudes of column vectors of $\mathbf{W}_1$ and the angles between them. If we take a perturbation on the linear path $\tilde{\mathbf{W}} = (1-\mu)\mathbf{W} + \mu\bar{\mathbf{W}}, \mu > 0$, such perturbation cannot guarantee to keep $\mathbf{\Omega}_k$ and $\mathbf{\Lambda}_k$ invariant simultaneously, unless this perturbation is occured at $\mathbf{W}$ or $\bar{\mathbf{W}}$. For such perturbation on $\mathbf{W}_2$, $\tilde{\mathbf{W}}_2^\mathsf{T}\tilde{\mathbf{W}}_2$ cannot satisfy Eqn. (4). Therefore $\tilde{\mathbf{W}}$ cannot be a critical point. ■

4.9. Proof of Theorem 4

Proof Let $\mathbf{M} = \sum_k \mathbf{P}_k \mathbf{W}_1^\mathsf{T} (\mathbf{W}_1\mathbf{\Omega}_k + \mathbf{B}\mathbf{\Lambda}_k)$ and $\mathbf{Q} = \sum_k \mathbf{P}_k \mathbf{W}_1^\mathsf{T} (\mathbf{W}_1^*\mathbf{\Omega}_k^* + \mathbf{B}\mathbf{\Lambda}_k^*)\mathbf{W}_2^{*\mathsf{T}}$. Note that $\mathbf{M}$ may be not full rank, then we have

$$\mathbf{W}_2^\mathsf{T} = \mathbf{M}^-\mathbf{Q} + (\mathbf{I} - \mathbf{M}^-\mathbf{M})\mathbf{L}, \ \forall\, \mathbf{L} \in \mathbb{R}^{d_1 \times d_1}, \tag{13}$$

where $\mathbf{M}^-$ is a generalized inverse of $\mathbf{M}$. We can choose a $\bar{\mathbf{W}}_2$ to satisfy Eqn. (13) for different $\mathbf{L}$. Given any fixed $\mathbf{W}_1$, and $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2\}$ is a critical point, then there exists a perturbation on the straight line $\tilde{\mathbf{W}}_2 = \mathbf{W}_2 + \mu(\bar{\mathbf{W}}_2 - \mathbf{W}_2), \mu > 0$ such that $\tilde{\mathbf{W}}_2$ also satisfies Eqn. (5), hence $\tilde{\mathbf{W}}$ is also a critical point.

Any critical point $\tilde{\mathbf{W}} \in \mathcal{C}(\mathcal{L}, \mathbf{W}, \epsilon)$ can be obtained by perturbing to a different yet infinitely close $\mathbf{W}_1$ when $\mu \to 0$. Therefore, $\tilde{\mathbf{W}}$ is a non-isolated critical point. ■

5. Conclusion and Future works

In this paper, we have theoretically analyzed the properties of a two-layered ReLU network trained by minimizing the expected squared loss between its prediction function and a target function with known optimal parameters. To characterize the behaviors of critical points and loss surface, we propose a normal equation for critical points, and study the invariances under three kinds of transformations, namely, scale transformation, rotation transformation, and perturbation transformation. We find that these transformations keep the loss of a critical point invariant, thus can form flat regions. Our results indicate that the flat loss surfaces indeed exist, thus it is very important to deal with the flatness issue when training neural networks.

There are many open questions to be further explored in the future. How to apply similar analysis to general distributions and how to generalize the analysis of two-layered network to multi-layered network are also open problems.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) under Grants 61502177 and 61602185, Recruitment Program for Young Professionals, Fundamental Research Funds for the Central Universities under Grants D2172500 and D2172480, Guangdong Provincial Scientific and Technological Funds under Grants 2017B090901008 and 2017A010101011 and CCF-Tencent Open Research Fund.

References

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Zachary C Lipton. Stuck in a what? adventures in weight space. *arXiv preprint arXiv:1602.07320*, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Ruhi Sarikaya, Geoffrey E Hinton, and Anoop Deoras. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(4):778–784, 2014.

Ohad Shamir. Distribution-specific hardness of learning neural networks. *arXiv preprint arXiv:1609.01037*, 2016.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3404–3413, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/tian17a.html.