EXISTENCE OF A BAD LOCAL MINIMUM OF NEURAL NETWORKS WITH GENERAL SMOOTH ACTIVATION FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the loss surface of neural networks is essential to the understanding of deep learning. However, the existence of a bad local minimum has not yet been fully identified. We investigate the existence of a bad local minimum of the 2-layer and 3-layer neural networks with general smooth activation functions. We provide constructive proof using the algebraic nature of the activation functions. We show this for realistic settings where the data (X, Y) have a positive measure. We hope that such results give theoretical foundations for studies related to local minima and loss surfaces.

1 INTRODUCTION

Modern machine learning with neural networks has shown remarkable results in many real-world applications. However, little is known about the theoretical foundation of how the neural network works. In particular, most modern machine learning models rely on gradient descent-based optimization algorithms which minimize the difference between the output of the neural network and the target function. In this context, understanding the loss surface of neural networks is of fundamental importance.

The question of the existence of a bad local minimum is also very important because it provides whether the gradient descent-based algorithms can stably reach the global minimum without falling into the local minimum (in this context, a bad local minimum means it is not a global minimum). Moreover, because various optimization-based studies on loss surfaces are based on the premise of the existence of a bad local minimum, investigating its existence will support these studies as a theoretical background. (Jastrzebski et al., 2017; Kleinberg et al., 2018; Zhu et al., 2018; Xie et al., 2020; Ziyin et al., 2021; Mori et al., 2022).

For convex loss function, it is widely known that the loss surface has a unique global minimum. For a general neural network, it is not easy to investigate the loss surface because of its strong non-convexity. Several works suggest that there exists no bad local minimum in a deep linear network. Kawaguchi (2016) and Lu & Kawaguchi (2017) show that there are only global minima and saddle points in a deep linear network with squared error. Laurent & Brecht (2018) show that every local minimum of the deep linear network is global under any differentiable convex loss function.

On the other hand, the existence of a bad local minimum is reported in the deep non-linear network. Yun et al. (2018) and He et al. (2020) show that a bad local minimum exists in the neural network with piece-wise linear activations. If the activations have partial linearity like piece-wise linear functions, a local minimum can be constructed by borrowing some weights from a linear minimizer. For general smooth activation functions, the problem is more difficult. Petzka & Sminchisescu (2021) show that a bad local minimum exists in the deep neural network with sigmoid activation functions using the local minimum embedding. Ding et al. (2022) also construct a bad local minimum in the 2-layer network with sigmoid activations. Although these two studies succeed in finding a bad local minimum, they only show a single example of the sigmoid function, rather than the general smooth activations. This is because properties that should be verified experimentally for each activation are required in the proof. Therefore, the existence of a bad local minimum for general smooth activations remains unclear. The summary of the studies of the existence of a bad local minimum, as well as our own results, is presented in Table 1.

Reference	Depth	Activation	Bad local minima	Condition
Baldi & Hornik (1989)	2	linear	Not exist	
Lu & Kawaguchi (2017)	L	linear	Not exist	almost all (X, Y)
Yun et al. (2018)	2	2-piece linear	Exist	almost all (X, Y)
He et al. (2020)	L	piece-wise linear	Exist	almost all (X, Y)
Ding et al. (2022)	L	almost all smooth[1]	Exist	zero measure
Ding et al. (2022)	2	sigmoid	Exist	positive measure
Petzka & Sminchisescu (2021)	L	sigmoid	Exist	zero measure
Ours	2	almost all analytic[2]	Exist	$N \geq 7$, positive measure
Ours	3	some analytic[3]	Exist	$N \geq 34$, positive measure

Table 1: A summary of the studies on the existence of a bad local minimum. N denotes the number of samples.

[1] For some $a \in \mathbb{R}$, $\sigma(x)$ is twice differentiable on $[a-\delta, a+\delta]$ and $\sigma(a), \sigma'(a), \sigma''(a) \neq 0$.
[2] Assumption 1.
[3] Assumption 2.

In this context, we first present a pure mathematical proof of the existence of a bad local minimum for the general smooth activations and for data of positive measure (Theorem 3, 4). In this proof, we use the differential Galois theory, and no computational experiments are required. We find the mild conditions of the smooth activations, which allow the neural network to have a bad local minimum (Assumption 1, 2). Fortunately, despite the existence of a bad local minimum, it is not very bad because it is non-attracting (there is a non-increasing path to the global minimum). The existence of a really bad local minimum in the training (it is an attracting local minimum) needs to be investigated in the future. We hope that this approach of finding a local minimum will inspire future studies on loss surfaces.

We summarize our key contributions below:

- For almost all analytic functions $\sigma(x)$, we show that the 2-layer network $1 - d_1 - 1$ ($d_1 \geq 2$) with $\sigma(x)$ activation function has a bad local minimum for a positive measure of data $(X, Y) \in (\mathbb{R}^{1\times N}, \mathbb{R}^{1\times N})$ with $N \geq 7$.

- For some analytic functions $\sigma(x)$ satisfying Assumption 2, including famous activation functions such as $sigmoid$, $GELU$, and $Swish$, we show that the 3-layer network $1 - d_1 - d_2 - 1$ ($d_1, d_2 \geq 2$) with $\sigma(x)$ activation function has a bad local minimum for a positive measure of data $(X, Y) \in (\mathbb{R}^{1\times N}, \mathbb{R}^{1\times N})$ with $N \geq 34$.

The paper is organized as follows. We first show that we construct a local minimum in a deep neural network with partially linear activation by borrowing parameters from the linear model. Then for $N \geq 7$ and L^2 loss, we find that there exists a strict local minimum in the 2-layer network of width 1 with smooth activations which satisfy some mild assumptions. We show that most of widely-used activation functions satisfy this assumption. We extend these results for data of positive measure. Using the local minimum embedding, we show that a bad local minimum exists in the 2-layer network of width ≥ 2 . Furthermore, for $N \geq 34$, in the 3-layer network of width ≥ 2 with smooth activations, we also find a bad local minimum for data of positive measure.

2 Related works

2.1 Existence and Non-Existence of a local minimum

There is considerable literature to study the loss surface in the neural network. First, Goodfellow et al. (2016) remark that Baldi & Hornik (1989) show that every local minimum is a global minimum for shallow linear networks. Kawaguchi (2016) extends this result by showing that every local minimum is a global minimum in the deep linear network. Lu & Kawaguchi (2017) advance the result by relaxing the assumption. Zhou & Liang (2018) provides an analytic formulation of critical

points in a deep linear network. Laurent & Brecht (2018) show that every local minimum is global even for any convex loss function. Soltanolkotabi et al. (2018) shows every local minimum is global for the quadratic activations.
For the piece-wise linear activations, Yun et al. (2018) show that bad local minima exist in the deep neural network with two-piece linear activation. Goldblum et al. (2019) show a local minimum exists in L-layer network with Affine activations. He et al. (2020) generalize the results of Yun et al. (2018) by improving the two-piece linear to the piece-wise linear activation function, and the 2-layer to the general L-layer neural network.
For smooth activation functions, Petzka & Sminchisescu (2021) and Ding et al. (2022) show that a bad local minimum exists in the deep neural network with sigmoid activation functions.
For population loss, Safran & Shamir (2018) empirically show the existence of local minima in the 2-layer ReLU network. Wu et al. (2018) shows that there is no bad local minimum on the manifold $\|w_1\| = \|w_2\| = 1$.

2.2 Existence of Local Valley

From another perspective, there are several studies on local valleys (*i.e.* sub-level sets of loss surface). The absence of a bad local valley guarantees that there is a non-increasing path to a global minimum. Poston et al. (1991) show the existence of a non-increasing path to the global minimum for an extremely wide 2-layer network with sigmoid activations in probability. Venturi et al. (2019) show the existence of a non-increasing path to the global minimum for a 2-layer network with non-polynomial activations and $d_1 \geq N$. Nguyen (2019; 2021) show the existence of non-increasing path to the global minimum for general L-layer network if $d_1 \geq N$ and $d_1 > d_2 > ... > d_L$, and every sub-level set is connected if $d_1 \geq N + 1$.

3 Notation and setup

We begin by defining the notation. Let L be the number of layers. Let (X, Y) be the training dataset with $X \in \mathbb{R}^{d_X \times N}$, and $Y \in \mathbb{R}^{d_Y \times N}$, where N is the number of samples. d_X and d_Y denote the dimension of the inputs and outputs, respectively. $d_1, d_2, ...d_{L-1}$ denote the width of the i-th layer. $B(x_0, r)$ denotes a ball with a center at x_0 and a radius of r.

Consider the N-layer neural network and the weights $W = [W_1, b_1, ..., W_N, b_N] \in [\mathbb{R}^{d_1 \times d_X}, \mathbb{R}^{d_1}, ..., \mathbb{R}^{d_Y \times d_L}, \mathbb{R}^{d_N}]$ of the network:

$$F_1 = \sigma(W_1 X + b_1) \tag{1}$$

$$F_{j+1} = \sigma(W_{j+1} F_j) + b_{j+1} \tag{2}$$

$$\hat{Y}_i = F_L = W_L F_{L-1} + b_L. \tag{3}$$

Then define the empirical loss function of the network, $\mathcal{R}(W)$ as

$$\mathcal{R}(W) = L([W_1, b_1, ..., W_N, b_N]) = \sum_i \ell(Y_i, F_L), \tag{4}$$

where ℓ is the loss function.
We denote $\mathbf{n}(l, r)$ a neuron of the network with r-th index in layer l, and $\mathbf{n}(l, r; x_\alpha)$ the output value before the activation with input x_α. We denote $\mathbf{act}(l, r; x_\alpha)$ the output value after the activation with input x_α.
Let $A\{B_1, B_2, ..., B_n\}$ denote $AB_1, AB_2, ..., AB_n$.
If property P is generic if it holds almost everywhere (except measure zero). For instance, if P holds for generic $X \in \mathbb{R}^N$, then P holds for almost every $X \in \mathbb{R}^N$.

Additionally, we classify local minima as follows.

Definition 1. *Let $\hat{W}$ be a local minimum of the loss $\mathcal{L}$.*

- *A bad local minimum $\hat{W}$ is called attracting if there is no non-increasing path to the global minimum from $\hat{W}$.*

- *A bad local minimum $\hat{W}$ is called non-attracting if there exists a non-increasing path to the global minimum from $\hat{W}$*

- *A global minimum is considered attracting.*

4 EXISTENCE OF A LOCAL MINIMUM FOR PARTIALLY LINEAR ACTIVATIONS

As studied in the works (Yun et al., 2018; He et al., 2020), a local minimum can be constructed by borrowing the parameters from the linear model, using the local linearity of the piece-wise linear activation function. Similarly, we investigate the existence of a local minimum in the network with partially linear activation functions using the partial linearity of the activation functions.

Proposition 1. *Suppose σ is partially linear with $\sigma(x) = cx + d$ on a open interval (α, β). Then $\mathcal{R}(W)$ has a local minimum.*

Proof. The proof is provided in Appendix A.1.1. □

However, if $\sigma(x)$ has no linearity, the borrowing technique is no longer available. Therefore, we have to approach the problem in a new way.

5 EXISTENCE OF A LOCAL MINIMUM IN THE 2-LAYER NETWORKS FOR GENERAL N

5.1 EXISTENCE OF A LOCAL MINIMUM IN THE NARROW 2-LAYER NETWORKS

In this section, we study the existence of a bad local minimum on the loss surface of 2-layer networks with smooth activation functions for general N. First, consider the following 2-layer network of width 1.

$$F_2(W) = w_2\sigma(w_1x + b_1) + b_2. \tag{5}$$

We present the following assumptions on the activation functions. We borrow the idea of the assumption from the proof of Theorem 2 in Ding et al. (2022).

Assumption 1. *Assume $\sigma(x)$ is analytic and $B_2(x) = \{1, \sigma(x), \sigma'(x), x\sigma'(x), \sigma''(x), x\sigma''(x), x^2\sigma''(x)\}$ is linearly independent. In fact, this implies that $\sigma(x)$ is not a solution to any second order linear ODE with polynomial coefficient of the following form:*

$$(Ax^2 + Bx + C)y'' + (Dx + E)y' + Fy + G = 0, \tag{6}$$

where A, B, C, D, E, F, and G are scalar and are not zero at the same time.

Assumption 1 describes the linear independence of the set $B_2(x)$ to be used in proving Proposition 2. Although the statement of Assumption 1 might seem unfamiliar, we discover that it is actually a mild assumption that holds for almost all analytic functions (in sense of measure). In fact, we prove that the most of widely used activation functions such as Tanh, Sigmoid, SiLU (Elfwing et al., 2018), SoftPlus, GELU (Hendrycks & Gimpel, 2016), Swish (Ramachandran et al., 2017), and Mish (Misra, 2019) actually satisfy Assumption 1 (Lemma 1). Therefore, Assumption 1 can be considered to cover almost all smooth activation functions widely used in machine learning so far. To show linear Independence, we use the more generalized concept, algebraic independence, in other words, transcendence. The disciple of differential Galois theory studies the algebraic properties of functions with derivations. Please see Appendix B for explanations for more details.

Lemma 1. *Tanh, Sigmoid, SiLU, SoftPlus, GELU, Swish, and Mish activation functions satisfy Assumption 1.*

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{7}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

$$SiLU(x) = \frac{x}{1 + e^{-x}} \tag{9}$$

$$SoftPlus(x) = \log(1 + e^x) \tag{10}$$

$$GELU(x) = \frac{x}{2}(1 + erf(\frac{x}{\sqrt{2}})) = \frac{x}{2}(1 + \frac{2}{\sqrt{\pi}} \int_0^{x/\sqrt{2}} e^{-t^2} dt) \tag{11}$$

$$Swish(X) = x \cdot sigmoid(\beta x) \tag{12}$$

$$Mish(x) = x \cdot tanh(softPlus(x)) = x \cdot tanh(\log(1 + e^x)). \tag{13}$$

Generally, it holds for almost every analytic function (i.e. except measure zero).

Proof. In this proof, we present the proof for $\sigma(x) = sigmoid(x)$. Please see Appendix A.2.1 for the other activation functions.

Suppose $\sigma(x)$ is sigmoid. Suppose $\sigma(x)$ is a solution of some second order linear ODE of form equation 6. Then since

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)),\ \sigma''(x) = \sigma(x)(1 - \sigma(x))(1 - 2\sigma(x)),$$

we have

$$\begin{aligned}
&(Ax^2 + Bx + C)(\sigma(x)(1 - \sigma(x))(1 - 2\sigma(x))) + (Dx + E)\sigma(x)(1 - \sigma(x)) + F\sigma(x) + G \\
=&(2Ax^2 + 2Bx + 2C)(\sigma(x))^3 + (-3Ax^2 + (-3B - D)x + (-3C - E))(\sigma(x))^2 + \\
&(Ax^2 + (B + D)x + (C + E + F))(\sigma(x)) + G = 0.
\end{aligned}$$

Since $\sigma(x)$ can be viewed as the root of a cubic equation with polynomial coefficients, we can consider the field extension of the quotient field of the polynomial ring. Let Q denote the quotient field of the polynomial ring and $Q(\sigma(x))$ be an extension field of Q with $\sigma(x)$. Let L be a Galois extension of Q including element $\sigma(x)$. Since $\sigma(x)$ is the root of the cubic equation, the degree of field extension is finite.

$$[Q(\sigma(x)) : Q] \leq [L : Q] \leq |S_3| = 6.$$

However, since $\sigma(x) = \frac{1}{1+e^{-x}}$ is transcendental function, it cannot be expressed in terms of a finite sequence of algebraic operations, hence

$$[Q(\sigma(x)) : Q] = \infty.$$

This is a contradiction, therefore the sigmoid function satisfies Assumption 1.

To show that Assumption 1 holds for almost every analytic function, we utilize the following Picard–Lindelöf theorem, which guarantees the uniqueness of the solution of the ODE.

Theorem 1 (Picard–Lindelöf theorem)**.** *Let $D \subseteq \mathbb{R} \times \mathbb{R}^n$ be a closed rectangle with $(t_0, y_0) \in D$ $(t_0, y_0) \in D$. Let $f : D \to \mathbb{R}^n$ be a function that is continuous in t and Lipschitz continuous in y. Then, there exists some $\epsilon > 0$ such that the initial value problem*

$$y'(t) = f(t, y(t)), \qquad y(t_0) = y_0. \tag{14}$$

has a unique solution $y(t)$ on the interval $[t_0 - \varepsilon, t_0 + \varepsilon]$

By taking $\mathbf{y} = (y(t), y'(t))$, the ODE of the form 6 has a unique solution with the initial value $(y(t_0), y'(t_0)) = (y_0, y_0') \in \mathbb{R} \times \mathbb{R}$. Since the ODE 6 has 7 degrees of freedom, the solution set of 6 has 14 degrees of freedom, *i.e*, is isomorphic to $\mathbb{R}^{14}$. Because the space of the analytic functions has infinite degrees of freedom, *i.e.*, isomorphic to $\mathbb{R}^{\mathbb{Z}_{\geq 0}}$, we conclude that the solutions set has measure zero in the space of analytic functions. □

Remark 1. *The exception set of Assumption 1 is the set of solutions of equation 6, including e^x, $sin(x)$. Since e^x, $sin(x)$ are solutions of $y' = y, y'' = -y$, they do not satisfy Assumption 1.*

Now we show that there exists a strict local minimum of the 2-layer neural network of width 1 by constructing suitable $Y \in \mathbb{R}^{1\times N}$ for generic $X \in \mathbb{R}^{1\times N}$.

Proposition 2. *Consider a 2-layer network with $N \geq 7$, and $d_X = d_1 = d_Y = 1$. Suppose ℓ is L^2 loss function, and $\sigma(x)$ satisfies Assumption 1. Then $\mathcal{R}(W)$ has a strict local minimum $\hat{W}$ for a generic dataset $X \in \mathbb{R}^{1\times N}$ with $\mathcal{R}(\hat{W}) > 0$. Moreover, Hessian matrix $H(\hat{W})$ of $\mathcal{R}(W)$ at $\hat{W}$ is strictly positive definite.*

Proof. The proof is provided in Appendix A.2.2. □

Now we have a pair (X, Y) where the neural networks can have a local minimum. By giving a small perturbation to (X, Y), we attain a perturbed $(\tilde{X}, \tilde{Y})$. If the neural network have a local minimum for the perturbed $(\tilde{X}, \tilde{Y})$, we can conclude that the network has a local minimum for positive measure of data, which is realistic condition. To show this, we need the following lemma.

Lemma 2. *Let $F(a, b)$ be a smooth function for $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$. Suppose for fixed $b_0 \in \mathbb{R}^n$, $F(\cdot, b_0)$ has a strict local minimum at $a = a_0$ and the Hessian $[\frac{\partial^2 F}{\partial a_i \partial a_j}](a_0, b_0)$ is strictly positive definite. Then for any $\epsilon > 0$, there exists $\delta > 0$, such that for any $\tilde{b} \in B(b_0, \delta)$, $F(\cdot, \tilde{b})$ has a strict local minimum at some $a = \tilde{a} \in B(a_0, \epsilon)$ and the Hessian $[\frac{\partial^2 F}{\partial a_i \partial a_j}](\tilde{a}, \tilde{b})$ is strictly positive definite.*

Proof. The proof is provided in Appendix A.2.3. □

By setting $a = W, b = (X, Y)$ in Lemma 2, together with Proposition 2, we show that a strict local minimum exists for data of positive measure.

Proposition 3. *Consider a 2-layer network with $N \geq 7$, and $d_X = d_1 = d_Y = 1$. Suppose ℓ is L^2 loss function, and $\sigma(x)$ satisfies Assumption 1. Then there exists a positive measure of $X \in \mathbb{R}^{1\times N}$ and $Y \in \mathbb{R}^{1\times N}$, such that $\mathcal{R}(W)$ has a strict local minimum with strictly positive definite Hessian matrix.*

Proof. Define $F(W, (X, Y)) = \|F_2(W)(X) - Y\|_2^2$. For a 2-layer network, by Proposition 2, there exists $\hat{W}$ and (X_0, Y_0) such that $F(\cdot, (X_0, Y_0))$ has a strict local minimum at $W = \hat{W}$ with the strictly positive definite Hessian $[\frac{\partial^2 F}{\partial W_i \partial W_j}](\hat{W}, (X_0, Y_0))$. By Lemma 2, for any $\epsilon > 0$ there exists $\delta > 0$, such that for any $(\tilde{X}, \tilde{Y}) \in B((X_0, Y_0), \delta)$, $F(\cdot, (\tilde{X}, \tilde{Y}))$ has a strict local minimum at some $\tilde{W} \in B(\hat{W}, \epsilon)$ and the Hessian $[\frac{\partial^2 F}{\partial W_i \partial W_j}](\tilde{W}, (\tilde{X}, \tilde{Y}))$ is strictly positive definite. □

5.2 Local minimum embedding and a bad local minimum in the wider networks

In the previous subsection, we show the existence of a local minimum in 2-layer networks of width 1. In this subsection, we introduce a technique called local minimum embedding which embeds a local minimum from the smaller network into the larger network. Using this technique, we show that a bad local minimum can exist in the wider networks. We introduce the definition of the local minimum embedding.

Definition 2 (Local minimum embedding)**.** *Consider a L-layer neural network $F^{small}(X)$ with the width $d_1, d_2, ...d_L$ and the weights W^{small}. Consider a neuron $\mathbf{n}(l, r)$ with index r in layer l. Let $[u_{r,i}^{small}]_{i=1}^{d_{l-1}}$ be the incoming weights into $\mathbf{n}(l, r)$ and $[v_{s,r}^{small}]_{s=1}^{d_{l+1}}$ be the outgoing weights of $\mathbf{n}(l, r)$. The weights of the smaller network W^{large} can be represented as*

$$W^{small} = ([u_{r,i}^{small}]_{i=1}^{d_{l-1}}, [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \bar{W}^{small}), \tag{15}$$

where $\bar{W}^{small}$ denote the collection of all remaining weights of the smaller network.
Consider the larger network $F_{large}(X)$ by adding a new neuron $\mathbf{n}(l, -1)$ referring $\mathbf{n}(l, r)$ with new

weights $[u_{-1,i}^{large}]_{i=1}^{d_{l-1}}$ *and* $[v_{s,-1}^{large}]_{s=1}^{d_{l+1}}$. *The weights of the larger network* W^{large} *can be represented as*

$$W^{large} = ([u_{-1,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,-1}^{large}]_{s=1}^{d_{l+1}}, [u_{r,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,r}^{large}]_{s=1}^{d_{l+1}}, \bar{W}^{large}). \tag{16}$$

Note that $[u_{-1,i}^{large}]_{i=1}^{d_{l-1}}$ *and* $[v_{s,-1}^{large}]_{s=1}^{d_{l+1}}$ *are the incoming and outgoing weights of the new neuron* $\mathbf{n}(l,-1)$ *in the larger network.*
Then define the local minimum embedding function γ_λ^r *mapping* W^{small} *to* W^{large} *as*

$$\gamma_\lambda^r([u_{r,i}^{small}]_{i=1}^{d_{l-1}}, [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \bar{W}^{small}) = ([u_{-1,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,-1}^{large}]_{s=1}^{d_{l+1}}, [u_{r,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,r}^{large}]_{s=1}^{d_{l+1}}, \bar{W}^{large})$$

with

$$[u_{-1,i}^{large}]_{i=1}^{d_{l-1}} = [u_{r,i}^{small}]_{i=1}^{d_{l-1}},\ [v_{s,-1}^{large}]_{s=1}^{d_{l+1}} = \lambda [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \tag{17}$$

$$[u_{r,i}^{large}]_{i=1}^{d_{l-1}} = [u_{r,i}^{small}]_{i=1}^{d_{l-1}},\ [v_{s,r}^{large}]_{s=1}^{d_{l+1}} = (1-\lambda)[v_{s,r}^{small}]_{s=1}^{d_{l+1}},\ \bar{W}^{large} = \bar{W}^{small} \tag{18}$$

The key idea of the local minimum embedding is to construct the larger network that works the same as the smaller network.

Remark 2. *By definition, the smaller network and the larger network have the same output.*

$$F_{large}(x) = F_{small}(x) \tag{19}$$

for all $x \in \mathbb{R}^{d_X}$. *This is independent of the embedding parameter* $\lambda \in \mathbb{R}$.

Under what conditions is the local minimum in the smaller network still the local minimum in the larger network? The following theorem describes the condition.

Theorem 2 ((Petzka & Sminchisescu, 2021))**.** *Define the matrices* $\mathfrak{B}_{i,j}$ *and* $\mathfrak{D}_i^{r,s}$ *as*

$$\mathfrak{B}_{i,j} = \sum_{\alpha=1}^{N} \sum_{k=1}^{d_{l+1}} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,k;x_\alpha)} \cdot v_{k,r} \cdot \sigma''(\mathbf{n}(l,r;x_\alpha))\mathbf{act}(l-1,i;x_\alpha)\mathbf{act}(l-1,j;x_\alpha), \tag{20}$$

and

$$\mathfrak{D}_i^{r,s} := \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,s;x_\alpha)} \sigma'(\mathbf{n}(l,r;x_\alpha))\mathbf{act}(l-1,i;x_\alpha). \tag{21}$$

Then, assume $\mathfrak{B}_{i,j}$ *is either*

- *positive definite and* $\lambda \in (0,1)$, *or*
- *negative definite and* $\lambda \in (-\infty, 0) \cup (1, \infty)$.

Then the embedding $\gamma_\lambda^r(\cdot)$ *determines a local minimum in the larger network if and only if* $D_i^{r,s} = 0$ *for all* i, s.

Now, together with Proposition 3, we show that a bad local minimum exists in the wide network.

Theorem 3. *Suppose* $d_X = d_Y = 1, d_1 \geq 2,\ N \geq 7$, ℓ *is* L^2 *function and the activation function* $\sigma(x)$ *satisfies Assumption 1. Then there exists a positive measure of* $X \in \mathbb{R}^{1\times N}$ *and* $Y \in \mathbb{R}^{1\times N}$, *such that the network* $1 - d_1 - 1$ *has a bad local minimum.*

Proof. The proof is provided in Appendix A.2.4. □

Remark 3. *The* $\hat{W}^{d_1}$ *found in Theorem 3 is the local minimum by definition (i.e.,* $\mathcal{R}(\hat{W}^{d_1}) \leq \mathcal{R}(\tilde{W})$ *for the neighborhood* $\tilde{W}$*). However, its type is actually non-attracting (i.e., there exists non-increasing path to the global minimum.)*

6 EXISTENCE OF A LOCAL MINIMUM IN 3-LAYER NETWORKS FOR GENERAL N

So far, we construct a local minimum for the 2-layer neural network. In the next section, we stretch the existence of a local minimum to the 3-layer network. Unlike the 2-layer, the problem becomes very complex because of the composition of $\sigma(x)$. In the case of 3-layer, the assumption becomes:

Assumption 2. *Assume $\sigma(x)$ is analytic and satisfies Assumption 1. Additionally, assume that $\sigma(x)$ satisfies the following claims.*

Claim 1. *$\sigma(\sigma(x))$ is transcendental over $Q(\sigma(x))$.*

Claim 2. *If*

$$\alpha_1 + \alpha_2\sigma(\sigma(x)) + \alpha_3\sigma'(\sigma(x)) + \alpha_4\sigma''(\sigma(x)) = 0 \tag{22}$$

for some $\alpha_i \in Q(\sigma(x))$, then $\alpha_i = 0$ for all i.

Lemma 3. *Assume $\sigma(x)$ satisfies Assumption 2. Define $B_3(x), B_{3,1}, B_{3,2}$ as*

$$\begin{aligned}B_3(x) = \{&1, \sigma(\sigma(x)), \sigma'(\sigma(x))\{1, \sigma(x), \sigma'(x)\{1, x\}, \sigma''(x)\{1, x, x^2\}\},\\ &\sigma''(\sigma(x))\{1, \sigma(x), \sigma(x)^2, \sigma'(x)\{1, x\}, \sigma''(x)\{1, x, x^2\}, \sigma'(x)^2\{1, x, x^2\},\\ &\sigma(x)\sigma'(x)\{1, x\}, \sigma(x)\sigma''(x)\{1, x, x^2\}, \sigma'(x)'\sigma''(x)\{1, x, x^2, x^3\}\}, \sigma''(x)^2\{1, x, x^2, x^3, x^4\}\},\end{aligned}$$

$$B_{3,1}(x) = \{\sigma'(\sigma(x))\sigma''(x)x^2\}, \tag{23}$$

$$B_{3,2}(x) = \{\sigma'(\sigma(x))\sigma''(x)\}, \tag{24}$$

$$B_{3,3}(x) = \{\sigma''(\sigma(x))\}, \tag{25}$$

and $\tilde{B_3}(x) = B_3(x) - B_{3,1}(x) - B_{3,2}(x) - B_{3,3}(x)$.
Then, we have

$$span\{B_{3,i}(x)\} \cap span\{\tilde{B}(x)\} = \{0\} \tag{26}$$

$$span\{B_{3,i}(x)\} \cap span\{B_{3,j}(x)\} = \{0\}, \tag{27}$$

for $i, j = 1, 2, 3$ and $i \neq j$.

Proof. The proof is provided in Appendix A.2.5. □

We need to check that .

Lemma 4. *Tanh, Sigmoid, SiLU, SoftPlus, GELU, Swish, and Mish functions satisfy Assumption 2.*

Proof. The proof is provided in Appendix A.2.6. □

We believe Assumption 2 holds for almost every analytic function, but due to the mathematical difficulties with the composition of functions, we leave it a conjecture.

Conjecture 1. *Assumption 2 holds for almost every analytic function (i.e except measure zero).*

Now we show that there exists a strict local minimum in the 3-layer neural network of width 1. Together with Lemma 2, a strict local minimum exists for data of positive measure.

Proposition 4. *Consider a 3-layer network with $N \geq 34$, and $d_X = d_1 = d_2 = d_Y = 1$. Suppose ℓ is L^2 loss function, and $\sigma(x)$ satisfies Assumption 1 and 2. Then there exists a positive measure of $X \in \mathbb{R}^{1\times N}$ and $Y \in \mathbb{R}^{1\times N}$, such that $\mathcal{R}(W)$ has a strict local minimum $\hat{W}$. Moreover, Hessian matrix $H(\hat{W})$ of $\mathcal{R}(W)$ at $\hat{W}$ is strictly positive definite.*

Proof. The proof is provided in Appendix A.2.7. □

Similar to the 2-layer network, using the local minimum embedding, we show that a bad local minimum exists in the wide 3-layer network.

Theorem 4. *Suppose $d_X = d_Y = 1, d_1 \geq 2, d_2 \geq 2,\ N \geq 34$, ℓ is L^2 function and the activation function $\sigma(x)$ satisfies Assumption 1 and 2. Then there exists a positive measure of $X \in \mathbb{R}^{1\times N}$ and $Y \in \mathbb{R}^{1\times N}$, such that the network $1 - d_1 - d_2 - 1$ has a bad local minimum.*

Proof. The proof is provided in Appendix A.2.8. □

7 CONCLUSION AND FUTURE WORK

We investigate the existence of a bad local minimum in the 2-layer and 3-layer neural networks with general smooth activations. In addition, We present techniques for handling the smooth activation functions. Together with the studies on the piece-wise linear activations, it is verified that the neural networks with widely-used activation functions such as ReLU, SiLU, GELU, Swish, and Mish can have a bad local minimum. A future research direction will be investigating the existence of a bad local minimum in the general L-layer network. Although several empirical pieces of evidence suggest that an attracting local minimum actually exists in many cases when the width is less than N, there it has not been theoretically elucidated. One can try to show the existence of an attracting bad local minimum and bad local valley in the neural network with smooth activations when $d_1 < N$.

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.

RC Churchill. Liouville's theorem on integration in terms of elementary functions. In *posted on the website of the Kolchin Seminar in Differential Algebra*. Citeseer, 2006.

Tian Ding, Dawei Li, and Ruoyu Sun. Suboptimal local minima exist for wide neural networks with smooth activations. *Mathematics of Operations Research*, 2022.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein. Truth or backpropaganda? an empirical investigation of deep learning theory. *arXiv preprint arXiv:1910.00359*, 2019.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Fengxiang He, Bohan Wang, and Dacheng Tao. Piecewise linear activations substantially shape the loss surfaces of neural networks. *arXiv preprint arXiv:2003.12236*, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

John H Hubbard and Benjamin E Lundell. A first look at differential algebra. *The American Mathematical Monthly*, 118(3):245–261, 2011.

Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Thomas Laurent and James Brecht. The multilinear structure of relu networks. In *International conference on machine learning*, pp. 2908–2916. PMLR, 2018.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 4(2):10–48550, 2019.

Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pp. 15959–15975. PMLR, 2022.

Quynh Nguyen. On connected sublevel sets in deep learning. In *International Conference on Machine Learning*, pp. 4790–4799. PMLR, 2019.

Quynh Nguyen. A note on connectivity of sublevel sets in deep learning. *arXiv preprint arXiv:2101.08576*, 2021.

Tohru Nitta. Resolution of singularities introduced by hierarchical structure in deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2282–2293, 2016.

Henning Petzka and Cristian Sminchisescu. Non-attracting regions of local minima in deep and wide neural networks. *J. Mach. Learn. Res.*, 22:143–1, 2021.

Timothy Poston, C-N Lee, Y Choie, and Yonghoon Kwon. Local minima and back propagation. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pp. 173–176. IEEE, 1991.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Marius Van der Put and Michael F Singer. *Galois theory of linear differential equations*, volume 328. Springer Science & Business Media, 2012.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20:133, 2019.

Chenwei Wu, Jiajun Luo, and Jason D Lee. No spurious local minima in a two hidden unit relu network. 2018.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

Yi Zhou and Yingbin Liang. Critical points of linear neural networks: Analytical forms and landscape properties. In *International Conference on Learning Representations*, 2018.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in sgd. *arXiv preprint arXiv:2102.05375*, 2021.

A APPENDIX

In this section, we provide detailed proofs of the theorems, propositions, and lemmas.

A.1 PROOFS IN SECTION 4

A.1.1 PROOF OF PROPOSITION 1

Proposition. *Suppose σ is partially linear with $\sigma(x) = cx + d$ on a open interval (α, β). Then $\mathcal{R}(W)$ has a local minimum.*

Proof. Consider the linear minimization problem

$$\sum_i \ell(Y_i, W \begin{bmatrix} X_i \\ 1 \end{bmatrix}). \tag{28}$$

Let $\bar{W}$ be a linear minimizer of the above problem.
Let $\bar{Y} \in \mathbb{R}^{d_Y}$ be the output of linear model and M, m be the maximum and minimum value of $\{\bar{Y}_i\}_{i=1}^N$

$$\bar{Y}_i = \bar{W} \begin{bmatrix} X_i \\ 1 \end{bmatrix}$$

$$M := \max_i \max(\bar{Y}_i)$$

$$m := \min_i \min(\bar{Y}_i).$$

Then we can find $f(x) = px + q,\ p \neq 0,\ p, q \in \mathbb{R}$ such that,

$$f(M), f(m) \in (\alpha, \beta).$$

Let define the weight $\hat{W} = [\hat{W}_1, \hat{b}_1, \hat{W}_2, \hat{b}_2]$ of 2-layer network as :

$$\hat{W}_1 = \begin{bmatrix} f(\bar{W}_{[1:d_X]}) \\ \mathbf{0} \end{bmatrix}, \hat{b}_1 = \begin{bmatrix} f([\bar{W}]_{[d_X+1]}) \\ \mathbf{0} \end{bmatrix} \tag{29}$$

$$\hat{W}_2 = \begin{bmatrix} (cp)^{-1} I_{d_Y} & 0 \end{bmatrix}, \hat{b}_2 = (-p^{-1}q - (cp)^{-1}d)\mathbf{1}. \tag{30}$$

Since $\hat{W}_1 X_i + \hat{b}_1 = f(\bar{Y}_i) \in [\alpha, \beta]$, we have $\sigma(\hat{W}_1 X_i + \hat{b}_1) = cf(\bar{Y}_i) + d$. Then we claim $\hat{W}$ is the local minimum. To show this, we introduce the small disturbance $\delta_W = (\delta_{w_1}, \delta_{b_1}, \delta_{w_2}, \delta_{b_2})$. Then, since $(\hat{W}_1 + \delta_{W_1})X_i + \hat{b}_1 + \delta_{b_1} \in (\alpha, \beta)$,

$$\sigma((\hat{W}_1 + \delta_{W_1})X_i + \hat{b}_1 + \delta_{b_1}) = \sigma(\begin{bmatrix} p\bar{Y}_i + q\mathbf{1}_{d_Y} \\ \mathbf{0} \end{bmatrix} + \delta_{W_1} X_i + \delta_{b_1}) = c(\begin{bmatrix} p\bar{Y}_i + q\mathbf{1}_{d_Y} \\ \mathbf{0} \end{bmatrix} + \delta_{W_1} X_i + \delta_{b_1}) + d$$

$$F_2 = (\hat{W}_2 + \delta_{W_2})(\begin{bmatrix} cp\bar{Y}_i + cq\mathbf{1}_{d_Y} \\ \mathbf{0} \end{bmatrix} + c\delta_{W_1} X_i + c\delta_{b_1} + d) + \delta_{b_2}$$

$$= \bar{Y}_i + [p^{-1}\delta_{W_1} X_i + p^{-1}\delta_{b_1}]_{[:d_Y]} + \delta_{W_2}([cp\bar{Y}_i + cq\mathbf{1}_{d_Y}]_{[:d_Y]} c\delta_{b_1} + d) + c\delta_{W_2}\delta_{W_1} X_i + \delta_{b_2}$$

$$= \bar{Y}_i + \delta \begin{bmatrix} X_i \\ 1 \end{bmatrix}, \tag{31}$$

where $\delta = [\delta_1, \delta_2]$ and

$$\delta_1 = p^{-1}\delta_{W_1[:,:d_Y]} + c\delta_{W_2}\delta_{W_1} \tag{32}$$

$$\delta_2 = p^{-1}\delta_{b_1[:,:d_Y]} + \delta_{W_2}([cp\bar{Y}_i + cq\mathbf{1}_{d_Y}]_{[:d_Y]} c\delta_{b_1} + d) + \delta_{b_2}. \tag{33}$$

Because, $\mathcal{R}(\hat{W}) = \mathcal{R}_{linear}(\bar{W}) < \mathcal{R}_{linear}(\bar{W} + \delta) = \mathcal{R}(\hat{W} + \delta_W)$, we can conclude that $\hat{W}$ is the local minimum. □

A.2 PROOFS IN SECTION 5 AND 6

A.2.1 PROOF OF LEMMA 1

Lemma. *Tanh, Sigmoid, SiLU, SoftPlus, GELU, Swish, and Mish activation functions satisfy Assumption 1.*

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{34}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{35}$$

$$SiLU(x) = \frac{x}{1 + e^{-x}} \tag{36}$$

$$SoftPlus(x) = \log(1 + e^x) \tag{37}$$

$$GELU(x) = \frac{x}{2}(1 + erf(\frac{x}{\sqrt{2}})) = \frac{x}{2}(1 + \frac{2}{\sqrt{\pi}} \int_0^{x/\sqrt{2}} e^{-t^2} dt) \tag{38}$$

$$Swish(X) = x \cdot sigmoid(\beta x) \tag{39}$$

$$Mish(x) = x \cdot tanh(softPlus(x)) = x \cdot tanh(\log(1 + e^x)). \tag{40}$$

Generally it holds for almost every analytic function (i.e. except measure zero).

Proof. Suppose $\sigma(x)$ is sigmoid. Suppose $\sigma(x)$ is a solution of some second order linear ODE of form equation 6. Then since

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)),\ \sigma''(x) = \sigma(x)(1 - \sigma(x))(1 - 2\sigma(x)),$$

we have

$$\begin{aligned}
&(Ax^2 + Bx + C)(\sigma(x)(1 - \sigma(x))(1 - 2\sigma(x))) + (Dx + E)\sigma(x)(1 - \sigma(x)) + F\sigma(x) + G \\
=&(2Ax^2 + 2Bx + 2C)(\sigma(x))^3 + (-3Ax^2 + (-3B - D)x + (-3C - E))(\sigma(x))^2 + \\
&(Ax^2 + (B + D)x + (C + E + F))(\sigma(x)) + G = 0.
\end{aligned}$$

Since $\sigma(x)$ can be viewed as the root of a cubic equation with polynomial coefficients, we can consider the field extension of the quotient field of polynomial ring. Let Q denote the quotient field of polynomial ring and $Q(\sigma(x))$ be a extension field of Q with $\sigma(x)$. Let L be a Galois extension of Q including element $\sigma(x)$. Since $\sigma(x)$ is the root of cubic equation, the degree of field extension is finite.

$$[Q(\sigma(x)) : Q] \leq [L : Q] \leq |S_3| = 6.$$

However, since $\sigma(x) = \frac{1}{1+e^{-x}}$ is transcendental function, it cannot be expressed in terms of a finite sequence of the algebraic operations, hence

$$[Q(\sigma(x)) : Q] = \infty.$$

This is a contradiction, therefore the sigmoid function satisfies Assumption 1.
Similarly, because $tanh(x) = 2sigmoid(2x) - 1$, the tanh function satisfies Assumption 1.

For $\sigma(x) = SiLU(x)$, since $SiLU(x)$ is a special case of $Swish(x)$, please refer to $Swish(x)$ part.

For $\sigma(x) = SoftPlus(x)$, we have $\sigma'(x) = s(x)$ where $s(x)$ denotes the sigmoid function. Then we have

$$\begin{aligned}
\sigma'(x) &= s(x) \\
\sigma''(x) &= s'(x) = s(x)(1 - s(x)).
\end{aligned}$$

By substituting into equation 6, we have

$$\sigma(x) = P_2(s(x)),$$

for some quadratic polynomial $P_2(X)$. By substituting into equation 6 again, we have

$$\sigma(x) = P_4(s(x)),$$

for some quartic polynomial $P_4(x)$. Similar to $SiLU$ case, we have a contradiction.
Therefore we conclude that for sigmoid activation function $s(x)$, if $\sigma(x)$ is a form of polynomial of $s(x)$, k-th derivative, k-th indefinite integral or their linear combination, *i.e*,

$$\sigma(x) \in span\{P_i(s(x)), s^{(k)}(x), \int^{(k)} s(s)dx^{(k)}\},$$

then $\sigma(x)$ satisfies Assumption 1.
For $\sigma(x) = GELU(x)$, note that

$$\sigma'(x) = \frac{1}{2}(1 + \text{erf}(\frac{x}{\sqrt{2}})) + x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{41}$$

$$\sigma''(x) = \frac{\sqrt{2}}{\sqrt{\pi}}e^{-\frac{x^2}{2}} - x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \tag{42}$$

Direct substitution into equation 6 induces

$$\begin{aligned}
&(Ax^2 + Bx + C)(\frac{\sqrt{2}}{\sqrt{\pi}}e^{-\frac{x^2}{2}} - x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}) \\
&+ (Dx + E)(\frac{1}{2}\text{erf}(\frac{x}{\sqrt{2}}) + x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} + \frac{1}{2}) + \frac{F}{2}(x\text{erf}(\frac{x}{\sqrt{2}}) + x) + G = 0 \qquad (43) \\
&\frac{1}{2}\text{erf}(\frac{x}{\sqrt{2}})(Dx + E + Fx) + \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}(-Ax^4 - Bx^3 + (2A - C + D)x^2 + (2B + E)x + 2C) \\
&+ (\frac{D + F}{2})x + \frac{E}{2} + G = 0. \qquad (44)
\end{aligned}$$

Since $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is not an elementary function by Liouville's theorem (Theorem 5), we have

$$(Dx + E + Fx) = 0.$$

Otherwise, $\text{erf}(x)$ would be an elementary function. Therefore we have

$$D = -F,\ E = 0.$$

Similarly, since $e^{-\frac{x^2}{2}}$ is transcendental over Q, we have

$$A = 0,\ B = 0,\ (2A - C + D) = 0,\ (2B + E) = 0,\ C = 0, D + F = 0,\ E = 0,\ G = 0.$$

Therefore every coefficient is zero. Hence we have a contradiction.
For $\sigma(x) = Swish(x)$, the proof is a generalized case of $SiLU(x)$. Let $\sigma(x) = xs(\beta x)$, where $s(x)$ denotes the sigmoid function. Then we have

$$\begin{aligned}
\sigma'(x) &= s(\beta x)(1 + \beta x(1 - s(\beta x))) \\
\sigma''(x) &= \beta s(\beta x)(1 - s(\beta x))(2 + \beta x - 2\beta x s(\beta x)).
\end{aligned}$$

By substituting into equation 6, we get the cubic equation with polynomial coefficient, of which $s(x)$ is a solution. Similarly, let L be a Galois extension of Q including element $s(x)$, the degree of field extension is

$$[Q(s(x)) : Q] \leq [L : Q] \leq |S_3| = 6.$$

Since $s(x)$ is transcendental, this is a contraction.

For $\sigma(x) = Mish(x) = x \cdot tanh(\log(1+e^x))$, it is sufficient to show that $\tau(x) := tanh(\log(1+e^x))$ is not a solution of the following second order linear ODE

$$(Ax^3 + Bx^2 + Cx + D)y'' + (Ex^2 + Fx + G)y' + (Hx + I)y + J = 0. \tag{45}$$

To show this, let $f(x) = \log(1 + e^x)$ and $s(x) = sigmoid(x)$. Since $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}, e^x = \frac{s(x)}{s(x)-1}$, we have

$$\tanh(f(x)) = \frac{(1+e^x)^2 - 1}{(1+e^x)^2 + 1} = \frac{(1 + \frac{s(x)}{s(x)-1})^2 - 1}{(1 + \frac{s(x)}{s(x)-1})^2 + 1} = \frac{3s(x)^2 - 2s(x)}{5s(x)^2 - 6s(x) + 2}. \tag{46}$$

Then we have

$$\tau'(x) = (\frac{3s(x)^2 - 2s(x)}{5s(x)^2 - 6s(x) + 2})' = \frac{4s(s-1)^2(2s-1)}{(5s^2 - 6s + 2)^2} \tag{47}$$

$$\tau''(x) = \frac{-4s(s-1)^2(s^3 + 7s^2 - 8s + 2)}{(5s^2 - 6s + 2)^3}. \tag{48}$$

By substituting into equation 45, $\sigma(x)$ is a solution of the equation of degree 6. Since $s(x)$ is transcendental, this is a contraction. □

A.2.2 Proof of Proposition 2

Proposition. *Consider a 2-layer network with $N \geq 7$, and $d_X = d_1 = d_Y = 1$. Suppose ℓ is L^2 loss function, and $\sigma(x)$ satisfies Assumption 1. Then $\mathcal{R}(W)$ has a strict local minimum $\hat{W}$ for a generic dataset $X \in \mathbb{R}^{1\times N}$ with $\mathcal{R}(\hat{W}) > 0$. Moreover, Hessian matrix $H(\hat{W})$ of $\mathcal{R}(W)$ at $\hat{W}$ is strictly positive definite.*

Proof. By assumption 1, $\{1, \sigma(x), \sigma'(x), x\sigma'(x), \sigma''(x), x\sigma''(x), x^2\sigma''(x)\}$ is linearly independent. Let $A(X) = \{\mathbf{1}_N, [\sigma(x_i)], [\sigma'(x_i)], [x_i\sigma'(x_i)], [\sigma''(x_i)], [x_i\sigma''(x_i)], [x_i^2\sigma''(x_i)]\}$. Consider the mapping

$$(x_1, x_2, ..., x_7) \mapsto det(A([x_i]_{i=1}^7)). \tag{49}$$

Because of linear independence, this map is not zero map. Since $\sigma(x)$ is analytic, this map is also analytic. Therefore, zero set of the map has measure zero.
Therefore, for generic $X = (x_1, x_2, ...x_N)$, $rank(A([x_i]_{i=1}^N)) \geq 7$. we have seven linearly independent N-dimensional vectors

$$\{\mathbf{1}_N, [\sigma(x_i)], [\sigma'(x_i)], [x_i\sigma'(x_i)], [\sigma''(x_i)], [x_i\sigma''(x_i)], [x_i^2\sigma''(x_i)]\}_{i=1}^N.$$

Then we can find $N - 5$ linearly independent N-dimensional vectors $\{v_k\}$ such that

$$\langle v_k, \mathbf{1}_N\rangle = 0 \tag{50}$$

$$\langle v_k, [\sigma(x_i)]\rangle = 0 \tag{51}$$

$$\langle v_k, [\sigma'(x_i)]\rangle = 0 \tag{52}$$

$$\langle v_k, [x_i\sigma'(x_i)]\rangle = 0 \tag{53}$$

$$\langle v_k, [x_i\sigma''(x_i)]\rangle = 0 \tag{54}$$

$$\langle v_k, [\sigma''(x_i)]\rangle > 0 \tag{55}$$

$$\langle v_k, [x_i^2\sigma''(x_i)]\rangle > 0. \tag{56}$$

Select data points $Y = [y_i]_{i=1}^N$ as

$$y_i = F_2(x_i) - w_2 \sum_{k=1}^{N-5} c_k [v_k]_i, \tag{57}$$

for some positive $c_k \in \mathbb{R}$.
Then pick $\hat{W} = (w_1, b_1, w_2, b_2) = (1, 0, w_2, 0)$, where $w_2 > 0$ can be fixed arbitrarily.

Define $\Delta y \in \mathbb{R}^N$ such that

$$[\Delta Y]_i = F_2(x_i) - y_i = w_2 \sum_{k=1}^{N-5} c_k [v_k]_i.$$

Then we get

$$\langle \Delta Y, \mathbf{1}_N \rangle = \langle \Delta Y, [\sigma(x_i)] \rangle = \langle \Delta Y, [\sigma'(x_i)] \rangle = \langle \Delta Y, [x_i \sigma'(x_i)] \rangle = \langle \Delta Y, [\sigma''(x_i)] \rangle = 0 \tag{58}$$

$$\langle \Delta Y, [x_i \sigma''(x_i)] \rangle > 0 \tag{59}$$

$$\langle \Delta Y, [x_i^2 \sigma''(x_i)] \rangle > 0. \tag{60}$$

Then for the loss $\mathcal{R} = \|F_2(x_i) - y_i\|_2^2 = \langle F_2(X) - Y, F_2(X) - Y \rangle$, derivatives are

$$\frac{\partial \mathcal{R}}{\partial w_1} = 2\langle \Delta Y, X\sigma'(X) \rangle = 0 \tag{61}$$

$$\frac{\partial \mathcal{R}}{\partial b_1} = 2\langle \Delta Y, \sigma'(X) \rangle = 0 \tag{62}$$

$$\frac{\partial \mathcal{R}}{\partial w_2} = 2\langle \Delta Y, \sigma(X) \rangle = 0 \tag{63}$$

$$\frac{\partial \mathcal{R}}{\partial b_2} = 2\langle \Delta Y, \mathbf{1} \rangle = 0. \tag{64}$$

Hence $\hat{W}$ is a stationary point. In addition, $\mathcal{R}(\hat{W}) = \langle \Delta Y, \Delta Y \rangle > 0$ by construction.

To show $\hat{W}$ is a strict local minimum, we need to show that Hessian $H(\hat{W})$ of $\mathcal{R}(W)$ is strictly positive definite at $W = \hat{W}$. Therefore we need to show that

$$\mathbf{u}^T H(\hat{W}) \mathbf{u} = \lim_{t \to 0} \frac{\mathcal{R}(\hat{W} + t\mathbf{u}) + \mathcal{R}(\hat{W} - t\mathbf{u}) - 2\mathcal{R}(\hat{W})}{t^2} > 0, \tag{65}$$

for all $\mathbf{u} = (u_{w_1}, u_{b_1}, u_{w_2}, u_{b_2})$ with $\|\mathbf{u}\|_2 = 1$.

First, note that

$$\begin{aligned}
&\mathcal{R}(\hat{W} + t\mathbf{u}) - \mathcal{R}(\hat{W}) = \langle F_2(\hat{W} + t\mathbf{u}) - Y, F_2(\hat{W} + t\mathbf{u}) - Y \rangle - \langle F_2(\hat{W}) - Y, F_2(\hat{W}) - Y \rangle \\
&= \langle F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W}), F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W}) \rangle + 2\langle \Delta Y, F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W}) \rangle \\
&= \|F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W})\|_2^2 + 2\langle \Delta Y, F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W}) \rangle,
\end{aligned} \tag{66}$$

and because $\hat{W} = (w_1, b_1, w_2, b_2) = (1, 0, w_2, 1)$,

$$\begin{aligned}
&F_2(\hat{W} + t\mathbf{u})(x) - F_2(\hat{W})(x) \\
&= ((w_2 + tu_{w_2})\sigma((w_1 + tu_{w_1})x + b_1 + tu_{b_1}) + b_2 + tu_{b_2}) - (w_2\sigma(w_1 x + b_1) + b_2) \\
&= (w_2 + tu_{w_2})(\sigma((1 + tu_{w_1})x + tu_{b_1}) - \sigma(x)) + tu_{w_2}\sigma(x) + tu_{b_2} \\
&= (w_2 + tu_{w_2})(\sigma(x + t(u_{w_1}x + u_{b_1})) - \sigma(x)) + tu_{w_2}\sigma(x) + tu_{b_2}.
\end{aligned} \tag{67}$$

Let $u_1 = u_{w_1}x + u_{b_1}$. By Taylor theorem, we have

$$\begin{aligned}
&\sigma(x + t(u_{w_1}x + u_{b_1})) - \sigma(x) = \sigma'(x)tu_1 + \frac{1}{2}\sigma''(x)t^2u_1^2 + o(t^2) \\
&= \sigma'(x)tu_{w_1}x + \sigma'(x)tu_{b_1} + \frac{1}{2}\sigma''(x)t^2(u_{w_1}^2x^2 + 2u_{w_1}u_{b_1}x + u_{b_1}^2) + o(t^2).
\end{aligned} \tag{68}$$

Therefore, we have

$$\begin{aligned}
&F_2(\hat{W} + t\mathbf{u})(x) - F_2(\hat{W})(x) \\
&= (w_2 + tu_{w_2})(\sigma'(x)tu_{w_1}x + \sigma'(x)tu_{b_1} + \frac{1}{2}\sigma''(x)t^2(u_{w_1}^2x^2 + 2u_{w_1}u_{b_1}x + u_{b_1}^2)) + tu_{w_2}\sigma(x) + tu_{b_2} + o(t^2).
\end{aligned} \tag{69}$$

By equation 66, we can calculate $\mathcal{R}(\hat{W} + t\mathbf{u}) + \mathcal{R}(\hat{W} - t\mathbf{u}) - 2\mathcal{R}(\hat{W})$.

$$\begin{aligned}
&\mathcal{R}(\hat{W} + t\mathbf{u}) + \mathcal{R}(\hat{W} - t\mathbf{u}) - 2\mathcal{R}(\hat{W}) = (\mathcal{R}(\hat{W} + t\mathbf{u}) - \mathcal{R}(\hat{W})) + (\mathcal{R}(\hat{W} - t\mathbf{u}) - \mathcal{R}(\hat{W})) \\
&= \|F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W})\|_2^2 + \|F_2(\hat{W} - t\mathbf{u}) - F_2(\hat{W})\|_2^2 + \\
&2\langle \Delta Y, F_2(\hat{W} + t\mathbf{u}) - F_2(\hat{W}) \rangle + 2\langle \Delta Y, F_2(\hat{W} - t\mathbf{u}) - F_2(\hat{W}) \rangle.
\end{aligned} \tag{70}$$

By equation 58 and equation 69, we have

$$\langle \Delta Y, F_2(\hat{W}+t\mathbf{u}) - F_2(\hat{W})\rangle = \\ \frac{1}{2}(w_2+tu_{w_2})t^2u_{w_1}^2\langle \Delta Y, \sigma''(X)X^2\rangle + \frac{1}{2}(w_2+tu_{w_2})t^2u_{b_1}^2\langle \Delta Y, \sigma''(X)\rangle + (w_2+tu_{w_2})\langle \Delta Y, o(t^2)\rangle. \tag{71}$$

We consider the following two cases.
Case 1: $(u_{w_1}, u_{b_1}) \neq (0,0)$.
In this case, by equation 71

$$2\langle \Delta Y, F_2(\hat{W}+t\mathbf{u}) - F_2(\hat{W})\rangle + 2\langle \Delta Y, F_2(\hat{W}-t\mathbf{u}) - F_2(\hat{W})\rangle \tag{72}$$
$$=(w_2+tu_{w_2})(t^2u_{w_1}^2\langle \Delta Y, \sigma''(X)X^2\rangle + t^2u_{b_1}^2\langle \Delta Y, \sigma''(X)\rangle) + o(t^2). \tag{73}$$

Since $\|F_2(\hat{W}+t\mathbf{u}) - F_2(\hat{W})\|_2^2 + \|F_2(\hat{W}-t\mathbf{u}) - F_2(\hat{W})\|_2^2 \geq 0$, we have

$$\lim_{t\to 0} \frac{\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W})}{t^2} \tag{74}$$
$$\geq \lim_{t\to 0} \frac{(w_2+tu_{w_2})(t^2u_{w_1}^2\langle \Delta Y, \sigma''(X)X^2\rangle + t^2u_{b_1}^2\langle \Delta Y, \sigma''(X)\rangle) + o(t^2)}{t^2} \tag{75}$$
$$= w_2(u_{w_1}^2\langle \Delta Y, \sigma''(X)X^2\rangle + u_{b_1}^2\langle \Delta Y, \sigma''(X)\rangle > 0. \tag{76}$$

Case 2: $(u_{w_1}, u_{b_1}) = (0,0)$.
In this case, we have $\|F_2(\hat{W}+t\mathbf{u}) - F_2(\hat{W})\|_2^2 + \|F_2(\hat{W}-t\mathbf{u}) - F_2(\hat{W})\|_2^2 = \|tu_{w_2}\sigma(X) + tu_{b_2}\mathbf{1}_N\|_2^2$ and $\langle \Delta Y, F_2(\hat{W}+t\mathbf{u}) - F_2(\hat{W})\rangle + \langle \Delta Y, F_2(\hat{W}-t\mathbf{u}) - F_2(\hat{W})\rangle = o(t^2)$. Since $\sigma(X)$ and $\mathbf{1}_N$ are independent, $\|u_{w_2}\sigma(X) + u_{b_2}\mathbf{1}_N\|_2^2 > 0$. Therefore, we have

$$\lim_{t\to 0} \frac{\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W})}{t^2} \tag{77}$$
$$= \lim_{t\to 0} \frac{t^2\|u_{w_2}\sigma(X) + u_{b_2}\mathbf{1}_N\|_2^2 + o(t^2)}{t^2} = \|u_{w_2}\sigma(X) + u_{b_2}\mathbf{1}_N\|_2^2 > 0. \tag{78}$$

In both cases, we have $\mathbf{u}^T H(\hat{W})\mathbf{u} > 0$ for all $\mathbf{u}$. Therefore, Hessian $H(\hat{W})$ is strictly positive definite at $\hat{W}$ and $\hat{W}$ is a strict local minimum of $\mathcal{R}(W)$. □

A.2.3 Proof of Lemma 2

Lemma. *Let $F(a,b)$ be a smooth function for $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$. Suppose for fixed $b_0 \in \mathbb{R}^n$, $F(\cdot, b_0)$ has a strict local minimum at $a = a_0$ and the Hessian $[\frac{\partial^2 F}{\partial a_i \partial a_j}](a_0, b_0)$ is strictly positive definite. Then for any $\epsilon > 0$, there exists $\delta > 0$, such that for any $\tilde{b} \in B(b_0, \delta)$, $F(\cdot, \tilde{b})$ has a strict local minimum at some $a = \tilde{a} \in B(a_0, \epsilon)$ and the Hessian $[\frac{\partial^2 F}{\partial a_i \partial a_j}](\tilde{a}, \tilde{b})$ is strictly positive definite.*

The first part of this proof follows the proof of Lemma B.2 in Ding et al. (2022).

Proof. Let $\epsilon > 0$ be given. Since a_0 is a strict local minimum of $F(\cdot, b_0)$, there exists ϵ_1 such that $F(\tilde{a}, b_0) > F(a, b_0)$, for all $\tilde{a} \in B(a_0, \epsilon_1)$.
Let $\epsilon_2 = \min\{\epsilon, \epsilon_1/2\}$. Then we have $F(\tilde{a}, b_0) > F(a, b_0)$ for all $a \in \partial\bar{B}(a_0, \epsilon_2)$. Then define

$$\eta = \inf_{\tilde{a}\in\partial\bar{B}(a_0,\epsilon_2)} F(\tilde{a}, b_0) - F(a_0, b_0) > 0,\ a^* = \arg\inf_{\tilde{a}\in\partial\bar{B}(a_0,\epsilon_2)} F(\tilde{a}, b_0) - F(a_0, b_0). \tag{79}$$

Since $\partial\bar{B}(a_0, \epsilon_2)$ is compact and F is continuous, such a^* exits. Then, since F is uniformly continuous on $\bar{B}((a_0, b_0), \epsilon_2)$, there exists $0 < \delta \leq \epsilon_2$ such that

$$|F(\alpha_1, \beta_1) - F(\alpha_2, \beta_2)| < \frac{\eta}{3}, \tag{80}$$

for all $(\alpha_1, \beta_1), (\alpha_2, \beta_2) \in \bar{B}((a_0, b_0), \epsilon_2)$ with $\|(\alpha_1, \beta_1) - (\alpha_2, \beta_2)\|_2 < \delta$. Then for any $\hat{b} \in B(b_0, \delta)$, we have

$$F(a_0, \hat{b}) < F(a_0, b_0) + \frac{\eta}{3} = (F(a^*, b_0) - \eta) + \frac{\eta}{3} = F(a^*, b_0) - \frac{2}{3}\eta \tag{81}$$

$$< F(a^*, \hat{b}) - \frac{1}{3}\eta < \inf_{\tilde{a} \in \partial \bar{B}(a_0, \epsilon_2)} F(\tilde{a}, \hat{b}). \tag{82}$$

Since $\bar{B}(a_0, \epsilon_2)$ is compact, $\hat{a} := \arg\inf_{\tilde{a} \in \bar{B}(a_0, \epsilon_2)} F(\tilde{a}, \hat{b})$ exists. Then, since

$$F(\hat{a}, \hat{b}) \leq F(a_0, \hat{b}) < \inf_{\tilde{a} \in \partial \bar{B}(a_0, \epsilon_2)} F(\tilde{a}, \hat{b}), \tag{83}$$

$\hat{a}$ is in interior of $\bar{B}(a_0, \epsilon_2)$. Therefore, there exists $\epsilon_3 > 0$ such that $\bar{B}(\hat{a}, \epsilon_3) \subset B(a_0, \epsilon_2)$. Thus $F(\hat{a}, \hat{b}) \leq F(\tilde{a}, \hat{b})$ for all $\tilde{a} \in \bar{B}(\hat{a}, \epsilon_3)$ with $\epsilon_3 < \epsilon_2 \leq \epsilon$.

To show the strictness, we need the following lemma.

Lemma 5. *Let $A \in \mathbb{R}^{n \times n}$ be a strictly positive definite matrix. Let $\sigma_{max}(A)$ and $\sigma_{min}(A)$ denote the largest and smallest singular values of A. Then for a small symmetric perturbation matrix ΔA with $|\sigma_{max}(\Delta A)| < |\sigma_{min}(A)|$, $A + \Delta A$ is also strictly positive definite.*

Proof. Let $x \in \mathbb{R}^n$ be a vector. Then we have

$$x^T(A + \Delta A)x = x^T A x + x^T \Delta A x \geq \sigma_{min}(A)\|x\|^2 - \sigma_{max}(\Delta A)\|x\|^2 > 0. \tag{84}$$

Therefore $A + \Delta A$ is strictly positive definite. □

Note that since F is smooth, the function $[\frac{\partial^2 F}{\partial a_i \partial a_j}](a, b)$ is also smooth. Since $[\frac{\partial^2 F}{\partial a_i \partial a_j}](a, b)$ is smooth at (a_0, b_0) and strictly positive definite, for any $\epsilon_4 > 0$, there exists $\delta_4 > 0$ such that

$$\|[\frac{\partial^2 F}{\partial a_i \partial a_j}](a_0, b_0) - [\frac{\partial^2 F}{\partial a_i \partial a_j}](\alpha, \beta)\| < \epsilon_4, \text{ if } \|(a_0, b_0) - (\alpha, \beta)\| < \delta_4. \tag{85}$$

Therefore, by picking small enough $\epsilon_4 > 0$ such that $\epsilon_4 < \sigma_{min}[\frac{\partial^2 F}{\partial a_i \partial a_j}](a_0, b_0)$, $[\frac{\partial^2 F}{\partial a_i \partial a_j}](\alpha, \beta)$ is strictly positive definite by Lemma 5. By defining $\epsilon_2 = \min\{\epsilon, \epsilon_1/2, \delta_4\}$, instead of $\epsilon_2 = \min\{\epsilon, \epsilon_1/2\}$ we can say that $[\frac{\partial^2 F}{\partial a_i \partial a_j}](\hat{a}, \hat{b})$ is strictly positive definite. Therefore $\hat{a}$ is a strict local minimum of $F(\cdot, \hat{b})$. □

A.2.4 Proof of Theorem 3

Theorem. *Suppose $d_X = d_Y = 1, d_1 \geq 2,\ N \geq 7$, ℓ is L^2 function and the activation function $\sigma(x)$ satisfies Assumption 1. Then there exists a positive measure of $X \in \mathbb{R}^{1 \times N}$ and $Y \in \mathbb{R}^{1 \times N}$, such that the network $1 - d_1 - 1$ has a bad local minimum $\hat{W}^{d_1}$.*

Proof. Consider the small 2-layer network with $d_X = d_1 = d_Y = 1$. By Proposition 3, there exists a strict local minimum $\hat{W}$ with $\mathcal{R}(\hat{W}) > 0$ and strictly positive definite Hessian matrix, for positive measure of (X, Y). Now, we calculate $\mathfrak{B}_{1,1}$ in equation 20 in Thereom 2

$$\mathfrak{B}_{1,1} = \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial f(x_\alpha)} \cdot v_{s,r} \cdot \sigma''(\mathbf{n}(1, r; x_\alpha)) x_{\alpha,1}^2 \tag{86}$$

$$= \sum_i [\Delta Y]_i w_2 \sigma''(x_i) x_i^2 \tag{87}$$

$$= w_2 \sum_i [\Delta Y]_i \sigma''(x_i) x_i^2 \tag{88}$$

$$= w_2 \langle \Delta Y, \sigma''(X) X^2 \rangle > 0. \tag{89}$$

Now, we consider the local minimum embedding on the hidden layer with λ as described in Definition 2. We add a new neuron $\mathbf{n}(1,-1)$ referring $\mathbf{n}(1,1)$ on the hidden layer using embedding function. Denote $W^{(2)}$ be parameters of the larger network at the first step (The larger network has a width of 2 at the first step). By Lemma 13, we have the Hessian matrix H (equation 165) and $\mathfrak{D}_i^{r,s}=0$ by Lemma 14.
Without loss of generality, suppose $w_2>0$. Pick $\lambda\in(0,1)$. Since $\hat{W}$ has the strictly positive definite Hessian, H^{small} is also strictly positive definite, and $\alpha\beta[B_{1,1}]$ is positive. On the last axis $[v_{s,-1}-v_{s,1}]$ of $\mathcal{B}$, the loss of the larger network is constant on the last axis by direct calculation. Therefore, we conclude $W^{(2)}$ is a local minimum of the larger network.
Consider the path λ from $(0,1)$ to $(-\infty,0)\cup(1,\infty)$. Note that the loss is constant along the path. Additionally, $\alpha\beta[B_{1,1}]$ becomes positive to negative along the path, hence we conclude that the point become saddle finally. Because there exists a non-increasing path from $W^{(2)}$, $W^{(2)}$ is not global minimum, *i.e.* the bad local minimum.

We add a new neuron every step. At step t, we have

$$\mathfrak{B}_{1,1}^{(t+1)}=\sum_{\alpha=1}^{N}\frac{\partial\ell_\alpha(x_\alpha,y_\alpha)}{\partial f(x_\alpha)}\cdot(1-\lambda)^{t-1}v_{s,r}\cdot\sigma''(\mathbf{n}(1,r;x_\alpha))x_{\alpha,1}x_{\alpha,1} \tag{90}$$

$$=w_2(1-\lambda)^{t-1}\langle\Delta Y,\sigma''(X)X^2\rangle>0. \tag{91}$$

Therefore by similar argument, we conclude that W^{t+1} is a local minimum.
Finally, we construct sufficiently wide neural network (1-$(t+1)$-1) which has a bad local minimum for each t. □ □

A.2.5 Proof of Lemma 3

Lemma. *Assume $\sigma(x)$ satisfies Assumption 2. Define $B_3(x), B_{3,1}, B_{3,2}$ as*

$$\begin{aligned}B_3(x)=\{&1,\sigma(\sigma(x)),\sigma'(\sigma(x))\{1,\sigma(x),\sigma'(x)\{1,x\},\sigma''(x)\{1,x,x^2\}\},\\&\sigma''(\sigma(x))\{1,\sigma(x),\sigma(x)^2,\sigma'(x)\{1,x\},\sigma''(x)\{1,x,x^2\},\sigma'(x)^2\{1,x,x^2\},\\&\sigma(x)\sigma'(x)\{1,x\},\sigma(x)\sigma''(x)\{1,x,x^2\},\sigma'(x)'\sigma''(x)\{1,x,x^2,x^3\}\},\sigma''(x)^2\{1,x,x^2,x^3,x^4\}\},\end{aligned}$$

$$B_{3,1}(x)=\{\sigma'(\sigma(x))\sigma''(x)x^2\}, \tag{92}$$

$$B_{3,2}(x)=\{\sigma'(\sigma(x))\sigma''(x)\}, \tag{93}$$

$$B_{3,3}(x)=\{\sigma''(\sigma(x))\}, \tag{94}$$

and $\tilde{B_3}(x)=B_3(x)-B_{3,1}(x)-B_{3,2}(x)-B_{3,3}(x)$.
Then, we have

$$span\{B_{3,i}(x)\}\cap span\{\tilde{B}(x)\}=\{0\} \tag{95}$$

$$span\{B_{3,i}(x)\}\cap span\{B_{3,j}(x)\}=\{0\}, \tag{96}$$

for $i,j=1,2,3$ and $i\neq j$.

Proof. First, we need the following lemma.

Lemma 6. *Suppose that $g(x)$ is not zero constant function and $g(x), f_1(x), f_2(x), ..., f_n(x)$ are analytic and a set of analytic functions $\{f_1(x), f_2(x), ..., f_n(x)\}$ is linearly independent. Then $\{g(x)f_1(x), g(x)f_2(x), ..., g(x)f_n(x)\}$ is also linearly independent.*

Proof. Since $g(x)$ is analytic function, the zero set $Z(g(x))$ of $g(x)$ has no limit points. Hence there exists an open interval $I=(a,b)$, such that $g(x)\neq 0$ on I. Suppose $\{g(x)f_1(x), g(x)f_2(x), ..., g(x)f_n(x)\}$ is not linearly independent. Then there exist $\{a_i\}_i=1^n$, such that $\sum_{i=1}^n a_i g(x)f_i(x)=0$ and not all $\{a_i\}_i=1^n$ are zero. Since $g(x)\neq 0$ on I, by multiplying $\frac{1}{g(x)}$, we have $\sum_{i=1}^n a_i f_i(x)=0$ on I. Since $h(x):=\sum_{i=1}^n a_i f_i(x)=0$ is analytic and $h(x)=0$ on I, $h(x)=0$ on $\mathbb{R}$. This is a contradiction. Therefore $\{g(x)f_1(x), g(x)f_2(x), ..., g(x)f_n(x)\}$ is linearly independent. □

Since Claim 1 and 2 hold, we have

$$span\{B\} = span\{1\} \oplus span\{\sigma(\sigma(x))\{...\}\} \oplus span\{\sigma'(\sigma(x))\{...\}\} \oplus span\{\sigma''(\sigma(x))\{...\}\}.$$

Therefore, $span\{\sigma'(\sigma(x))\{...\}\}$ is linearly independent with others.
Since $\sigma(x)$ satisfies Assumption 1, $B_2(x)$ is linearly independent. Since $\sigma'(\sigma(x))$ is analytic, $\{\sigma'(\sigma(x))\{1, \sigma(x), \sigma'(x)\{1, x\}, \sigma''(x)\{1, x, x^2\}\}$ is linearly independent by Lemma 6. Since constant function 1 is linearly independent with all other non-constant functions, $\sigma''(\sigma(x))$ is linearly independent with all other functions in $\{\sigma''(\sigma(x))\{...\}\} - \{\sigma''(\sigma(x))\}$ by Lemma 6. Hence equation 95-equation 96 hold. □

A.2.6 Proof of Lemma 4

Lemma. *Tanh, Sigmoid, SiLU, SoftPlus, GELU, Swish, and Mish functions satisfy Assumption 2.*

Proof. Suppose $\sigma(x) = sigmoid(x)$. By Lemma 12, Claim 1 holds. For Claim 2, suppose

$$\alpha_1 + \alpha_2\sigma(\sigma(x)) + \alpha_3\sigma'(\sigma(x)) + \alpha_4\sigma''(\sigma(x)) = 0, \tag{97}$$

where $\alpha_i \in Q(\sigma(x))$ are not zero at the same time. Since $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, $\sigma'(x)$ and $\sigma''(x)$ are the second and third order polynomials in terms of $\sigma(x)$. Since $\sigma''(\sigma(x)) = P_3(\sigma(x)), \sigma'(\sigma(x)) = P_2(\sigma(x))$, for some third and second order polynomials P_3, P_2, we can say $\sigma(\sigma(x))$ is a solution of a cubic polynomial in $Q(\sigma(x))$. This is a contradiction since $\sigma(\sigma(x))$ is transcendental over $Q(\sigma(x))$. Therefore, Claim 2 holds.
For the case of $\sigma(x) = Tanh$, since $\sigma'(x) = -\sigma(x)^2$, we can apply the similar argument with $sigmoid$.
For the case of $\sigma(x) = SiLU(x)$, we have $\sigma(\sigma(x)) = \log(2+e^x)$. Since $\log(1+e^x)$ and $\log(2+e^x)$ are algebraically independent, Claim 1 and 2 hold.
For the case of $\sigma(x) = GELU(x)$, consider a the following differential field extension

$$Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}}) \subset Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}}, \text{erf}(\frac{x}{2\sqrt{2}}(1+\text{erf}(\frac{x}{\sqrt{2}})))) = Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}}, \sigma(\sigma(x))). \tag{98}$$

Let $p(x) = \text{erf}(\frac{x}{2\sqrt{2}}(1 + \text{erf}(\frac{x}{\sqrt{2}})))$. Then since

$$p'(x) = \frac{2}{\sqrt{\pi}}e^{-\frac{x^2}{8}(1+\text{erf}(\frac{x}{\sqrt{2}}))^2}(\frac{1}{2\sqrt{2}}(1 + \text{erf}(\frac{x}{\sqrt{2}})) + \frac{x}{2\sqrt{2}}\frac{1}{\sqrt{2}\pi}e^{-\frac{x^2}{2}}), \tag{99}$$

$p'(x)$ is transcendental over $Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}})$. Therefore, $p(x)$ is transcendental over $Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}})$. Therefore Claim 1 holds. Since $p(x)$ is not elementary by Liouville's theorem and $e^{-\frac{x^2}{8}(1+\text{erf}(\frac{x}{\sqrt{2}}))^2}$ is transcendental in $Q(\text{erf}(\frac{x}{\sqrt{2}}), e^{-\frac{x^2}{2}})$, Claim 2 holds.
For $\sigma(x) = Swish(x) = xs(\beta x)$, where $\sigma(x) = sigmoid(x)$, because it is similar with the sigmoid, Claim 1 and 2 hold.
For $\sigma(x) = Mish(x)$, Claim 1 holds by Lemma 12. Claim 2 also holds by similar arguments used in the proof of Lemma 1. □

A.2.7 Proof of Proposition 4

Proposition. *Consider a 3-layer network with $N \geq 34$, and $d_X = d_1 = d_2 = d_Y = 1$. Suppose ℓ is L^2 loss function, and $\sigma(x)$ satisfies Assumption 1 and 2. Then there exists a positive measure of $X \in \mathbb{R}^{1\times N}$ and $Y \in \mathbb{R}^{1\times N}$, such that $\mathcal{R}(W)$ has a strict local minimum $\hat{W}$. Moreover, Hessian matrix $H(\hat{W})$ of $\mathcal{R}(W)$ at $\hat{W}$ is strictly positive definite.*

Proof. To begin the proof, we need the following lemmas.

Lemma 7. *Suppose that $\sigma(x)$ is not a constant function and $\sigma(x), f_1(x), f_2(x), ..., f_n(x)$ are analytic, and a set of analytic functions*
$B(x) = \{f_1(x), f_2(x), ..., f_n(x)\}$ is linearly independent,
then $B(\sigma(x)) = \{f_1(\sigma(x)), f_2(\sigma(x)), ..., f_n(\sigma(x))\}$ is also linearly independent.

Proof. Since $\sigma(x)$ is not constant function, there exists $p \in \mathbb{R}$, such that $\sigma'(p) \neq 0$. By Inverse Function Theorem for analytic functions, there exists an open neighborhood $U \in \mathbb{R}$ of p so that $\sigma(x)$ is injective in U and the inverse $\sigma^{-1} : \sigma(U) \to U$ exists and is analytic.
Now, suppose $B(\sigma(x))$ is not linearly independent. Then there exist $\{a_i\}_{i=1}^n$ such that $\sum_{i=1}^n a_i f_i(\sigma(x)) = 0$ and not all $\{a_i\}_{i=1}^n$ are zero. For any $y \in \sigma(U)$, there exists the unique $x \in U$, such that $\sigma(x) = y$. Since $f_i(y) = f_i(\sigma(x))$, we have $\sum_{i=1}^n a_i f_i(y) = \sum_{i=1}^n a_i f_i(\sigma(x)) = 0$ for all $y \in \sigma^{-1}(U)$. Define $h(y)$ as $h(y) := \sum_{i=1}^n a_i f_i(y)$. The fact $h(y) = 0$ on $\sigma^{-1}(U)$ implies that the zero set of $h(y)$ has a limit point. Because $h(y)$ is analytic, $h(y) = 0$ in $\mathbb{R}$. This contradicts that $B(\sigma(x))$ is not linearly independent. Therefore $B(\sigma(x))$ is linearly independent. □

Lemma 8. *If $\sigma(x)$ satisfies Assumption 1, then*
$B_2(\sigma(x)) = \{1, \sigma(\sigma(x)), \sigma'(\sigma(x)), \sigma(x)\sigma'(\sigma(x)), \sigma''(\sigma(x)), \sigma(x)\sigma''(\sigma(x)), \sigma(x)^2\sigma''(\sigma(x))\}$ *is linearly independent.*

Proof. Since $\sigma(x)$ satisfies Assumption 1,
$B_2(x) = \{1, \sigma(x), \sigma'(x), x\sigma'(x), \sigma''(x), x\sigma''(x), x^2\sigma''(x)\}$ is linearly independent.
By Lemma 7, $B_2(\sigma(x))$ is linearly independent. □

Consider 3-layer network.

$$F_3(W) = w_3\sigma(w_2\sigma(w_1x + b_1) + b_2) + b_3. \tag{100}$$

First, recall that

$$\begin{aligned}B_3(x) = \{&1, \sigma(\sigma(x)), \sigma'(\sigma(x))\{1, \sigma(x), \sigma'(x)\{1, x\}, \sigma''(x)\{1, x, x^2\}\},\\ &\sigma''(\sigma(x))\{1, \sigma(x), \sigma(x)^2, \sigma'(x)\{1, x\}, \sigma''(x)\{1, x, x^2\}, \sigma'(x)^2\{1, x, x^2\},\\ &\sigma(x)\sigma'(x)\{1, x\}, \sigma(x)\sigma''(x)\{1, x, x^2\}, \sigma'(x)'\sigma''(x)\{1, x, x^2, x^3\}\}, \sigma''(x)^2\{1, x, x^2, x^3, x^4\}\},\end{aligned}$$

$$B_{3,1}(x) = \{\sigma'(\sigma(x))\sigma''(x)x^2\}, \tag{101}$$
$$B_{3,2}(x) = \{\sigma'(\sigma(x))\sigma''(x)\}, \tag{102}$$
$$B_{3,3}(x) = \{\sigma''(\sigma(x))\}, \tag{103}$$

and $\tilde{B}_3(x) = B_3(x) - B_{3,1}(x) - B_{3,2}(x) - B_{3,3}(x)$.
By Lemma 3, since $\sigma(x)$ satisfies Assumption 2, we have

$$span\{B_{3,i}(x)\} \cap span\{\tilde{B}(x)\} = \{0\} \tag{104}$$
$$span\{B_{3,i}(x)\} \cap span\{B_{3,j}(x)\} = \{0\}, \tag{105}$$

for $i, j = 1, 2, 3$ and $i \neq j$. Then we can decompose $\mathcal{B}(x) = span\{B_3(x)\}$ as

$$\mathcal{B}_3(x) = \mathcal{B}_{3,1}(x) \oplus \mathcal{B}_{3,2}(x) \oplus \mathcal{B}_{3,3}(x) \oplus \mathcal{B}_3^{\perp}(x). \tag{106}$$

where $\mathcal{B}_{3,1}(x) = span\{B_{3,1}(x)\}$, $\mathcal{B}_{3,2}(x) = span\{B_{3,2}(x)\}$, $\mathcal{B}_{3,3}(x) = span\{B_{3,3}(x)\}$, and $\mathcal{B}_3^{\perp}(x)$ is the space in $\mathcal{B}(x)$ which is orthogonal to $\mathcal{B}_{3,1}(x)$, $\mathcal{B}_{3,2}(x)$, and $\mathcal{B}_{3,3}(x)$. In addition, we can say

$$\tilde{\mathcal{B}}_3(x) \subseteq \mathcal{B}_3^{\perp}(x), \tag{107}$$

where $\tilde{\mathcal{B}}_3(x) = span\{\tilde{B}_3(x)\}$.
Consider the mapping

$$(x_1, x_2, ..., x_N) \mapsto det([B_2(\sigma(x_i))]_{i=1}^7), \tag{108}$$

where $[B_2(\sigma(x_i))]_{i=1}^7$ is

$$\begin{aligned}&[B_2(\sigma(x_i))]_{i=1}^7\\ &= \{\mathbf{1}_N, [\sigma(\sigma(x_i))], [\sigma'(\sigma(x_i))], [\sigma(x_i)\sigma'(\sigma(x_i))], [\sigma''(\sigma(x_i))], [\sigma(x_i)\sigma''(\sigma(x_i))], [\sigma(x_i)^2\sigma''(\sigma(x_i))]\}_{i=1}^7.\end{aligned}$$

Note that only first seven x_is are used in the mapping. Since $B(\sigma(x))$ is linearly independent by Lemma 8, the mapping is not a zero map. Because the mapping is not a zero map and is analytic, all seven vectors in $B_2([\sigma(x_i)]_{i=1}^7)$ are linearly independent for generic $X = (x_1, x_2, ..., x_N)$. Therefore $B_2([\sigma(x_i)]_{i=1}^N)$ is also linearly independent for generic $X = (x_1, x_2, ..., x_N)$.

We define a reduction set $red(A)$ of a set $A = \{a_1, a_2, ..., a_n\}$ as

$$red(A) = \{a_i | a_i \notin span\{a_1, a_2, ..., a_{i-1}\}\}. \tag{109}$$

Note that $dim(span\{A\}) = dim(span\{red(A)\}) = |red(A)|$.
Consider the mapping

$$(x_1, x_2, ..., x_{N_0}) \mapsto det([red(B_3)(x_i)]_{i=1}^{N_0}), \tag{110}$$

where $red(B_3(x))$ is a reduction set of $B_3(x)$ and $N_0 = dim(\mathcal{B}_3(x))$. Note that $N_0 = dim(\mathcal{B}_3(x))$ $= dim(span\{B_3(x)\}) \leq 34$. Since the mapping is not a zero map and is analytic,

$$[\mathcal{B}_3(x_i)]_{i=1}^{N_0} = [\mathcal{B}_{3,1}(x_i)]_{i=1}^{N_0} \oplus [\mathcal{B}_{3,2}(x_i)]_{i=1}^{N_0} \oplus [\mathcal{B}_{3,3}(x_i)]_{i=1}^{N_0} \oplus [\mathcal{B}_3^{\perp}(x_i)]_{i=1}^{N_0}, \tag{111}$$

$$[\tilde{\mathcal{B}}_3(x)]_{i=1}^{N_0} \subseteq [\mathcal{B}_3^{\perp}(x)]_{i=1}^{N_0} \tag{112}$$

hold for generic $X = (x_1, x_2, ..., x_N)$.
Then we can find $(N - N_0))$ independent N-dimensional vectors $\{v_k\}$ such that

$$\langle v_k, [b_{j,\mathcal{B}_3^{\perp}}(x_i)]\rangle = 0 \tag{113}$$

$$\langle v_k, [b_{\mathcal{B}_{3,1}}(x_i)]\rangle > 0 \tag{114}$$

$$\langle v_k, [b_{\mathcal{B}_{3,2}}(x_i)]\rangle > 0 \tag{115}$$

$$\langle v_k, [b_{\mathcal{B}_{3,3}}(x_i)]\rangle > 0, \tag{116}$$

where $\{b_{j,\mathcal{B}^{\perp}(x)}\}$, $\{b_{\mathcal{B}_{3,1}}(x)\}$, $\{b_{\mathcal{B}_{3,2}}(x)\}$, and $\{b_{\mathcal{B}_{3,3}}(x)\}$ are basis of $\mathcal{B}^{\perp}(x)$, $\mathcal{B}_{3,1}(x)$, $\mathcal{B}_{3,2}(x)$, and $\mathcal{B}_{3,3}(x)$, respectively. Without loss of generality, we set

$$b_{\mathcal{B}_{3,1}}(x) = \sigma'(\sigma(x))\sigma''(x)^2 x^2 \tag{117}$$

$$b_{\mathcal{B}_{3,2}}(x) = \sigma'(\sigma(x))\sigma''(x)^2 \tag{118}$$

$$b_{\mathcal{B}_{3,3}}(x) = \sigma''(\sigma(x)). \tag{119}$$

Therefore we conclude

$$\langle v_k, [b_{j,\tilde{\mathcal{B}}_3}(x_i)]\rangle = 0, \tag{120}$$

where $b_{j,\tilde{\mathcal{B}}_3}$ are basis of $\tilde{\mathcal{B}}_3 = span(\tilde{B}_3)$.

Then select data point $Y = [y_i]_{i=1}^N$ as

$$y_i = F_3(x_i) - w_3 \sum_{k=1}^{N-N_0} c_k [v_k]_i. \tag{121}$$

for some $c_k \in \mathbb{R}$.
Define $\Delta Y \in \mathbb{R}^N$ such that

$$[\Delta Y]_i = F_3(x_i) - y_i = w_3 \sum_{k=1}^{N-N_0} c_k [v_k]_i.$$

Then we have:

$$\langle \Delta Y, [b_{j,\mathcal{B}_3^{\perp}}(x_i)]\rangle = 0 \tag{122}$$

$$\langle \Delta Y, [b_{j,\mathcal{B}_{3,1}}(x_i)]\rangle > 0 \tag{123}$$

$$\langle \Delta Y, [b_{j,\mathcal{B}_{3,2}}(x_i)]\rangle > 0. \tag{124}$$

$$\langle \Delta Y, [b_{j,\mathcal{B}_{3,3}}(x_i)]\rangle > 0. \tag{125}$$

Now pick $\hat{W} = (w_1, b_1, w_2, b_2, w_3, b_3) = (1, 0, 1, 0, w_3, 0)$, where $w_3 > 0$.
First, we claim that $\hat{W}$ is a stationary point. To show this, we compute the gradients of $\mathcal{R}(W) =$

$\|F_3(W)-Y\|^2 = \|w_3\sigma(w_2\sigma(w_1x+b_1)+b_2)+b_3-Y\|^2$ at $W=\hat{W}$. By equation 113-equation 116 the gradients are computed as

$$\frac{\partial \mathcal{R}}{\partial w_3} = 2\langle \Delta Y, \sigma(\sigma(X))\rangle = 0 \tag{126}$$

$$\frac{\partial \mathcal{R}}{\partial b_3} = 2\langle \Delta Y, \mathbf{1}_N\rangle = 0 \tag{127}$$

$$\frac{\partial \mathcal{R}}{\partial w_2} = 2\langle \Delta Y, \sigma'(\sigma(X))\sigma(X)\rangle = 0 \tag{128}$$

$$\frac{\partial \mathcal{R}}{\partial b_2} = 2\langle \Delta Y, \sigma'(X)\rangle = 0 \tag{129}$$

$$\frac{\partial \mathcal{R}}{\partial w_1} = 2\langle \Delta Y, \sigma'(\sigma(X))\sigma'(X)X\rangle = 0 \tag{130}$$

$$\frac{\partial \mathcal{R}}{\partial b_1} = 2\langle \Delta Y, \sigma'(\sigma(X))\sigma'(X)\rangle = 0. \tag{131}$$

Therefore $\hat{W}$ is a stationary point of $\mathcal{R}(W)$.
To show $\hat{W}$ is a strict local minimum, we need to show that Hessian $H(\hat{W})$ of $\mathcal{R}(W)$ is strictly positive definite at $W=\hat{W}$. Therefore we need to show that

$$\mathbf{u}^T H(\hat{W})\mathbf{u} = \lim_{t\to 0} \frac{\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W})}{t^2} > 0, \tag{132}$$

for all $\mathbf{u} = (u_{w_1}, u_{b_1}, u_{w_2}, u_{b_2}, u_{w_3}, u_{b_3})$ with $\|\mathbf{u}\|_2 = 1$. Similar to equation 70, we compute $\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W})$ as

$$\begin{aligned}
&\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W}) \\
&= \|F_3(\hat{W}+t\mathbf{u}) - F_3(\hat{W})\|_2^2 + \|F_3(\hat{W}-t\mathbf{u}) - F_3(\hat{W})\|_2^2 + \\
&2\langle \Delta Y, F_3(\hat{W}+t\mathbf{u}) - F_3(\hat{W})\rangle + 2\langle \Delta Y, F_3(\hat{W}-t\mathbf{u}) - F_3(\hat{W})\rangle.
\end{aligned} \tag{133}$$

We consider the following two cases.
Case 1: $(u_{w_1}, u_{b_1}) = (0,0)$
In this case, we have a similar to the case of $L=2$ network where x is replaced by $\sigma(x)$.

$$\begin{aligned}
&F_3(\hat{W}+t\mathbf{u})(x) - F_3(\hat{W})(x) \\
&= ((w_3 + tu_{w_3})\sigma((1+tu_{w_2})\sigma(x) + tu_{b_2}) + tu_{b_3}) - (w_3\sigma(\sigma(x))).
\end{aligned} \tag{134}$$

Since $B_2([\sigma(x_i)]_{i=1}^N)$ is linearly independent and $\langle \Delta Y, \sigma''(\sigma(X))\rangle > 0$, we can apply the similar argument used in the proof of Proposition 2 (the case of $L=2$). Therefore, we conclude

$$\mathbf{u}^T H(\hat{W})\mathbf{u} = \lim_{t\to 0} \frac{\mathcal{R}(\hat{W}+t\mathbf{u}) + \mathcal{R}(\hat{W}-t\mathbf{u}) - 2\mathcal{R}(\hat{W})}{t^2} > 0. \tag{135}$$

Case 2: $(u_{w_1}, u_{b_1}) \neq (0,0)$
We calculate $F_3(\hat{W}+t\mathbf{u})(x)$. By Taylor theorem, we have

$$\sigma((1+tu_{w_1})x + tu_{b_1}) = \sigma(x) + \sigma'(x)(tu_{w_1}x + tu_{b_1}) + \frac{1}{2}\sigma''(x)(tu_{w_1}x + tu_{b_1})^2 + o(t^2), \tag{136}$$

Then, by Taylor theorem again, we have

$$\begin{aligned}
&\sigma((1+tu_{w_2})\sigma((1+tu_{w_1})x + tu_{b_1}) + tu_{b_2}) \\
=&\sigma((1+tu_{w_2})(\sigma(x) + \sigma'(x)(tu_{w_1}x + tu_{b_1}) + \frac{1}{2}\sigma''(x)(tu_{w_1}x + tu_{b_1})^2 + o(t^2)) + tu_{b_2}) \\
=&\sigma(\sigma(x) + tu_{w_2}\sigma(x) + (1+tu_{w_2})(\sigma'(x)(tu_{w_1}x + tu_{b_1}) + \frac{1}{2}\sigma''(x)(tu_{w_1}x + tu_{b_1})^2 + o(t^2)) + tu_{b_2}) \\
=&\sigma(\sigma(x)) + \sigma'(\sigma(x))u_2 + \frac{1}{2}\sigma''(\sigma(x))u_2^2 + o(t^2),
\end{aligned} \tag{137}$$

where $u_2 = tu_{w_2}\sigma(x)+(1+tu_{w_2})(\sigma'(x)(tu_{w_1}x+tu_{b_1})+\frac{1}{2}\sigma''(x)(tu_{w_1}x+tu_{b_1})^2)+tu_{b_2}+o(t^2)$. Therefore $F_3(\hat{W}+t\mathbf{u})(x)$ is

$$F_3(\hat{W}+t\mathbf{u})(x) = (w_3+tu_{w_3})\sigma((1+tu_{w_2})\sigma((1+tu_{w_1})x+tu_{b_1})+tu_{b_2})+tu_{b_3} \tag{138}$$

$$=(w_3+tu_{w_3})(\sigma(\sigma(x))+\sigma'(\sigma(x))u_2+\frac{1}{2}\sigma''(\sigma(x))u_2^2+o(t^2))+tu_{b_3}. \tag{139}$$

so we have

$$\begin{aligned}&F_3(\hat{W}+t\mathbf{u})(x)-F_3(\hat{W})(x)\\&=tu_{w_3}\sigma(\sigma(x))+(w_3+tu_{w_3})(\sigma'(\sigma(x))u_2+\frac{1}{2}\sigma''(\sigma(x))u_2^2+o(t^2))+tu_{b_3}.\end{aligned} \tag{140}$$

Note that

$$\begin{aligned}u_2^2 &= t^2u_{w_2}^2\sigma(x)^2\\&+2tu_{w_2}(1+tu_{w_2})(\sigma(x)\sigma'(x)(tu_{w_1}x+tu_{b_1})+\frac{1}{2}\sigma(x)\sigma''(x)(tu_{w_1}x+tu_{b_1})^2+2t^2u_{w_2}u_{b_2}\sigma(x))\\&+(1+tu_{w_2})^2(\sigma'(x)^2(tu_{w_1}x+tu_{b_1})^2+\sigma(x)\sigma'(x)(tu_{w_1}x+tu_{b_1})(tu_{w_1}x+tu_{b_1})^2\\&+\sigma'(x)(tu_{w_1}x+tu_{b_1})tu_{b_2}+\frac{1}{4}\sigma''(x)^2(tu_{w_1}x+tu_{b_1})^4)\\&+2(1+tu_{w_2})tu_{b_2}(\sigma'(x)(tu_{w_1}x+tu_{b_1})+\frac{1}{2}\sigma''(x)(tu_{w_1}x+tu_{b_1})^2)+t^2u_{b_2}^2+o(t^2).\end{aligned}$$

By expanding equation 140 and utilizing equation 113-equation 116, we can calculate

$$2\langle\Delta Y, F_3(\hat{W}+t\mathbf{u})-F_3(\hat{W})\rangle+2\langle\Delta Y, F_3(\hat{W}-t\mathbf{u})-F_3(\hat{W})\rangle \tag{141}$$

$$\begin{aligned}&=(w_3+tu_{w_3})t^2(1+tu_{w_2})u_{w_1}^2\langle\Delta Y,\sigma'(\sigma(X))\sigma''(X)X^2\rangle\\&+(w_3+tu_{w_3})t^2(1+tu_{w_2})u_{b_1}^2\langle\Delta Y,\sigma'(\sigma(X))\sigma''(X)\rangle\\&+(w_3+tu_{w_3})t^2u_{b_2}^2\langle\Delta Y,\sigma''(\sigma(X))\rangle+o(t^2).\end{aligned} \tag{142}$$

Therefore $\mathbf{u}^TH(\hat{W})\mathbf{u}$ is

$$\begin{aligned}\mathbf{u}^TH(\hat{W})\mathbf{u} &= \lim_{t\to0}\frac{\mathcal{R}(\hat{W}+t\mathbf{u})+\mathcal{R}(\hat{W}-t\mathbf{u})-2\mathcal{R}(\hat{W})}{t^2}\\&\geq\lim_{t\to0}\frac{(w_3+tu_{w_3})t^2(1+tu_{w_2})u_{w_1}^2\langle\Delta Y,\sigma'(\sigma(X))\sigma''(X)X^2\rangle}{t^2}\\&+\lim_{t\to0}\frac{(w_3+tu_{w_3})t^2(1+tu_{w_2})u_{b_1}^2\langle\Delta Y,\sigma'(\sigma(X))\sigma''(X)\rangle}{t^2}\\&+\lim_{t\to0}\frac{(w_3+tu_{w_3})t^2u_{b_2}^2\langle\Delta Y,\sigma''(\sigma(X))\rangle}{t^2}>0.\end{aligned} \tag{143}$$

In both cases, we have $\mathbf{u}^TH(\hat{W})\mathbf{u}>0$ for all $\mathbf{u}$. Therefore, Hessian $H(\hat{W})$ is strictly positive definite at $\hat{W}$ and $\hat{W}$ is a strict local minimum of $\mathcal{R}(W)$.

Similar to the proof of Proposition 3, using Lemma 2, we show that a strict local minimum exists for data of positive measure. □

A.2.8 PROOF OF THEOREM 4

Theorem. *Suppose $d_X=d_Y=1, d_1\geq2, d_2\geq2,\ N\geq34$, ℓ is L^2 function and the activation function $\sigma(x)$ satisfies Assumption 1 and 2. Then there exists a positive measure of $X\in\mathbb{R}^{1\times N}$ and $Y\in\mathbb{R}^{1\times N}$, such that the network $1-d_1-d_2-1$ has a bad local minimum.*

Proof. Consider the small 3-layer network with $d_X=d_1=d_2=d_Y=1$. Then by Proposition 4, there exists a strict local minimum $\hat{W}$ with $\mathcal{R}(\hat{W})>0$. Let $\mathfrak{B}_{1,1}(1),\mathfrak{D}_{1,1}(1)$ and $\mathfrak{B}_{1,1}(2),\mathfrak{D}_{1,1}(2)$

be the $\mathfrak{B}$-matrix and $\mathfrak{D}$-matrix in Theorem 2 at layer 1 and 2. Then we have $\mathfrak{D}_{1,1}(1) = 0$, $\mathfrak{D}_{1,1}(2) = 0$, and

$$\mathfrak{B}_{1,1}(1) = \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(2,1;x_\alpha)} \cdot v_{1,r} \cdot \sigma''(\mathbf{n}(1,r;x_\alpha)) x_{\alpha,1}^2 \tag{144}$$

$$= \sum_{i=1}^{N} [\Delta Y]_i w_3 \sigma'(\sigma(x_i)) \sigma''(x_i) x_i^2 \tag{145}$$

$$= w_3 \langle \Delta Y, \sigma'(\sigma(X)) \sigma''(X) X^2 \rangle > 0. \tag{146}$$

$$\mathfrak{B}_{1,1}(2) = \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(3,1;x_\alpha)} \cdot v_{1,r} \cdot \sigma''(\mathbf{n}(2,r;x_\alpha)) \mathbf{act}(1,1;x_\alpha)^2 \tag{147}$$

$$= \sum_{i=1}^{N} [\Delta Y]_i w_3 \sigma''(\sigma(x_i)) \sigma(x_i)^2 \tag{148}$$

$$= w_3 \langle \Delta Y, \sigma''(\sigma(X)) \sigma(X)^2 \rangle = 0. \tag{149}$$

Given $d_1 > 1$ and $d_2 > 1$, our strategy is as follows. We refer the step of constructing the network $1 - a - b - 1$ as step (a, b).

In the first step $(2, 1)$, we consider the local minimum embedding on the layer 1 with $\lambda_1 \in (0, 1)$. Similar to Theorem 2, because $\mathfrak{B}_{1,1}(1)$ is positive, the weights $W^{(2,1)}$ is the local minimum. By adding a new neuron every step, we construct the network $1 - d_1 - 1 - 1$ with the local minimum $W^{(d_1,1)}$. Along the path of changing λ_1 from $(0, 1)$ to $(\infty, 0) \cup (1, \infty)$, because there exists a non-increasing path to the global minimum, $W^{(d_1,1)}$ is the bad local minimum.

Next, we add a new neuron in layer 2. In the step $(d_1, 2)$, we consider the local minimum embedding on the layer 2 with $\lambda_2 \in (0, 1)$. Then $\mathfrak{D}(2)^{(d_1,2)} = 0$ by Lemma 14. Since $\mathfrak{B}(2)^{(1,1)} = 0$, to show the local minimality, we see the Hessian matrix (equation 165).

Along the path $[\alpha u_{-1,i} - \beta u_{r,i}, v_{s,-1} - v_{s,r}]$, $(\lambda_2 = \frac{\beta_2}{\alpha_2 + \beta_2})$, the loss is constant. Therefore $W^{(d_1,2)}$ is the bad local minimum of the network $1 - d_1 - 2 - 1$. By adding a new neuron every step, we finally construct the network $1 - d_1 - d_2 - 1$ with the local minimum $W^{(d_1,d_2)}$. As shown above, $W^{(d_1,d_2)}$ is the bad local minimum along the path of changing λ_1 from $(0, 1)$ to $(\infty, 0) \cup (1, \infty)$. $\square$

B DIFFERENTIAL GALOIS THEORY

In this section, we give a brief introduction to the differential Galois theory. For more information, please see Hubbard & Lundell (2011); Churchill (2006); Van der Put & Singer (2012).

Definition 3. *Let K be a field. An additive group homomorphism $(') : K \to K$ is a derivation, if the Leibniz rule*

$$(k_1 k_2)' = k_1' k_2 + k_1 k_2', \tag{150}$$

holds for all $k_1, k_2 \in K$.
K is called differential field if K is equipped with the derivation.
The subfield $Con(K)$ is called the constants of K if

$$Con(K) = \{k \in K : k' = 0\}. \tag{151}$$

Definition 4. *Let L be a field and K be a subfield of K. $K \subset L$ is called a field extension. The larger field L is a K-vector space. The degree of a field extension $K \subset L$ is the dimension of the vector space, i.e.,*

$$[L : K] = \dim_K L. \tag{152}$$

α is algebraic if α is a root of a non-zero polynomial with coefficients in K. If every element of L is algebraic over K, then an extension $K \subset L$ is called an algebraic extension. If α is not a root of any polynomial with coefficients in K, α is transcendental. An extension $K \subset L$ is a transcendental extension if L has a transcendental element over K.

Proposition 5. *For an algebraic extension $K \subset K(\alpha)$, the extension degree $[K(\alpha) : K]$ equals the degree of the minimal polynomial $p(x)$, such that $p(\alpha) = 0$. If the extension $K \subset K(\alpha)$ is transcendental, the the extension degree is infinite.*

Lemma 9. *The functions e^x and $\log(x)$ are transcendental.*

Definition 5. *Let $K \subset L$ be a algebraic field extension. The extension $K \subset L$ is called normal extension if every irreducible polynomial over K which has a root in L, splits into linear factors in L. The extension $K \subset L$ is called separable extension if for every $\alpha \in L$, the minimal polynomial of α has no repeated roots. The extension $K \subset L$ is called Galois extension if it is normal and separable.*
If the extension $K \subset L$ is Galois, then its corresponding Galois group $Gal(L/K)$ is defined as the group of field automorphisms of L which fixes K.

Remark 4. *If field K is finite field or a field of characteristic zero, then every algebraic extension of K is separable.*

Proposition 6. *Let $K \subset L = K(\alpha_1, \alpha_2, ..., \alpha_n)$ be a Galois extension, where $\alpha_1, \alpha_2, ..., \alpha_n$ are roots of a irreducible polynomial $p(x)$ of degree n. Then its corresponding Galois group $Gal(L/K)$ is a subgroup of symmetric group S_n.*

Definition 6. *Let K be a differential field. $L = K(l)$ is called a logarithmic extension of K if l is transcendental over K and*

$$l' = \frac{k'}{k}, \tag{153}$$

for some $k \in K$.
Similarly, $L = K(l)$ is called a exponential extension of K if l is transcendental over K and

$$\frac{l'}{l} = k', \tag{154}$$

for some $k \in K$.
This is analogues of the logarithm and exponential where $l = \log(k)$ and $l = e^k$, respectively.
Then a differential field extension $K \subset L$ is elementary if there exists a finite sequence of intermediate differential field extensions

$$K = K_0 \subset K_1 \subset ... \subset K_n = L, \tag{155}$$

such that each $K_i \subset K_{i+1}$ is algebraic, logarithmic, or exponential extension. l is called elementary if $K \subset K(l)$ is an elementary extension.

Proposition 7. *If $\frac{l'}{l} = k'$ for some nonzero $k \in K$, then $l \notin K$, and moreover l cannot be algebraic over K. Similarly if $l' = \frac{k'}{k}$ for some nonzero $k \in K$, then $l \notin K$, and moreover l cannot be algebraic over K.*

Theorem 5 (Liouville). *Let K and L be differential fields of characteristic 0 with $Con(K) = Con(L)$, and $K \subset L$ be an elementary extension. Suppose $k \in K$ has no anti-derivative in K, and there exists $l \in L$ such that $l' = k$ (i.e. l has an anti-derivative in L). Then there exist $c_1, ..., c_n \in Con(K)$ and $k_1, ..., k_n, \gamma \in K$ such that*

$$k = c_1 \frac{k_1'}{k_1} + ... + c_n \frac{k_n'}{k_n} + \gamma'. \tag{156}$$

In other words, if k has an elementary anti-derivative, then k must have this form.

Corollary 1. *For $K = \mathbb{R}(x)$ or $\mathbb{C}(x)$, the functions e^{-x^2} and $\frac{\sin(x)}{x}$ have no elementary anti-derivatives.*

Therefore the error function $\text{erf}(x) = \int_0^x e^{-t^2} dt$ has no elementary anti-derivative.

Definition 7. *Let $K \subset L$ be a differential field extension. Its corresponding differential Galois group $G := DGal(L/K)$ is defined as the group of differential field automorphisms of L which fixes K, and such that*

$$g(l') = g(l)', \tag{157}$$

for all $g \in DGal(L/K)$ and $l \in L$.

Lemma 10, 11, and 12 describe some lemmas for exponential and logarithmic extensions.

Lemma 10. *Let $K = \mathbb{C}(x)$. Consider differential field extensions*

$$K \subset K(\alpha) \subset K(\alpha, \beta), \tag{158}$$

where $K \subset K(\alpha)$ is an exponential extension, and $\beta' = \frac{t'}{t}$ for some $t \in K(\alpha) \setminus K$. Then the extension $K(\alpha) \subset K(\alpha, \beta)$ is transcendental (therefore it is a logarithmic extension), if and only if t is not a monomial in terms of α in $K(\alpha)$.

Lemma 11. *Let $K = \mathbb{C}(x)$. Consider differential field extensions*

$$K \subset K(\alpha) \subset K(\alpha, \beta), \tag{159}$$

where $K \subset K(\alpha)$ is an logarithmic extension, and $\frac{\beta'}{\beta} = t'$ for some $t \in K(\alpha) \setminus K$. Then the extension $K(\alpha) \subset K(\alpha, \beta)$ is transcendental (therefore it is an exponential extension), if and only if β is not a form of $a\alpha + b$ for some $a, b \in \mathbb{C}$ in $K(\alpha)$.

Lemma 12. *Let $K = \mathbb{C}(x)$. Consider differential field extensions*

$$K \subset K(\alpha) \subset K(\alpha, \beta), \tag{160}$$

where $K \subset K(\alpha)$ is an exponential extension, and $\frac{\beta'}{\beta} = t'$ for some $t \in K(\alpha) \setminus K$. Then the extension $K(\alpha) \subset K(\alpha, \beta)$ is transcendental (therefore it is an exponential extension).

C LOCAL MINIMUM EMBEDDING

In this section, we give a brief introduction to the technique called local minimum embedding. For more information, please see Fukumizu & Amari (2000); Nitta (2016); Petzka & Sminchisescu (2021).

Definition (Local minimum embedding)**.** *Consider a L-layer neural network $F^{small}(X)$ with the width $d_1, d_2, ...d_L$ and the weights W^{small}. Consider a neuron $\mathbf{n}(l,r)$ with index r in layer l. Let $[u_{r,i}^{small}]_{i=1}^{d_{l-1}}$ be the incoming weights into $\mathbf{n}(l,r)$ and $[v_{s,r}^{small}]_{s=1}^{d_{l+1}}$ be the outgoing weights of $\mathbf{n}(l,r)$. The weights of the smaller network W^{large} can be represented as*

$$W^{small} = ([u_{r,i}^{small}]_{i=1}^{d_{l-1}}, [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \bar{W}^{small}), \tag{161}$$

where $\bar{W}^{small}$ denote the collection of all remaining weights of the smaller network.
Consider the larger network $F_{large}(X)$ by adding a new neuron $\mathbf{n}(l,-1)$ referring (l,r) with new weights $[u_{-1,i}^{large}]_{i=1}^{d_{l-1}}$ and $[v_{s,-1}^{large}]_{s=1}^{d_{l+1}}$. The weights of the larger network W^{large} can be represented as

$$W^{large} = ([u_{-1,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,-1}^{large}]_{s=1}^{d_{l+1}}, [u_{r,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,r}^{large}]_{s=1}^{d_{l+1}}, \bar{W}^{large}). \tag{162}$$

where $\bar{W}^{small}$ denote the collection of all remaining weights of the smaller network. Note that $[u_{-1,i}^{large}]_{i=1}^{d_{l-1}}$ and $[v_{s,-1}^{large}]_{s=1}^{d_{l+1}}$ are the incoming and outgoing weights of the new neuron $\mathbf{n}(l,-1)$ in the larger network.
Then define the local minimum embedding function γ_λ^r mapping W^{small} to W^{large} as

$$\gamma_\lambda^r([u_{r,i}^{small}]_{i=1}^{d_{l-1}}, [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \bar{W}^{small}) = ([u_{-1,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,-1}^{large}]_{s=1}^{d_{l+1}}, [u_{r,i}^{large}]_{i=1}^{d_{l-1}}, [v_{s,r}^{large}]_{s=1}^{d_{l+1}}, \bar{W}^{large})$$

with

$$[u_{-1,i}^{large}]_{i=1}^{d_{l-1}} = [u_{r,i}^{small}]_{i=1}^{d_{l-1}},\ [v_{s,-1}^{large}]_{s=1}^{d_{l+1}} = \lambda [v_{s,r}^{small}]_{s=1}^{d_{l+1}}, \tag{163}$$

$$[u_{r,i}^{large}]_{i=1}^{d_{l-1}} = [u_{r,i}^{small}]_{i=1}^{d_{l-1}},\ [v_{s,r}^{large}]_{s=1}^{d_{l+1}} = (1-\lambda)[v_{s,r}^{small}]_{s=1}^{d_{l+1}},\ \bar{W}^{large} = \bar{W}^{small}. \tag{164}$$

From here, we omit the superscript $large$ for the weights of the larger network. In fact, we can explicitly represent the Hessian of the larger network.

Lemma 13 (Petzka & Sminchisescu (2021))**.** *Let $\mathcal{L}$ denote the loss function of the larger network and ℓ be the loss function of smaller network. Let $\lambda = \frac{\beta}{\alpha+\beta}$. Then the Hessian H of the loss $\mathcal{L}$ with respect to the basis $\mathcal{B} = [u_{-1,r} + u_{r,i}, v_{s,-1} + v_{s,r}, \bar{w}, \alpha u_{-1,i} - \beta u_{r,i}, v_{s,-1} - v_{s,r}]$ is given by:*

$$H = \begin{bmatrix} \frac{\partial^2 \ell}{\partial u_{r,i} \partial u_{r,j}} & 2\frac{\partial^2 \ell}{\partial u_{r,i} \partial v_{s,r}} & \frac{\partial^2 \ell}{\partial \bar{w} \partial u_{r,i}} & 0 & 0 \\ 2\frac{\partial^2 \ell}{\partial u_{r,i} \partial v_{s,r}} & 4\frac{\partial^2 \ell}{\partial v_{s,r} \partial v_{t,v}} & 2\frac{\partial^2 \ell}{\partial \bar{w} \partial v_{s,r}} & (\alpha-\beta)[\mathfrak{D}_i^{r,s}] & 0 \\ \frac{\partial^2 \ell}{\partial \bar{w} \partial u_{r,i}} & 2\frac{\partial^2 \ell}{\partial \bar{w} \partial v_{s,r}} & \frac{\partial^2 \ell}{\partial \bar{w} \partial \bar{w}'} & 0 & 0 \\ 0 & (\alpha-\beta)[\mathfrak{D}_i^{r,s}] & 0 & \alpha\beta[\mathfrak{B}_{i,j}^r] & (\alpha+\beta)[\mathfrak{D}_i^{r,s}] \\ 0 & 0 & 0 & (\alpha+\beta)[\mathfrak{D}_i^{r,s}] & 0 \end{bmatrix} \tag{165}$$

where

$$\mathfrak{B}_{i,j} = \sum_{\alpha=1}^{N} \sum_{k=1}^{d_{l+1}} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,k;x_\alpha)} \cdot v_{k,r} \cdot \sigma''(\mathbf{n}(l,r;x_\alpha)) \mathbf{act}(l-1,i;x_\alpha) \mathbf{act}(l-1,j;x_\alpha) \tag{166}$$

and

$$\mathfrak{D}_i^{r,s} := \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,s;x_\alpha)} \sigma'(\mathbf{n}(l,r;x_\alpha)) \mathbf{act}(l-1,i;x_\alpha). \tag{167}$$

Theorem ((Petzka & Sminchisescu, 2021))**.** *Define the matrices $\mathfrak{B}_{i,j}$ and $\mathfrak{D}_i^{r,s}$ as*

$$\mathfrak{B}_{i,j} = \sum_{\alpha=1}^{N} \sum_{k=1}^{d_{l+1}} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,k;x_\alpha)} \cdot v_{k,r} \cdot \sigma''(\mathbf{n}(l,r;x_\alpha)) \mathbf{act}(l-1,i;x_\alpha) \mathbf{act}(l-1,j;x_\alpha) \tag{168}$$

and

$$\mathfrak{D}_i^{r,s} := \sum_{\alpha=1}^{N} \frac{\partial \ell_\alpha(x_\alpha, y_\alpha)}{\partial \mathbf{n}(l+1,s;x_\alpha)} \sigma'(\mathbf{n}(l,r;x_\alpha)) \mathbf{act}(l-1,i;x_\alpha). \tag{169}$$

Then, assume $\mathfrak{B}_{i,j}$ is either

- *positive definite and* $\lambda \in (0, 1)$*, or*
- *negative definite and* $\lambda \in (-\infty, 0) \cup (1, \infty)$*.*

Then the embedding $\gamma_\lambda^r(\cdot)$ *determines a local minimum in the larger network if and only if* $\mathfrak{D}_i^{r,s} = 0$ *for all* i, s*.*

Lemma 14 (Conditions on $\mathfrak{D}_i^{r,s} = 0$)**.** *Suppose for the outgoing weights* $v_{r,s}$ *of* $\mathbf{n}(l, r; x)$*, we have* $\sum_s v_{s,r} \neq 0$*. Then* $\mathfrak{D}_i^{r,s} = 0$ *if one of the following holds.*

- *The layer* l *is the last hidden layer.*
- *For all* t, t', α*, we have*

$$\frac{\partial \ell_\alpha}{\partial \mathbf{n}(l+1, t; x_\alpha)} = \frac{\partial \ell_\alpha}{\partial \mathbf{n}(l+1, t'; x_\alpha)}. \tag{170}$$

- *For each* α, t*,*

$$\frac{\partial \ell_\alpha}{\partial \mathbf{n}(l+1, t; x_\alpha)} = 0. \tag{171}$$

Remark 5. *Lemma 14 holds for* $1 - 1 - 1$ *neural network. In other words,* $D_i^{r,s} = 0$ *for the case* $d_X = d_1 = d_Y = 1$*.*

Definition 8. *Let*

$$H^{small} = \begin{bmatrix} \frac{\partial^2 \ell}{\partial u_{r,i} \partial u_{r,j}} & 2\frac{\partial^2 \ell}{\partial u_{r,i} \partial v_{s,r}} & \frac{\partial^2 \ell}{\partial \bar{w} \partial u_{r,i}} \\ 2\frac{\partial^2 \ell}{\partial u_{r,i} \partial v_{s,r}} & 4\frac{\partial^2 \ell}{\partial v_{s,r} \partial v_{t,v}} & 2\frac{\partial^2 \ell}{\partial \bar{w} \partial v_{s,r}} \\ \frac{\partial^2 \ell}{\partial \bar{w} \partial u_{r,i}} & 2\frac{\partial^2 \ell}{\partial \bar{w} \partial v_{s,r}} & \frac{\partial^2 \ell}{\partial \bar{w} \partial \bar{w}'} \end{bmatrix} \tag{172}$$

be the smaller matrix of Hessian H*.*