WHAT ILLNESS OF LANDSCAPE CAN OVER-PARAMETERIZATION ALONE CURE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Over-parameterized networks are widely believed to have nice landscape, but what rigorous results can we prove? In this work, we prove that: (i) from under-parameterized to over-parameterized networks, there is a phase transition from having sub-optimal basins to no sub-optimal basins; (ii) over-parameterization alone cannot eliminate bad non-strict local minima. Specifically, we prove that for any continuous activation functions, the loss surface of a class of over-parameterized networks has no sub-optimal basin, where "basin" is defined as the setwise strict local minimum. Furthermore, for under-parameterized network, we construct loss landscape with strict local minimum that is not global. We then show that it is impossible to prove "all over-parameterized networks have no sub-optimal local minima", by giving counter-examples for 1-hidden-layer networks with a class of neurons.

Viewing various bad patterns of landscape as illnesses (bad basins, flat regions, etc.), our results indicate that over-parameterization is not a panacea for every "illness" of the landscape, but it can cure one practically annoying illness (bad basins).

1 INTRODUCTION

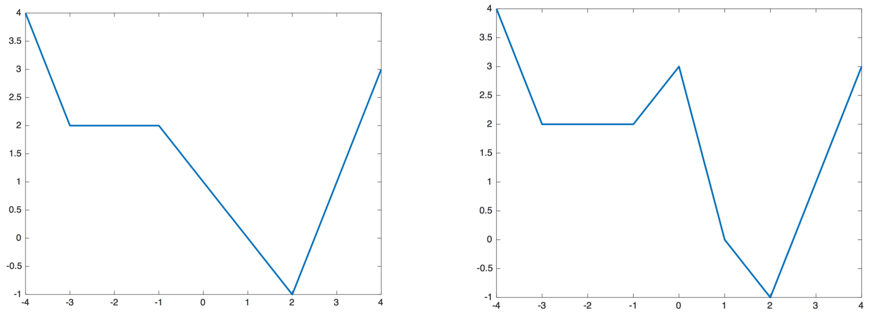
It is empirically observed that over-parameterized networks can achieve small training error Bengio et al. (2006); Zhang et al. (2017); Livni et al. (2014); Neyshabur et al. (2017); Geiger et al. (2018). These observations are often explained by the intuition that more parameters can "smooth the landscape" Livni et al. (2014); Lopez-Paz & Sagun (2018). While the intuition looks reasonable, can we provide some rigorous theory to formalize this intuition that "over-parameterized networks have nice landscape"?

Before answering this question, we would like to draw an analogy to the deep learning landscape. If we view the landscape as a human body, the bad properties (e.g., sub-optimal minima, saddle points, flat regions) can be regarded as different types of illness. Then, the aforementioned question become: Can over-parameterization cure all these diseases? If not, which specific type of illness can over-parameterization cure?

In this work, we aim at clearly identifying the "pharmacodynamics" of over-parameterization alone. To this end, we derive a serial of results on the loss landscape. Interestingly, our results are established in a way similar to a scientific experiment.

First, we hope to find out a property that holds for over-parameterized networks but not under-parameterized networks. Here the over-parameterized networks are like the treatment groups, while the under-parameterized networks are like the control group. In this way, we can claim that over-parameterization does really benefit the landscape.

Second, our theory should has minimal assumptions on other aspects except over-parameterization. Potentially, there are many other factors that affects the landscape. For example, with linear neurons Kawaguchi (2016) "no bad local inima" can be proved, implying that the choice of activation functions can impact the landscape. Thus, we should treat factors like activation function, loss function and training data as "control variables". We would like to minimize the effect of these factors, and see how far we can go based on merely the condition of over-parameterization.



(a)

(b)

Figure 1: An example of a weakly global function (a) and a non-weakly-global function (b). Both functions have bad non-strict local minima, consisting a plateau of $(-3, -1)$. The plateau in the right figure is the bottom of a basin, entailing a bad strict local minimum in the sense of sets.

To conclude, in this work we endeavor to explore the **"true benefit of over-parameterization alone"**.

- **"True"**: In terms of "true", we specify that the advantage of over-parameterization is not the commonly-believed "elimination of bad local minima", but "elimination of bad basin". On the positive side, we prove that bad basin does not exist in over-parameterized neural networks. On the negative side, we show that bad local minima do exist.
- **"Benefit"**: In terms of "benefit", we emphasize that "elimination of bad basin" is indeed a property brought by over-parameterization. To demonstrate this, we give an example to show that bad strict local minima exist in "exactly" under-parameterized neural networks, and analyze how over-parameterization makes a difference in the landscape of loss surface.
- **"Alone"**: In terms of "alone", we want to remove other confounding factors that may help the landscape. For instance, linear networks are known to have nice landscape, implying that activation functions can have a large impact on the landscape. Thus we want to reduce the impact of neuron activations by imposing mild condition on the activations.

More details are given in the next subsection.

1.1 Main Contributions

We first explain some geometrical concepts. A function is called "global function" if no sub-optimal local minimum exists. A function is called "weakly global function" if it admits no setwise bad strict local minima (defined in Josz et al. (2018)), as illustrated in Figure 1(a). Intuitively, weakly global functions may have flat regions of local minima (Figure 1(a)), but do not have "truly bad" local minima that are surrounded by barriers (Figure 1(b)).

A key geometrical property studied in our paper is called "Property PT": at any point, after a small perturbation there is a strictly decreasing path to the global minima. Property PT was proposed in the early work Yu & Chen (1995), and was claimed to imply "no bad local minima", but we provide a counter-example to show that this is not true. We show that Property PT actually implies "no bad setwise strict local minima" (i.e. weakly global function).

Our first result roughly states that for any continuous activation function, the loss function of over-parameterized networks (with a wide last hidden layer) is a weakly global function. Therefore, over-parameterized networks have nice landscape in the sense that truly bad local minima do not exist. We also extend this result to the case with one wide layer (not necessarily the last hidden layer) [1], though with extra assumptions on the structure and some neurons (see Assumption 2).

[1] This setting of network structure is more general than the first setting, and is quite close to practice: for instance, ImageNet dataset in ILSVRC competition has around 1.2 million training samples, while VGG and Inception network have 3 and 1.3 million neurons in the widest layer respectively.

Last but not least, we show by counter-example that, with other assumptions unchanged, under-parameterized networks may have sub-optimal basins, i.e, not weakly global.

We provide a brief summary of our contributions as follows:

- We give an example to show over-parameterized neural networks can have bad local minima for a large class of neurons (Section 6.2), which reveals the limitation of over-paramterized networks.
- We give an example to show that Property PT does not imply "no bad local minima" (Section 6.1). This shows that the proof of a classical work Yu & Chen (1995) on over-paramterized networks is incorrect. This example improves our geometrical understanding of the landscape.
- When the last hidden layer has at least N neurons (N is the number of samples), for any continuous activations, we prove that the loss function is weakly global (Theorem 1).
- When there is one wide layer with at least N neurons, under extra assumptions the loss function is a weakly global function (Theorem 2).
- We show that under-parameterized neural networks do have bad strict local minima, for any number of neurons and any generic data x. Our results clarify the role of over-parameterization in eliminating bad strict local minima.

1.2 Paper Organization

This paper is organized as follows. We first provide a geometric explanation of our main idea in Section 2 and review some related works in Section 3. In Section 4, we present the main results. In Section 5, we highlight the main proof ideas. Section 6 provides various examples of bad local minima. Conclusions are presented in Section 7.

2 Geometric Explanation

2.1 Landscape Analysis of the 1-Dimensional Case

Simple examples can go far: the phenomenons revealed by simple examples can often offer great insight about very general problems. One good example is $\min_{v,w\in\mathbb{R}}(vw-1)^2$: it is a non-convex function but it does not have bad local-min as the reparameterization does not destroy the nice landscape of the outer function. This insight can be extended to even the general deep linear networks Kawaguchi (2016). In this part, we discuss the 1-dim non-linear problem $(1-v\sigma(w))^2$. Interestingly, the geometrical phenomenons for this general 1-dim case still occur in deep over-parameterized networks (but not in under-parameterized networks). Therefore, to understand the major geometrical properties discussed in this paper, one only needs to understand this 1-dimensional example.

Consider a 1-hidden-layer-1-neuron network with network output $\hat{y} = v\sigma(wx)$. Assume that only one data sample pair $(x, y) = (1, 1)$ is given and the quadratic loss is employed. Thus, the empirical loss is given by $E(w, v) = (y - \hat{y})^2$. Then, with the given sample pair, the empirical loss is represented by $(1 - v\sigma(w))^2$.

The landscape of the considered network is determined by the activation function. In particular, we discuss the following three cases:

1. The activation function is linear, i.e., $\sigma(z) = z$. The empirical loss is thus $(1 - uv)^2$, which obviously does not contain bad local minima.
2. The activation function is nonlinear and strictly monotone. Bad local minima also do not exist in this case.
3. The activation function is nonlinear and non-monotone. In this case, bad local minima may exist. For example, the activation function is quadratic, i.e. , $\sigma(z) = z^2$. Points on line segment $l : \{(w, v) \mid w = 0, v < 0\}$ are all spurious local minima with a loss of 1. To see this, note that vw^2 is non-positive in the neighborhood of line segment l, so $(1 - vw^2)^2$ is always at least 1. One important feature in this example is that $w = 0$ is a local minimum of

$\sigma(w)$. This means that after perturbing w, $\sigma(w)$ can only move in one direction so that the sign of vw^2 can be controlled.

Case 1 is generalizable to deep networks: as mentioned earlier, the geometrical property of deep linear networks is similar to 1-dim linear problem. Which case can reflect the geometrical property of a deep non-linear over-parameterized networks? Currently, it seems that Case 2 cannot be extended: it is still open whether increasing activation can eliminate bad local-min [2]. This is why we do not want to focus on increasing activations since there is a gap between the simplest case and the general case.

Our paper intends to show that Case 3 captures the major geometrical properties of a general deep non-linear over-parameterized networks: it has bad non-strict local-min, but it has no bad basin. The intuition comes as the following: in the example constructed, we can find that these bad local minima in fact lie in a plateau (so they are non-strict). Moreover, at the boundary of the plateau there is a decreasing direction. This implies that these bad local minima do not lie in a basin, so theoretically they are possible to escape with algorithms like noisy GD. In the rest of this paper, we will show that this is the case for all deep over-parameterized neural networks, i.e., all deep over-parameterized neural networks do not contain bad basins. This result characterizes what benefits over-parameterization alone brings to general neural networks.

2.2 Strict Local Minimum Exists for Under-parameterized Neural Networks

One may wonder whether the previous 1-dim example actually reveals the insight of *all* networks, not just over-parameterized networks. We show that the answer is no: for under-parameterized networks, strict bad local-min exists for fairly general settings. Our examples also explains the geometrical reasons how over-parameterization "cures" the illenss of bad basins.

Suppose there are two sample pairs $(x_1, y_1) = (1, 2), (x_2, y_2) = (2, 0)$. Consider a 1-hidden-layer-1-neuron neural network $\hat{y} = v\sigma(wx)$. Let $(v, w) = (1, 0)$ and $\sigma(z) = z^2(1 - z^2) + 1$. Then $(\hat{y}_1, \hat{y}_2) = (1, 1)$. We plot the figure of the output data and activation function $\sigma(\cdot)$ separately.

We will now show that $(v, w) = (1, 0)$ is a bad strict local minimum through these two figures. In Figure 2(a), we mark out the true data $(y_1, y_2) = (2, 0)$ and the current output $(\hat{y}_1, \hat{y}_2) = (1, 1)$. Note that $\sigma(wx_1) = \sigma(wx_2) = \sigma(0) = 1$, if we change only v and not w, $(\hat{y}'_1, \hat{y}'_2) = (v'\sigma(wx_1), v'\sigma(wx_2)) = (v', v')$. This means that if we only perturb v, the new output $(\hat{y}'_1, \hat{y}'_2)$ will move along line $h = \{(e_1, e_2) : e_1 = e_2\}$ (illustrated as the red line in Figure 2(a)). On the other hand, $(\hat{y}_1, \hat{y}_2) = (1, 1)$ is exactly the projection of (y_1, y_2) to line h. Consequently, perturbing v alone will move the new output $(\hat{y}'_1, \hat{y}'_2)$ away from the true data (y_1, y_2), thus strictly increasing the empirical loss (distance).

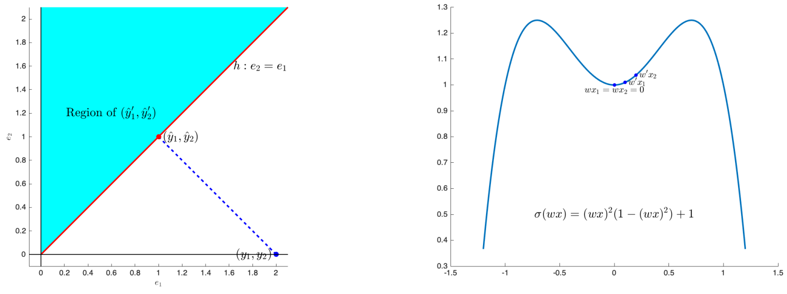
What if we perturb w simultaneously? In Figure 2(b) we plot the value of $\sigma(wx)$ before and after perturbation. Before perturbation, $\sigma(wx_1) = \sigma(wx_2) = 1$. After perturbation, both $\sigma(w'x_1)$ and $\sigma(w'x_2)$ rise above 1. However, since $x_2 = 2 > 1 = x_1$, $\sigma(w'x_2)$ is always larger than $\sigma(w'x_1)$ due to the special landscape of $\sigma(\cdot)$. This implies that we always have $\hat{y}'_2 = v'\sigma(w'x_2) > v'\sigma(w'x_1) = \hat{y}'_1$. Therefore, the feasible region of $(\hat{y}'_1, \hat{y}'_2)$ is above line h (illustrated as the cyan region in Figure 2(a)), and hence any perturbation of (v, w) will result in a larger distance between $(\hat{y}'_1, \hat{y}'_2)$ and (y_1, y_2), increasing the empirical loss.

From the analysis above, we can conclude that a bad strict local minimum occurs when

- Due to under-parameterization, the perturbation direction provided by the last hidden-layer lies in a hyperplane of the whole space.
- Due to landscape of activation function, the perturbation direction provided by the previous hidden layers lies in a halfspace.

The first condition demonstrates the key difference between under-parameterized and over-parameterized neural networks. For over-parameterized networks, one of the following two circumstances happens: a) the perturbation direction provided by the last hidden-layer is the whole

[2] assuming generic data. For very special data, constructing bad local-min is possible but it is not that interesting, since the proofs of positive results often assume generic data.



(a)

(b)

Figure 2: The figure of the output data (a) and activation function $\sigma(\cdot)$ (b) in the proposed 1-hidden-layer-1-neuron example. In (a), $(y_1, y_2) = (2, 0)$ is the true data, $(\hat{y}_1, \hat{y}_2) = (1, 1)$ is the current output and the cyan area is the feasible region of $(\hat{y}'_1, \hat{y}'_2)$ if we perturb both v and w. In (b), $wx_1 = wx_2 = 0$ is the current input of $\sigma(\cdot)$, while $w'x_1$ and $w'x_2$ are possible inputs after perturbation.

space, implying that a decreasing path exists if the current point is not a global minimum; (b) the perturbation direction provided by the last hidden-layer still lies in a hyperplane of the whole space. In this circumstance, since neurons are over-parameterized, the input of the last hidden layer is linearly dependent. Hence, there must exist a perturbation of the last hidden layer that keeps the objective function value unchanged, eliminating bad strict local minimum. Thus over-parameterized networks can not have bad strict local minimum. In contrast, for under-parameterized networks, even if the the input of the last hidden layer is linearly independent, the perturbation direction provided by the last hidden-layer still lies in a hyperplane of the whole space. If the true data does not lie in this hyperplane, bad strict local minimum may exist.

The second condition demonstrates that existence of bad local minimum is highly related to the perturbation direction determined by the landscape of activation function. The landscape that leads to a restricted perturbation direction, as analyzed in the proof, is in fact rather common and not specified to particular activation functions. This finding implies that existence of bad strict local minimum is commonly seen in under-parameterization networks. It is noteworthy that over-parameterization, on the contrary, does not play a decisive role in perturbation direction. That's why bad local minimum still exists in over-parameterization networks. However, due to the first condition, bad strict local minimum is eliminated by over-parameterization.

Finally, this example can be generalized to under-parameterized neural networks with any width through the following proposition:

Proposition 1 *For any N input data $x_1, \cdots, x_N \in \mathbb{R}$ with $x_n \neq x'_n, \forall n \neq n'$, there exists N output data $y_1, \cdots, y_N \in \mathbb{R}$ and a 1-hidden-layer neural network with $N - 1$ neurons such that the empirical loss $E(\cdot)$ has bad strict local minimum.*

It is noteworthy that the requirement for $\sigma(\cdot)$ is quite mild. We only need to characterize first-order and second-order derivatives of $\sigma(\cdot)$ at specific points rather than in the whole domain of $\mathbb{R}$. This implies that such $\sigma(\cdot)$ we construct can possess good properties such as monotonicity. Furthermore, we can relax the requirement for the second-order derivative to be $\sum_{i=1}^{N} x_i^2 \cdot a_i \sigma''(w_j x_i) > 0$ for all $j = 1, \cdots, N - 1$. With this relaxed requirement, $\sigma(\cdot)$ can even be convex.

3 Related Works and Discussions

We notice that our results may have a strong connection to the "lottery ticket hypothesis" Frankle & Carbin (2019). It was found in Frankle & Carbin (2019) that a large (and generally over-parameterized) feedforward network contains a subnetwork that, when properly initialized, can be trained to achieve similar performance to the original network, i.e., the "wining ticket". However, if such subnetwork is randomly initialized, it no longer matches the performance of the original network. This experiment is nice as it rules out the effect of representation power, and indicates that the worse performance of small networks is due to optimization, not due to representation power. Which part of the optimization makes the difference? Early works Dauphin et al. (2014) conjectured that saddle points cause bad performance, but there is no clear evidence that saddle points exist in small networks but not in large networks. Our result proved rigorously that at least bad basin is a difference between small and big networks. Thus based on our result, we conjectured that the lottery tickets are due to bad basins. The verification of this conjecture is beyond the scope of this paper (since checking basins is computatationally expensive), and we leave it to future work.

The loss surface of neural networks has been extensively studied in recent years Du & Lee (2018); Ge et al. (2018); Andoni et al. (2014); Sedghi & Anandkumar (2014); Janzamin et al. (2015); Haeffele & Vidal (2015); Gautier et al. (2016); Brutzkus & Globerson (2017); Soltanolkotabi (2017); Soudry & Hoffer (2017); Goel & Klivans (2017); Boob & Lan (2017); Du et al. (2017); Zhong et al. (2017); Li & Yuan (2017); Liang et al. (2018a); Mei et al. (2018); Sirignano & Spiliopoulos (2018a;b); Safran & Shamir (2018); Wang et al. (2018); Chizat & Bach (2018); Li & Liang (2018). Most works are only about one-hidden-layer networks (many assume second layer weights are fixed, and many have strong assumptions on data distributions). There are some works on deep neural networks, but they either only show a subset of local minima are global minima Haeffele et al. (2014); Haeffele & Vidal (2015); Soudry & Carmon (2016); Nguyen & Hein (2017a;b); Shamir (2018), or require special neurons such as linear neurons Baldi & Hornik (1989); Kawaguchi (2016); Freeman & Bruna (2016); Hardt & Ma (2017); Yun et al. (2017).

Some very recent works Du et al. (2018a); Allen-Zhu et al. (2018); Zou et al. (2018) prove that GD (gradient descent) converges to global minima at a linear rate for deep neural networks, which extend earlier works on 1-hidden-layer networks Li & Liang (2018); Du et al. (2018b). Their conclusions are stronger than our landscape analysis since they can prove linear convergence, but their assumption is also stronger. In particular, these works require a large number of neurons, e.g. at least $O(n^k)$ neurons in every layer for deep networks, where n is the number of samples and k is a certain integer such as $k = 24$ in Allen-Zhu et al. (2018). In practical neural networks such as VGG and ResNet, the number of neurons is in the order of $O(n)$, and even $O(n^2)$ neurons is difficult to implement in practice. It will be interesting to prove results that combine the strengths of those works and our landscape analysis.

Moreover, two recent works Nguyen et al. (2018); Liang et al. (2018b) have addressed similar issues to our work. Specifically, Nguyen et al. (2018) covers a rather broad range of network structures, but only holds for a limited family of activation functions, which does not include activation functions such as ReLU, leaky ReLU, quadratic and Swish activation Ramachandran et al. (2018). Liang et al. (2018b) eliminates bad local minima by adding a special type of exponential neuron. In contrast, our result holds for any continuous activation function, so is closer to practical settings in this sense. In addition, these works do not show a phase transition from under-parameterized to over-parameterized networks.

Finally, landscape analysis is just one part of the deep learning theory, which includes representation, optimization and generalization. In terms of generalization, many recent works Neyshabur et al. (2017); Bartlett et al. (2017); Poggio et al. (2018) try to understand why over-parameteriztion does not cause overfitting. This is a very interesting line of research, but its underlying assumption that over-parameterization can lead to small training error still requires rigorous justification. Our study is orthogonal and complimentary to the research on generalization error.

4 Main Theorems

4.1 Network Model

We consider a fully-connected neural network $f_W : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ that maps an input data $x \in \mathbb{R}^{d_x}$ to a predicted output $\hat{y} \in \mathbb{R}^{d_y}$. Specifically, the network output is given by

$$\hat{y} = f_W(x) = W_{H+1}\sigma_H(W_H \cdots \sigma_2(W_2\sigma_1(W_1 x))) \tag{1}$$

where H is the number of hidden layers, $\sigma : \mathbb{R}_h \to \mathbb{R}$ is the neuron activation function (sometimes simply called "activation" or "neuron") of the h-th hidden layer, $W_h \in \mathbb{R}^{d_h \times d_{h-1}}$ is the weight matrix to the h-th hidden layer for $h = 1, \cdots, H$, $W_H \in \mathbb{R}^{d_{H+1} \times d_H}$ is the weight matrix to the output layer. Note that we define $d_0 = d_x$ and $d_{H+1} = d_y$, and the scalar function σ_h is applied entry-wise if its input is a vector or a matrix. Throughout the paper, we omit the "bias" term in the expression of neural networks for simplicity of presentation.

Let $W = (W_1, \cdots, W_H)$ denote the collection of all network weights. Given N pairs of training data $(x^{(n)}, y^{(n)})$, $n = 1, \cdots, N$, the empirical loss of the considered network as

$$E(W) = l(Y, \hat{Y}) \tag{2}$$

where $Y \triangleq [y^{(1)}, \cdots, y^{(N)}] \in \mathbb{R}^{d_y \times N}$, $\hat{Y} \triangleq [f_W(x^{(1)}), \cdots, f_W(x^{(N)})] \in \mathbb{R}^{d_y \times N}$, and $l(\cdot, \cdot)$ is the loss function. Then, the training problem of the considered network is to find W that minimize the empirical loss $E(W)$.

4.2 Main Theorems

In this section, we present our main result on the absence of sub-optimal basin on the loss surface. Specifically, we define "basin" as the setwise strict local minimum, a notion borrowed from Josz et al. (2018).

Definition 1 (Setwise strict local minimum) *We say that a compact subset $X \in S$ is a strict local minimum of $f : S \to \mathbb{R}$ in the sense of sets if there exists $\varepsilon > 0$ such that for all $x \in X$ and for all $y \in S \setminus X$ satisfying $\|x - y\|_2 \leq \varepsilon$, it holds that $f(x) < f(y)$.*

Definition 1 generalizes the notion of strict local minimum from the sense of points to the sense of sets. Any strict local minimum must be setwise strict local minimum, but not vice versa. A pleateau at the bottom of a basin, as shown in the right part of Figure 1, is also a setwise strict local minimum.

Definition 2 (Weakly global function) *We say that $f : S \to \mathbb{R}$ is a weakly global function if it is continuous and all setwise strict local minima are setwise global minima.*

Definition 2 introduces an important class of continuous functions, termed weakly global functions, which admits no bad strict local minima in the sense of sets, and hence no sub-optimal basin.

Now we specify our assumptions for the first theorem. on the training dataset, the loss functions, the over-parameterization, and the activation functions.

Assumption 1

A1 There exists a dimension k such that $x_k^{(n)} \neq x_k^{(n')}, \forall n \neq n'$;

A2 The loss function $l(Y, \hat{Y})$ is convex respect to $\hat{Y}$;

A3 $d_H \geq N$;

A4 The activation function σ_h is continuous for all $h = 1, \cdots, H$.

Assumption A1 implies that the input data samples need to be distinguished with each other in one dimension. This can be always achieved if we allow an arbitrarily small perturbation on data. Assumption A2 is satisfied for almost all commonly-used loss functions, including quadratic, cross entropy, etc. Assumption A3 is the over-parameterization assumption, which only requires the last hidden layer to be wide. There is no assumption on the width of all other hidden layers. Assumption A4 is a very mild assumption on the neuron activation that it should be continuous.

We are now ready to present the first main theorem.

Theorem 1 *Suppose that a fully connected neural network satisfies Assumption 1. Then, the empirical loss $E(W)$ is a weakly global function.*

Theorem 1 states that the empirical loss function of an over-parameterized neural network is weakly global as long as the activation function is continuous. Note that the notion of weakly global function is distinct from that of “no bad local valleys" used in Nguyen et al. (2018); Venturi et al. (2018), but they both guarantee non-existence of bad strict local minimum. Formally, we have the following corollary.

Corollary 1 *The loss surface of a fully connected neural network satisfying Assumption 1 has no bad strict local minimum.*

Theorem 1 requires that the last hidden layer is sufficiently “wide". Next, we show that if we add a pyramid structure after the “wide” hidden layer, the resulting loss surface still has no sub-optimal basin. We specify such pyramid structure by the the following assumption.

Assumption 2 *There exists* $1 \leq h_0 \leq H$ *such that*

B1 $d_{h_0} \geq N$, $d_{h_0} \geq d_{h_0+1} \geq \cdots \geq d_{H+1}$;

B2 For all $h_0 + 1 \leq h \leq H$, *the activation function* σ_h *is non-increasing or non-decreasing over* $\mathbb{R}$.

Then it can be shown that our main result still holds after adding the pyramid architecture:

Theorem 2 *Suppose that a fully connected neural network satisfies Assumption A1, A2, A4, and 2. Then, the empirical loss* $E(W)$ *is a weakly global function.*

Corollary 2 *The loss surface of a fully connected neural network satisfying Assumption A1, A2, A4, and 2 has no bad strict local minimum.*

5 Proof Ideas

5.1 Property that Eliminates Bad Strict Local Minimum

Consider a 1-hidden-layer network with d input dimensions and m hidden-layer neurons. The output of the network is given by $t = v^\top \sigma(WX) \in \mathbb{R}^{1\times N}$, where $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{d\times N}$ is the input data matrix consisting of N data samples, $W \in \mathbb{R}^{m\times d}$ is the weights to the hidden layer, $v \in \mathbb{R}^m$ is the weights to the output layer. The loss function is quadratic, i.e., $l(y,t) = ||t-y||_2^2$. The network is assumed to be over-parameterized, i.e., $m > N$.

Now we denote $Z = \sigma(WX) \in \mathbb{R}^{m\times N}$ as the output matrix of the hidden layer. The training of the considered network can be formulated as

$$\min_{v,W} \left|\left|t - v^\top Z\right|\right|^2 \tag{3}$$

A simple but important observation Gori & Tesi (1992) is that if Z is of full column rank, i.e., rank-N, problem equation 3 (with fixed W) is strongly convex with respect to v. Then, we can find a strictly decreasing path to the global infimum (with respect to v) of equation 3. Moreover, for any t, there exists v such that $v^\top Z = t$, so the infimum equals 0, which is the global infimum ofequation 3. Thus, there is a strictly decreasing path to the global infimum of equation 3, and hence the network parameter (W, v) cannot be a bad local minimum if the corresponding Z is rank-N.

It is noteworthy that when Z is not full rank, the infimum (with respect to v) of equation 3 may not be equal to the global infimum of equation 3, so the previous decreasing-path approach does not work. (Obviously, matrix Z is not always rank-N. For example, if W is a zero matrix, the corresponding Z is at most rank-1 regardless of the number of hidden neurons and activation function.) Regarding this, we introduce a property that can eliminate all bad setwise strict local minima on the loss surface.

Property PT: After a generic perturbation of any initial weight, there exists a strictly decreasing path to the global infimum.

We claim that if Property PT holds, the loss function with respect to (v, W) is a weakly global function, which admits no setwise bad strict local minimum. Note that weakly global function also implies the non-existence of (point-wise) bad strict local minimum.

Whether Property PT holds highly depends on the activation function. In this paper, we identify a class of activation functions that guarantee Property PT. Specifically, we show that if the activation function is analytic and has non-vanishing derivatives at zero up to N-th order, Property PT holds. This enables us to prove that the loss surface is weakly global. However, to extend this result to all continuous activation functions remains a challenging task.

5.2 No Bad Strict Local Minimum for Continuous Activations

So far, we have identified a class of analytic activation functions such that the loss functions have Property PT. Although these functions do not cover many commonly used activations like quadratic, sigmoid, ReLU, etc, they constitute a dense set in the space of continuous functions. That is, for any continuous function f, there exists a sequence of analytic functions that uniformly converges to f, and each of the analytic functions has non-vanishing derivatives at zero up to N-th order. Further, we show that the uniform convergence of activation function implies the compact convergence of loss function. As the property of weakly global is preserved under compact convergence, we prove that the loss surface weakly global for any continuous activation.

To summarize, our proof can be sketched in the following three steps:

Step 1: If the activation function has non-vanishing derivatives at zero up to N-th order, then Property PT holds, and the loss surface is a weakly global function.

Step 2: The set of activation functions in Step 1 are dense in the space of continuous functions.

Step 3: By the “closedness" of weakly global functions, the loss surface for any continuous activation is a weakly global function.

We note that in Step 3, the “closedness" of weakly global functions means that if a sequence of weakly global functions converges compactly to a function, then that function is also weakly global. This is a slight extension of a proposition in Josz et al. (2018).

6 Examples of Bad Local Minima

In this section, we will give some counter-examples to illustrate the “tightness” of our conditions (cannot be weakened too much) and conclusion (cannot be strengthened under the current setting).

6.1 Property PT Does Not Imply “No Bad Local Minima”

It is worth mentioning that our result only implies the non-existence of bad *strict* local minima, instead of bad local minima. This is in constrast with what Yu et al. claimed in Yu & Chen (1995) that if Property PT holds, an over-parameterized single-hidden-layer neural network has no bad local minimum (i.e., Theorem 3 in Yu & Chen (1995)). While this claim seems correct (in fact, it is correct in one dimensional case), we could not provide a rigorous proof to this claim. Later, we found a counter-example to this claim, which implies that Property PT can only eliminate bad *strict* local minima, but cannot eliminate bad local minima.

This counter-example is drawn in Fig. 3, which illustrates the loss surface of Example 1 with one hidden neuron. The point $(v, W) = (1, 1)$ is a bad non-strict local minimum. However, after a generic perturbation within a sufficiently small neighbourhood of $(1, 1)$, there exists a strictly decreasing path to the global minimum $(-2, -1)$.

6.2 Bad (Non-strict) Local Minima Can Exist for Wide Networks

Although Property PT does not imply the non-existence of bad local minima, one may wonder whether it can be proved by other proof techniques. In particular, is it possible to prove the non-existence of bad local minima under the same setting of Yu et al. Yu & Chen (1995)? Below, we

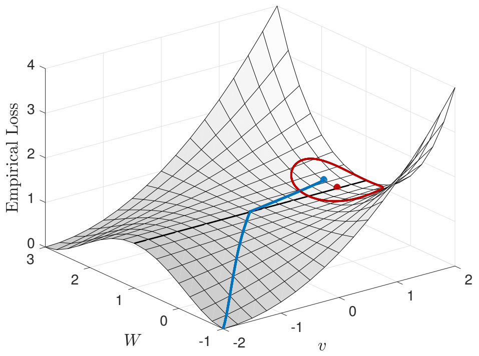


Figure 3: The loss surface of Example 1 with $d = 1$, $(x, y) = (1, 1)$, and $\sigma(z) = -(z-1)^2/8$. The point $(v, W) = (1, 1)$ is a bad non-strict local minimum.

give a counter-example of Theorem 3 in Yu & Chen (1995) that for a class of neuron activations that satisfy the condition of Yu et al. Yu & Chen (1995)[3], bad local minima exist.

Example 1 (Bad non-strict local minimum) *Consider a 1-hidden-layer neural network with 1 data sample, 1 input dimension, and d hidden-layer neurons. The data sample satisfies $x, y \neq 0$. The activation function σ is analytic and has non-vanishing derivatives at zero up to N-th order. Thus, Property PT holds. Moreover, suppose that there exists $t \neq 0$ and δ such that $\sigma(t) = 0$ and $\sigma(t') \leq 0$ for $t - \delta < t' < t + \delta$.*

Now, let $v^* = (sign(y), sign(y), \cdots, sign(y))^\top$ and $W^* = (t/x, t/x, \cdots, t/x)^\top$. Then, ${v^*}^\top \sigma(W^* x) = 0$ and the corresponding loss $l(v^*, W^*) = y^2 > 0$. In addition, since $y \cdot \big(v^\top \sigma(Wx)\big)$ is always non-positive in the neighborhood of (v^*, W^*), $l(v, W)$ is always at least y^2 in the neighborhood of (v^*, W^*), implying that (v^*, W^*) is a bad local minimum.

Example 1 shows that the loss surface can have bad non-strict local minimum even if the activation function has Property PT and the neural network is arbitrarily wide. It also shows the necessity of rigorous mathematical analysis: while it is widely believed that over-parameterized networks can “smooth” the landscape, the exact notion of “smoother landscape” was not clearly stated.

7 Conclusions

In this paper, we studied the loss surface of over-parameterized fully connected deep neural networks. We show that if the last hidden layer has no less neurons than the number of samples, for any continuous activation functions, the loss function is a weakly global function, i.e., the loss landscape has no setwise strict local minima (bad basins). We also show that for almost all analytic activation functions, starting from any point, after a small perturbation there exists a strictly decreasing path to the global infimum. We then extend our result to the deep neural network with one wide layer (not necessarily the last hidden layer) under extra assumptions. On the other hand, for a single-hidden-layer network with hidden neurons less than the data samples, we construct an example to show that sub-optimal strict local minimum exists. A geometric explanation is further provided to demonstrate how over-parameterization helps in eliminating bad basins.

In a word, our work shows that over-parameterization cannot bring perfect landscape to deep neural networks. However, it does really benefit the landscape by eliminating setwise strict local minimum. Our work clarify the role of over-parameterization alone in smoothing the loss landscape.

[3] Although Yu et al. Yu & Chen (1995) assumes sigmoid activation, their proof only uses the property that the activation has non-vanishing derivatives up to the N-th order, which holds for sigmoid. Such property also holds for a broad range of activations considered in the counter-example Example 1, thus this counter-example can be viewed as a counter-example to the setting of Yu & Chen (1995).

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang. Learning polynomials with neural networks. In *ICML*, 2014.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *NIPS*, pp. 123–130, 2006.

D. Boob and G. Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

A. Brutzkus and A. Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, pp. 2933–2941, 2014.

S. S Du and J. D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

S. S. Du, J. D. Lee, and Y. Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *ICLR*, 2019.

C D. Freeman and J. Bruna. Topology and geometry of half-rectified network optimization. *ICLR*, 2016.

A. Gautier, Q. N. Nguyen, and M. Hein. Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods. In *NIPS*, pp. 1687–1695, 2016.

R. Ge, J. D Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. *ICLR*, 2018.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *arXiv preprint arXiv:1809.09349*, 2018.

S. Goel and A. Klivans. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

Marco Gori and Alberto Tesi. On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):76–86, 1992.

B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *ICML*, 2014.

B. D Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

M. Hardt and T. Ma. Identity matters in deep learning. *ICLR*, 2017.

M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

C Josz, Y Ouyang, UC IEOR, RY Zhang, J Lavaei, and S Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. *NIPS*, 2018.

Wilfred Kaplan. Approximation by entire functions. *Michigan Math. J.*, 3(1):43–52, 1955. doi: 10.1307/mmj/1031710533. URL `https://doi.org/10.1307/mmj/1031710533`.

K. Kawaguchi. Deep learning without poor local minima. In *NIPS*, pp. 586–594, 2016.

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. In *NIPS*, pp. 597–607, 2017.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

S. Liang, R. Sun, Y. Li, and R. Srikant. Understanding the loss surface of neural networks for binary classification. 2018a.

Shiyu Liang, Ruoyu Sun, Jason D Lee, and R Srikant. Adding one neuron can eliminate all bad local minima. *NIPS*, 2018b.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pp. 855–863, 2014.

David Lopez-Paz and Levent Sagun. Easing non-convex optimization with neural networks. 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561*, 2018.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Q. Nguyen and M. Hein. The loss surface and expressivity of deep convolutional neural networks. *arXiv preprint arXiv:1710.10928*, 2017a.

Q. Nguyen and M. Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017b.

Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. *arXiv preprint arXiv:1809.10749*, 2018.

T Poggio, K Kawaguchi, Q Liao, B Miranda, L Rosasco, X Boix, J Hidary, and HN Mhaskar. Theory of deep learning iii: the non-overfitting puzzle. Technical report, Technical report, CBMM memo 073, 2018.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. 2018.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *ICML*, 2018.

H. Sedghi and A. Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

O. Shamir. Are resnets provably better than linear predictors? *arXiv preprint arXiv:1804.06739*, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018a.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *arXiv preprint arXiv:1808.09372*, 2018b.

M. Soltanolkotabi. Learning relus via gradient descent. In *NIPS*, pp. 2004–2014, 2017.

D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

D. Soudry and E. Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

Luca Venturi, Afonso Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

Gang Wang, Georgios B Giannakis, and Jie Chen. Learning relu networks on linearly separable data: Algorithm, optimality, and generalization. *arXiv preprint arXiv:1808.04685*, 2018.

Xiao-Hu Yu and Guo-An Chen. On the local minima free condition of backpropagation learning. *IEEE Transactions on Neural Networks*, 6(5):1300–1303, 1995.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. *arXiv preprint arXiv:1707.02444*, 2017.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

K. Zhong, Z. Song, P. Jain, P. L Bartlett, and I. S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *ICLR*, 2017.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

A Preliminaries

In order to provide rigorous proof and analysis, we re-introduce the considered network model and relevant notations in a more detailed ways.

Consider a fully connected neural network with H hidden layers. Assume that the h-th hidden layer contains d_h neurons for $1 \leq h \leq H$, and the input and output layers contain d_0 and d_{H+1} neurons, respectively. Given an input sample $x \in \mathbb{R}^{d_0}$, the output of the i-th neuron in the h-th hidden layer, denoted by $t_{h,i}$, is given by

$$t_{1,i}(x) = \sigma_1 \left(\sum_{k=1}^{d_0} w_{1,i,k} x_k + b_{1,i} \right), \quad 1 \leq i \leq d_1 \tag{4a}$$

$$t_{h,i}(x) = \sigma_h \left(\sum_{k=1}^{d_{h-1}} w_{h,i,k} t_{h-1,k}(x) + b_{h,i} \right), \quad 2 \leq h \leq H, \quad 1 \leq i \leq d_h, \tag{4b}$$

where $w_{h,i,k}$ is the weight from the k-th neuron of the $(h-1)$-th layer to the i-th neuron of the h-th layer, $b_{h,i}$ is the bias added to the i-th neuron of the h-th layer, and σ_h is the neuron activation function for the h-th hidden layer. Then, the i-th output of the network, denoted by $t_{H+1,i}$, is given by

$$t_{H+1,i}(x) = \sum_{k=1}^{d_H} w_{H+1,i,k} t_{H,i}(x), \quad 1 \leq i \leq d_{H+1} \tag{5}$$

where $w_{H+1,i,k}$ is the weight to the output layer, defined similarly to that of the hidden layers.

Consider a training dataset consisting of N samples. Denote the n-th sample by $(x^{(n)}, y^{(n)})$, $n = 1, \cdots, N$, where $x^{(n)} \in \mathbb{R}^{d_0}$ and $y^{(n)} \in \mathbb{R}^{d_{H+1}}$ are the input and output samples, respectively. In what follows, we rewrite all the training samples in matrix forms, which allows us to represent the input-output relation of the neural network in a more compact way. Specifically, let $X \triangleq [x^{(1)}, x^{(2)}, \cdots, x^{(N)}] \in \mathbb{R}^{d_0 \times N}$ and $Y \triangleq [y^{(1)}, y^{(2)}, \cdots, y^{(N)}] \in \mathbb{R}^{d_{H+1} \times N}$ as the input and output data matrices, respectively. Then, we define $W_h \in \mathbb{R}^{d_{h-1} \times d_h}$ as the weight matrix from the $(h-1)$-th layer to the h-th layer, $\mathbf{b}_h \in \mathbb{R}^{d_h}$ as the bias vector of the h-th layer, and $T_h \in \mathbb{R}^{d_h \times N}$ as the output matrix of the h-th layer. The entries of each matrix are given by

$$(W_h)_{i,k} = w_{h,i,k}, \quad (\mathbf{b}_h)_i = b_{h,i}, \quad (T_h)_{i,n} = t_{h,i}(x^{(n)}). \tag{6}$$

Based on the above definition, we can immediately rewrite the output of each layer as

$$T_1 = \sigma_1 \left([W_1 \quad \mathbf{b}_1] \begin{bmatrix} X \\ \mathbf{1}^\top \end{bmatrix} \right), \tag{7a}$$

$$T_h = \sigma_h \left([W_h \quad \mathbf{b}_h] \begin{bmatrix} T_{h-1} \\ \mathbf{1}^\top \end{bmatrix} \right), \quad h = 2, 3, \cdots, H, \tag{7b}$$

$$T_{H+1} = W_{H+1} T_H. \tag{7c}$$

where the activation function $\sigma_h(\cdot)$ is applied entry-wise to the input matrix and outputs a matrix with the same size. That is, $(\sigma(A))_{i,j} = \sigma(A_{i,j})$ for any input matrix A.

In the rest of this paper, we simplify the feed-forward operation equation 7 by ignoring all the bias neurons, yielding

$$T_1 = \sigma(W_1 X), \tag{8}$$

$$T_i = \sigma(W_i T_{i-1}), \quad i = 2, 3, \cdots, H, \tag{9}$$

$$T_{H+1} = W_{H+1} T_H. \tag{10}$$

We note that this simplification does not affect our analysis, and therefore the main results also hold for feed-forward deep neural networks with bias. Let $W = (W_1, \cdots, W_{H+1})$ denote all the weights and define the empirical loss as

$$E(W) = l(Y, T_{H+1}) = l(Y, W_{H+1} T_H) \tag{11}$$

where l is the loss function. Then, the training problem of the considered network is to find W to minimize the empirical loss $E(W)$.

B Proof of Proposition 1

Proof: Denote $x = (x_1, \cdots, x_N)^\top$. Let the neural network be defined by $\hat{y} = v^\top \sigma(wx^\top)$ where $v = (v_1, \cdots, v_{N-1})^\top$ and $w = (w_1, \cdots, w_{N-1})^\top$ are weights to the output layer and the hidden layer, respectively. Let $l(y, \hat{y}) = \|y - \hat{y}\|^2$, then $E(\cdot) = E(v, w) = \sum_{i=1}^N (y_i - v^\top \sigma(wx_i))^2$. In the following, we generically choose w and v, and then in turn determine the activation function $\sigma(\cdot)$ and the output data y to obtain the empirical loss $E(v, w)$ with bad strict local minimum.

Denote $z_j = (\sigma(w_j x_1), \cdots, \sigma(w_j x_N))^\top, j = 1, \cdots, N-1$. Since $z_1, \cdots, z_{N-1}$ are $N-1$ vectors in $\mathbb{R}^N$, they must be in a $N-1$-dimensional hyperplane. Assume that this hyperplane, namely $\mathcal{H}$, is represented as $a_1 e_1 + \cdots + a_N e_N = 0$, where $e_1, \cdots, e_N$ are the N orthogonal unit vectors. Denote $a = (a_1, \cdots, a_N)^\top$, then we have

$$a^\top z_j = \sum_{i=1}^N a_i \sigma(w_j x_i) = 0, \forall j = 1, \cdots, N-1.$$

Without loss of generosity, assume that

- $\sigma(\cdot)$ is twice-differentiable;
- $w_j x_i \neq w_{j'} x_{i'}$ if $i \neq i'$ or $j \neq j'$;

- $a_i \neq 0$ for all $i = 1, \cdots, N$.
- $z_1, \cdots, z_{N-1}$ are linearly independent.

Otherwise we can perform an arbitrarily small perturbation of w to make the assumptions hold.

We now come to the key step: determine $\sigma(\cdot)$ to guarantee that there exists a neighborhood $B(w, \delta)$ such that

$$\sum_{i=1}^{N} a_i \sigma(w'_j x_i) > 0, \forall j = 1, \cdots, N-1.$$

for all $w' \in B(w, \delta) \backslash \{w\}$. Since $\sigma(\cdot)$ is twice-differentiable, we can properly determine the first-order and second-order derivative of $\sigma(\cdot)$ at all points of $w_j x_i$ for $j = 1, \cdots, N-1$ and $i = 1, \cdots, N$. Specifically, let

$$\sum_{i=1}^{N} a_i x_i \sigma'(w_j x_i) = 0, \ \forall j = 1, \cdots, N-1,$$
$$a_i \cdot \sigma''(w_j x_i) > 0, \ \forall j = 1, \cdots, N-1, i = 1, \cdots, N.$$

Then, by Taylor expansion, we have

$$\begin{aligned}
&\sum_{i=1}^{N} a_i \sigma(w'_j x_i) \\
=&\sum_{i=1}^{N} \left[a_i \sigma(w_j x_i) + a_i \sigma'(w_j x_i) \cdot x_i (w'_j - w_j) + a_i \sigma''(w_j x_i) \cdot \frac{1}{2} x_i^2 (w'_j - w_j)^2 + o(w'_j - w_j)^2 \right] \\
=&\sum_{i=1}^{N} a_i \sigma(w_j x_i) + (w'_j - w_j) \cdot \sum_{i=1}^{N} a_i x_i \sigma'(w_j x_i) + \frac{1}{2}(w'_j - w_j)^2 \cdot \sum_{i=1}^{N} a_i x_i^2 \sigma''(w_j x_i) + o(w'_j - w_j)^2 \\
=&0 + 0 + \frac{1}{2}(w'_j - w_j)^2 \cdot \sum_{i=1}^{N} x_i^2 \cdot a_i \sigma''(w_j x_i) + o(w'_j - w_j)^2 > 0.
\end{aligned} \tag{12}$$

After $\sigma(\cdot)$ is determined, choose arbitrary $v > 0$ and obtain $\hat{y} = v^\top \sigma(wx)$. Let $y = \hat{y} - a$, then $E(v, w) = \|y - \hat{y}\|^2 = \|a\|^2$.

Now consider any (v', w') such that $w' \in B(w, \delta)$. Denote $\hat{y}' = (v')^\top \sigma(w'x)$ and $z'_j = (\sigma(w'_j x_1), \cdots, \sigma(w'_j x_N))^\top, j = 1, \cdots, N-1$. If $w' = w$, then $v' \neq v$ and thus $\hat{y}' \neq \hat{y}$ due to linear independency of $z_1, \cdots, z_{N-1}$. Since $z_j = z'_j$ for all j, $\hat{y}' = \sum_{j=1}^{N-1} v'_j z'_j$ is also in hyperplane $\mathcal{H}$. On the other hand, note that a is the normal vector of $\mathcal{H}$, we have $a^\top \hat{y} = a^\top \hat{y}' = 0$. Therefore, $(y - \hat{y})^\top (\hat{y} - \hat{y}') = -a^\top (\hat{y} - \hat{y}') = 0$. Thus

$$E(v', w') = \|y - \hat{y}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y}' - \hat{y}\|^2 > E(v, w).$$

If $w' \neq w$, since $w' \in B(w, \delta)$, we have $a^\top z'_j = \sum_{i=1}^{N} a_i \sigma(w'_j x_i) > 0$ by 12. Since $v > 0$, we have $a^\top \hat{y}' = \sum_{i=1}^{N-1} v_j \cdot a^\top z'_j > 0$. Therefore, $(y - \hat{y})^\top (\hat{y} - \hat{y}') = -a^\top (\hat{y} - \hat{y}') = a^\top \hat{y}' - a^\top \hat{y} > 0$. Thus

$$\begin{aligned}
E(v', w') &= \|y - \hat{y}\|^2 = \|(y - \hat{y}) + (\hat{y} - \hat{y}')\|^2 \\
&= \|y - \hat{y}\|^2 + \|\hat{y}' - \hat{y}\|^2 + (y - \hat{y})^\top (\hat{y} - \hat{y}') > E(v, w).
\end{aligned}$$

Therefore we show that $E(v', w') > E(v, w)$ for all $(v', w') \in B((v, w), \delta) \backslash \{(v, w)\}$, which means that (v, w) is a bad strict local min of $E(\cdot)$. □

C Relationship Between Activation Function and Property PT

Given the strong connection between Property PT and the loss surface, a natural but crucial question is that what activation functions can guarantee Property PT. Intuitively, it seems that Property PT can be easily met as long as the activation function is non-linear. Unfortunately, this is not true. In Section 5, we mention that it is necessary for the output matrix Z to be full rank so that Property PT holds in (3). Below we provide a simple rank-1 example of rank-deficient output matrix with activation function $\sigma(z) = z^p$.

Example 2 (Rank-deficient output matrix) *Consider a 1-hidden-layer neural network with* 2 *data samples,* 1 *input dimension,* 2 *hidden-layer neurons, and monomial activation* $\sigma(z) = z^p$*,* $p > 0$*. We have* $W \in \mathbb{R}^{2\times 1}, X \in \mathbb{R}^{1\times 2}$*, and therefore* WX *is always of rank* 1*. Denote* $WX = \begin{bmatrix} a_1 & a_2 \\ ta_1 & ta_2 \end{bmatrix}$ *where* $t \in \mathbb{R}$*. Then,* $Z = \sigma(WX) = \begin{bmatrix} a_1^p & a_2^p \\ t^p a_1^p & t^p a_2^p \end{bmatrix}$ *is always rank-deficient. This implies that we cannot find any* W *such that* Z *is of full column rank. Property PT does not hold.*

In fact, for polynomial activation functions, rank-deficient output is common if the number of samples is sufficiently greater than the input dimension. The following is an example of rank-deficient output matrix with general polynomial activation functions.

Example 3 (Rank-deficient output matrix) *Consider a* 1*-hidden-layer fully-connected neural network with* N *data samples,* d_0 *input dimension,* d_1 *hidden-layer neurons, and a polynomial activation* $\sigma_1(z) = a_0 + a_1 z + \cdots + a_n z^p$*. Assume* $N \gg d_0$*,* $d_1 = N$*. We have* $W_1 \in \mathbb{R}^{N\times d_0}, X \in \mathbb{R}^{d_0 \times N}$*, and therefore* $W_1 X$ *is at most rank-*d_0*. Denote* $\circ$ *as the Hadamard product, then*

$$T_1 = \sigma(W_1 X) = a_0 \cdot \mathbf{1}^{N\times N} + a_1 \cdot (W_1 X) + \cdots + a_n \cdot (W_1 X) \circ \cdots \circ (W_1 X). \tag{13}$$

By the property of Hadamard product, we have

$$\textit{rank}(T_1) \leq 1 + d_0 + d_0^2 + \cdots d_0^p. \tag{14}$$

This implies that T_1 *is rank-deficient if* $N \gg d_0$*.*

D Proof of Theorem 1: Single-Hidden-Layer Case

In this section, we provide the proof of Theorem 1 for 1-hidden-layer case. Following the ideas in Section 5, our proof strategy consists of three steps.

Step 1: Prove the result for a specific class of activation functions.

We first introduce a special class of activation functions, specified in the following assumption. Note that to ease notations, we will omit the subscript of activation function σ_1 when considering a 1-hidden-layer network.

Assumption 3 (Special Activation Functions) *The activation function* σ *is analytic, and its first* n *derivatives at 0, i.e.,* $\sigma(0)$*,* $\sigma'(0)$*,* $\cdots$*,* $\sigma^{(n-1)}(0)$*, are all non-zero.*

Assumption 3 covers many commonly used activation functions such as sigmoid and softplus, but it does not cover ReLU since it requires smoothness (as mentioned before, ReLU is covered by using the approximation trick). Activation functions satisfying Assumption 3 have the following nice property:

Lemma 1 *If Assumption 3 holds, then for any* N *scalars* $z_1, z_2, \cdots, z_N \in \mathbb{R}$ *such that* $z_n \neq z_{n'}$ *for all* $n \neq n'$*, the following matrix*

$$A = \begin{pmatrix} \sigma(0) & \sigma(0) & \cdots & \sigma(0) \\ z_1 \sigma'(0) & z_2 \sigma'(0) & \cdots & z_N \sigma'(0) \\ \vdots & \vdots & & \vdots \\ z_1^{N-1} \sigma^{(N-1)}(0) & z_2^{N-1} \sigma^{(N-1)}(0) & \cdots & z_N^{N-1} \sigma^{N-1}(0) \end{pmatrix} \tag{15}$$

is non-singular.

Next, we borrow an important result of Mityagin (2015) which states that the zero set of an analytic function is either the whole domain or zero-measure. The result is formally stated as the following lemma.

Lemma 2 *For any $m \in \mathbb{Z}^+$, let $f : \mathbb{R}^m \to \mathbb{R}$ be a real analytic function on $\mathbb{R}^m$. If f is not identically zero, then its zero set $\Omega = \{\mathbf{z} \in \mathbb{R}^m \mid f(\mathbf{z}) = 0\}$ has zero measure.*

Based on Lemma 2, we have the following result.

Lemma 3 *Suppose that σ is an analytic function satisfying Assumption 3. Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$, let $\Omega = \{\mathbf{z} \in \mathbb{R}^N \mid \sigma(\mathbf{a}^\top \mathbf{z}) = \sigma(\mathbf{b}^\top \mathbf{z})\}$. If $\mathbf{a} \neq \mathbf{b}$, then Ω is of measure zero.*

Based on the above lemmas, we show that the output matrix of the hidden layer has full column rank for almost all W_1, specified by the following proposition.

Proposition 2 (General two-layer Case) *Consider a two-layer neural network with output $T_1 = \sigma(W_1 X)$, where $X \in \mathbb{R}^{d_0 \times N}$ and $W_1 \in \mathbb{R}^{d_1 \times d_0}$. Let $\Omega_1 = \{W_1 \in \mathbb{R}^{d_1 \times d_0} \mid rank(T_1) < \min\{d_1, N\}\}$. If Assumption A1, A2, and 3 hold, then Ω_1 is of measure zero.*

Here we only provide the proof of Proposition 2 for a simple 1-dimensional-input case. We note that it can be readily extended to the general case, detailed in Appendix L.

Proposition 3 (1-dimensional-input Two-layer Square Case) *Consider a two-layer neural network with one input neuron and d_1 neurons in the hidden layer. Given an activation function σ and one-dimensional input data $x \in \mathbb{R}^N$, let $\Omega = \{w_1 \in \mathbb{R}^d \mid det(\sigma(w_1 x^\top)) = 0\}$. If $d_1 = N$ and Assumption A1, A2, and 3 hold, then Ω is of measure zero.*

Proof: We prove this result by induction on N. The conclusion is obvious when $N = 1$.

We consider the case with $N > 1$. Since $f(w_1) \triangleq \det(w_1 x^\top)$ is an analytic function with respect to w_1, from Lemma 2 we know that Ω is either $\mathbb{R}^N$ or a zero-measure set. We now prove that Ω cannot be $\mathbb{R}^N$.

Assume on the contrary that $\Omega = \mathbb{R}^N$, i.e., $f(w_1) = 0, \forall w_1 \in \mathbb{R}^N$. Denote $w_{1,i}$ as the i-th entry of w_1, $1 \le i \le d_1$. For any $k \ge 0$, the k-th order partial derivative of $f(w_1)$ with respect to $w_{1,1}$ is given by

$$G_k(w_1) \triangleq \frac{\partial^k f(w)}{\partial w_{1,1}^k} = \det \begin{pmatrix} x_1^k \sigma^{(k)}(w_{1,1}x_1) & x_2^k \sigma^{(k)}(w_{1,1}x_2) & \cdots & x_N^k \sigma^{(k)}(w_{1,1}x_N) \\ \sigma(w_{1,2}x_1) & \sigma(w_{1,2}x_2) & \cdots & \sigma(w_{1,2}x_N) \\ \vdots & \vdots & & \vdots \\ \sigma(w_{1,N}x_1) & \sigma(w_{1,N}x_2) & \cdots & \sigma(w_{1,N}x_N) \end{pmatrix} \tag{16}$$

As $f(w_1) = 0, \forall w_1 \in \mathbb{R}^N$, we have $G_k(w_1) = 0, \forall w_1 \in \mathbb{R}^N$ and $\forall k \ge 0$.

Denote the n-th row of $G_k(w_1)$ by $\mathbf{u}_n = [\sigma(w_{1,n}x_1), \cdots, \sigma(w_{1,n}x_N)]^\top, n = 2, \cdots, N$. We show there exist some $w_{1,2}, \cdots, w_{1,N}$ such that $\mathbf{u}_2, \cdots, \mathbf{u}_N$ are linearly independent. We denote $\hat{\mathbf{u}}_n = [\sigma(w_{1,n}x_1), \cdots, \sigma(w_{1,n}x_{N-1})]^\top, n = 2, \cdots, N$, and $\hat{G} = [\hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_N] \in \mathbb{R}^{(N-1)\times(N-1)}$. According to the induction hypothesis, the set $\{(w_{1,2}, \cdots, w_{1,N}) \mid \det(\hat{G}) \neq 0\}$ is zero-measure in $\mathbb{R}^{N-1}$, implying that there exist some $w_{1,2}, \cdots, w_{1,N}$ such that $\hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_N$ are linearly independent. This also implies that $\mathbf{u}_2, \cdots, \mathbf{u}_N$ are linearly independent.

Now we have found some $w_{1,2}, \cdots, w_{1,N}$ such that $\mathbf{u}_2, \cdots, \mathbf{u}_N$ are linearly independent. Fix $w_{1,2}, \cdots, w_{1,N}$ and let $w_{1,1} = 0$. Denote the first row of G_k as $\mathbf{a}_k$. Since $\det(G_k) = 0$ for any $k \ge 0$, $\mathbf{a}_k$ must be a linear combination of $\mathbf{u}_2, \cdots, \mathbf{u}_N$ for any $k \ge 0$, so all $\mathbf{a}_k$'s lie in an $(N-1)$-dimension subspace of $\mathbb{R}^N$. However, according to Lemma 1, the N vectors $\mathbf{a}_0, \cdots, \mathbf{a}_{N-1}$ are linearly independent, which is a contradiction.

Therefore we have proved that Ω cannot be $\mathbb{R}^N$, so it must be a zero-measure set. □

Proposition 3 states that the output matrix of the hidden layer is generically full-column-rank. This property implies that the loss surface is a weakly global function. We have the following theorem.

Theorem 3 *Consider a fully connected neural network with* 1 *hidden layer, activation function* σ *and empirical loss function* $E(W) = l(Y, W_2 T_1)$. *Suppose that Assumption 1 and Assumption 3 hold. Then* $E(W)$ *is a weakly global function.*

Step 2: Show that the activation function in Assumption 2 can approximate any continuous function.

In order to extend Theorem 3 to all continuous activation functions without dealing them directly, we use a mathematical trick that approximates the continuous activation by a class of analytical functions.

Lemma 4 *For any continuous function* $f : \mathbb{R} \to \mathbb{R}$, *there exists a sequence of functions* $(f_k)_{k\in\mathbb{N}}$, *all satisfying Assumption 3, such that* f_k *converges to* f *uniformly.*

Lemma 4 means that the analytic functions satisfying Assumption 3 constitute a dense set (in the sense of uniformly convergence) of the space of continuous function. As we will show in the next step, this property allows us to approximate a neural network with any continuous activation function by a sequence of neural networks under Assumption 3.

Step 3: Show that the property of weakly global function is preserved under compact convergence.

Having built the relation between the neural network with analytic activation functions and the neural network with continuous activation function, the last step is to show that the weakly global property is preserved under this relation. The following result is a modification of a result in Josz et al. (2018).

Proposition 4 *Consider a sequence of functions* $(f_k)_{k\in\mathbb{N}}$ *and a function* f, *all from* $S \subset \mathbb{R}^m$ *to* $\mathbb{R}$. *If,*

$$f_k \to f \quad \textit{compactly} \tag{17}$$

and if f_k *are weakly global functions on* S, *then* f *is a weakly global function on* S.

Proposition 4 is slightly different from its original version in Josz et al. (2018): here we assume that f_k are weakly global functions instead of global functions. Nevertheless, we can still prove that f is weakly global by using similar techniques as in Josz et al. (2018).

Based on Proposition 4, to prove Theorem 1 it suffices to find a sequence of weakly global functions that compactly converges to the loss surface. Below we present the formal proof of Theorem 1 for 1-hidden-layer case.

Proof: We denote the considered network by $\mathcal{N}$. From Lemma 4, there exists a sequence of activation functions $(\sigma_k)_{k\in\mathbb{N}}$ that uniformly converges to σ. For each $k \in \mathbb{N}$, we construct a neural network, denoted by $\mathcal{N}_k$, by replacing the activation function in $\mathcal{N}$ with σ_k. For all $\mathcal{N}_k$, we assume the training dataset to be identical to that of $\mathcal{N}$. We also denote the output matrix of the hidden layer by $T_1^{(k)}$ and the empirical loss by

$$E_k(W) = l\left(Y, W_2 T_1^{(k)}\right). \tag{18}$$

From Theorem 3, E_k is a weakly global function with respect to W, $\forall k \in \mathbb{N}$. In what follows, we prove that the sequence of the empirical loss functions $(E_k)_{k\in\mathbb{N}}$ compactly converges to E.

Consider an arbitrary compact subset S in the space of W. For any $W \in S$, denote $\tilde{t}_{1,i,n}^{(k)}(W) = (T_1^{(k)})_{i,n}$ and $\tilde{t}_{1,i,n}(W) = (T_1)_{i,n}$ for any $k \in \mathbb{N}$, $1 \le i \le d_1$, and $1 \le n \le N$. That is, we rewrite the output of each neuron in the hidden layer as a function of W. Then, we have

$$\tilde{t}_{1,i,n}^{(k)}(W) = \sigma_k\left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n}\right) \tag{19}$$

$$\tilde{t}_{1,i,n}(W) = \sigma\left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n}\right). \tag{20}$$

Since σ_k uniformly converges to σ, $\tilde{t}_{1,i,n}^{(k)}$ also uniformly converges to $\tilde{t}_{1,i,n}$ on S for all $1 \leq i \leq d_1$, $1 \leq n \leq N$.

Now we consider the empirical loss equation 18. As every component of $T_1^{(k)}$ uniformly converges to the corresponding component of T_1, it can be shown that $W_{H+1}T_1^{(k)}$ also uniformly converges to $W_{H+1}T_1$ on S. By the continuousness of the loss function l, we have that E_k uniformly converges to E on S.

Finally, noting that S is an arbitrary compact subset in the space of W, the empirical loss E_k compactly converges to E on the space of W. Since $E_k(W)$ is a weakly global function for every $k \in \mathbb{N}$, by Proposition 4, $E(W)$ is also a weakly global function. We complete the proof. □

E Proof of Theorem 1: Extension to Deep Networks

In this section, we extend the proof of Theorem 1 from 1-hidden-layer case to deep neural networks. It turns out that the extension is simpler than expected, as we only need to follow the same idea as Section D and generalize some of the propositions and lemmas.

In the first place, we show that if the activation function of each hidden layer satisfies Assumption 3, the output matrix of the last hidden layer, i.e., T_H, is of full column rank for almost all W. Formally, we have the following extended version of Proposition 2.

Proposition 5 *Given a fully connected neural network with H hidden layers, activation function σ_h for each hidden layer, and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Let $\Omega = \{(W_1, \cdots, W_H) \mid rank(T_H) < \min\{d_H, N\}\}$. Suppose that Assumption 1 hold and σ_h satisfies Assumption 3 for all $1 \leq h \leq H$, then Ω is a zero-measure set.*

Proposition 5 immediately gives rise to the following theorem, which generalizes Theorem 3 to deep neural networks.

Theorem 4 *Given a fully connected neural network with H hidden layers., activation function σ_h for each hidden layer, and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Suppose that Assumption 1 hold and σ_h satisfies Assumption 3 for all $1 \leq h \leq H$, then $E(W)$ is a weakly global function.*

To completely prove Theorem 1, what remains is to find a sequence of weakly global functions that compactly converges to the loss surface. We have the following lemmas.

Lemma 5 *Consider two continuous functions $f : S \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$, where $S \subset \mathbb{R}^m$ is a compact set. Suppose that there exists two sequences of functions $(f_k)_{k\in\mathbb{N}}$ and $(g_k)_{k\in\mathbb{N}}$, such that f_k uniformly converges to f on S, and g_k compactly converges to g on $\mathbb{R}^n$. Then, $g_k \circ f_k$ converges to $g \circ f$ uniformly on S.*

Lemma 6 *Given a fully connected neural network with H hidden layers, activation function σ_h for each hidden layer, and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Suppose that Assumption 1 holds. Then, there exists a sequence of weakly global functions that compactly converges to $E(W)$.*

Combining Proposition 4 and Lemma 6, we complete the proof of Theorem 1.

F Proof of Theorem 2

Similar to that of Theorem 1, the proof of Theorem 2 consists of three steps.

Step 1: Prove the result for networks with specific activation functions.

From Assumption 2, there exists a wide layer of the network, i.e., the h_0-th hidden layer. We show that if the activation functions satisfy some specific conditions, the resulting empirical loss function is a weakly global function. We first specify a class of activation functions for the pyramid structure.

Assumption 4 *(Special Activation for Pyramid Structure) The activation function σ_h is continuous, strictly increasing or strictly decreasing, and its range is $\mathbb{R}$, i.e., $\sigma(\mathbb{R}) = \mathbb{R}$.*

Clearly, the activation functions satisfying Assumption 4 also satisfies Assumption B2. Further, we have the following theorem.

Theorem 5 *Given a fully connected neural network with H hidden layers., activation function σ_h for each hidden layer, and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Suppose that Assumption 1 and 2 hold, and*

1. *For any $1 \leq h \leq h_0$, σ_h satisfies Assumption 3;*
2. *For any $h_0 < h \leq H$, σ_h satisfies Assumption 4.*

Then, $E(W)$ is a weakly global function.

Theorem 5 identifies a special class of deep over-parameterized networks with pyramid structure, whose empirical loss function is weakly global.

Step 2: Show that for each hidden layer, the activation function in Theorem 5 can approximate the activation function in Theorem 2.

We use the same approximation trick as in the proof of Theorem 1. For the activation functions satisfying Assumption 3, we have Lemma 4. Regarding the activation functions satisfying Assumption 4, we have the following lemma.

Lemma 7 *For any continuous and non-increasing (or non-decreasing) function $f : \mathbb{R} \to \mathbb{R}$, there exists a sequence of functions $(f_k)_{k\in\mathbb{N}}$, all continuous and satisfying Assumption 4, such that f_k converges to f compactly.*

Step 3: Under compact convergence, show that the neural network considered in Theorem 2 is weakly global.

We denote the considered network by $\mathcal{N}$. From Lemma 4, for any $1 \leq h \leq h_0$, there exists a sequence of activation functions $(\sigma_{h,k})_{k\in\mathbb{N}}$, each satisfying Assumption 3, that uniformly converges to σ_h. From Lemma 7, for any $h_0 < h \leq H$, there exists a sequence of activation functions $(\sigma_{h,k})_{k\in\mathbb{N}}$, each satisfying Assumption 4, that compactly converges to σ_h. Since uniform convergence implies compact convergence, for all $1 \leq h \leq H$, $\sigma_{h,k}$ compactly converges to σ_h.

In the following, we show that the empirical loss of $\mathcal{N}$ can be approximated by a sequence of weakly global function, which is identical to the analysis in Appendix E

For each $k \in \mathbb{N}$, we construct a neural network, denoted by $\mathcal{N}_k$, by replacing the activation function of the h-th hidden layer with $\sigma_{h,k}$, for all $h = 1, 2, \cdots, H$. For all $\mathcal{N}_k$, we assume the training dataset to be identical to that of $\mathcal{N}$. We also denote the output matrix of the h-th hidden layer by $T_h^{(k)}$ and the empirical loss by

$$E_k(W) = l\left(Y, W_{H+1}T_H^{(k)}\right). \tag{21}$$

From Theorem 3, E_k is a weakly global function with respect to W, $\forall k \in \mathbb{N}$.

Consider an arbitrary compact subset S in the space of W. For any $W \in S$, define $\tilde{t}_{h,i,n}^{(k)}(W) = (T_h^{(k)})_{i,n}$ and $\tilde{t}_{h,i,n}(W) = (T_h)_{i,n}$ for any $k \in \mathbb{N}$, $1 \leq h \leq H$, $1 \leq i \leq d_h$, and $1 \leq n \leq N$. That is, we rewrite the output of each neuron in the hidden layers as a function of W. We prove by induction that every sequence $(\tilde{t}_{h,i,n}^{(k)})_{k\in\mathbb{N}}$ converges to $\tilde{t}_{h,i,n}$ uniformly on S.

For $h = 1$, we have

$$\tilde{t}^{(k)}_{h,i,n}(W) = \sigma_k \left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n} \right) \tag{22}$$

$$\tilde{t}_{h,i,n}(W) = \sigma \left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n} \right). \tag{23}$$

Since σ_k uniformly converges to σ, $\tilde{t}^{(k)}_{1,j,n}$ also uniformly converges to $\tilde{t}_{1,j,n}$ on S for all $1 \leq j \leq d_1$, $1 \leq n \leq N$.

For $h > 1$, assume that $\tilde{t}^{(k)}_{h-1,i,n}$ uniformly converges to $\tilde{t}_{h-1,i,n}$ on S for all $1 \leq i \leq d_{h-1}$, $1 \leq n \leq N$. For the h-th layer, we have

$$\begin{aligned} \tilde{t}^{(k)}_{h,i,n}(W) &= \sigma_{h,k} \left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} \left(T^{(k)}_{h-1} \right)_{j,n} \right) \\ &= \sigma_{h,k} \left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} \tilde{t}^{(k)}_{h-1,j,n}(W) \right) \end{aligned} \tag{24}$$

$$\begin{aligned} \tilde{t}_{h,i,n}(W) &= \sigma_h \left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} (T_{h-1})_{j,n} \right) \\ &= \sigma_h \left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} \tilde{t}_{h-1,j,n}(W) \right). \end{aligned} \tag{25}$$

By the induction hypothesis, it is easy to show that $\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} \tilde{t}^{(k)}_{h-1,j,n}(W)$ uniformly converges to $\sum_{j=1}^{d_{h-1}} (W_h)_{i,j} \tilde{t}_{h-1,j,n}(W)$ on S. Note that $(\sigma_{h,k})_{k \in \mathbb{N}}$ converges to σ_h uniformly, and uniform convergence implies compact convergence. It directly follows from Lemma 5 that $\tilde{t}^{(k)}_{h,i,n}(W)$ converges to $\tilde{t}_{h,i,n}(W)$.

Therefore, we conclude that $\tilde{t}^{(k)}_{h,i,n}$ converges to $\tilde{t}_{h,i,n}$ uniformly on S for every $1 \leq h \leq H$, $1 \leq i \leq d_h$, and $1 \leq n \leq N$.

Now we consider the empirical loss

$$E_k(W) = l \left(Y, W_{H+1} T^{(k)}_H \right) \tag{26}$$

$$E(W) = l \left(Y, W_{H+1} T_H \right). \tag{27}$$

As every component of $T^{(k)}_H$ converges uniformly to the corresponding component of T_H on S, it can be shown that $W_{H+1} T^{(k)}_H$ converges uniformly to $W_{H+1} T_H$ on S. By Lemma 5, where we set both g_k and g to the loss function l, we have that E_k uniformly converges to E on S. Noting that S is an arbitrary compact subset in the space of W, the empirical loss E_k converges to E compactly on the space of W. Since $E_k(W)$ is a weakly global function for every $k \in \mathbb{N}$, by Proposition 4, $E(W)$ is also a weakly global function. We complete the proof.

G Proof of Theorem 3

Note that the proof of Theorem 3 is essentially the same with that of Theorem 4 if we set $H = 1$. We omit the proof of Theorem 3 here and refer the readers to Appendix H.

H Proof of Theorem 4

The proof consists of three steps. First, we show that for any W, we can perturb it to a point W' whose corresponding T_H is full rank. Second, we prove that starting from the perturbed point W', there exists a strictly decreasing path to the global infimum. Finally, we combine the two previous steps to show that the loss function is weakly global.

Proof: We first prove that from any initial weight $W^o = (W_1^o, \cdots, W_{H+1}^o)$, there exists a strictly decreasing path towards $\inf_W E(W)$ [4] after an arbitrarily small perturbation. According to Proposition 5, all W's that entail a non-full-rank T_H only constitute a zero-measure set. Therefore, for any initial weight W^o and an arbitrarily small $\delta > 0$, there exists $W^p = (W_1^p, W_2^p, \cdots, W_{H+1}^p) \in B(W^o, \delta)$ such that the corresponding T_H^p is full rank. Since $d_H \geq N$, we have $\text{rank}(T_H^p) = N$.

In what follows, we show that starting from W^p, there exists a continuous path along which the empirical loss $E(W)$ strictly decreases to $\inf_W E(W)$. Denote $W_{1:H} = (W_1, \cdots, W_H)$, i.e., the weights in the first H layers. By the feed-forward operation equation 8, T_H is a function of $W_{1:H}$. Thus, $E(W)$ can be rewritten as $l(Y, W_{H+1}T_H(W_{1:H}))$. Since $l(Y, \hat{Y})$ is convex to $\hat{Y}$, for any W_{H+1}, W'_{H+1}, and $\lambda \in [0, 1]$, we have

$$\begin{aligned} E(W) =& l\left(Y, \left(\lambda W_{H+1} + (1-\lambda)W'_{H+1}\right) T_H(W_{1:H})\right) \\ =& l\left(Y, \lambda W_{H+1}T_H(W_{1:H}) + (1-\lambda)W'_{H+1}T_H(W_{1:H})\right) \\ \leq& \lambda l\left(Y, W_{H+1}T_H(W_{1:H})\right) + (1-\lambda) l\left(Y, W'_{H+1}T_H(W_{1:H})\right) \end{aligned} \tag{28}$$

Thus, with the weights to the first H hidden layers fixed, $E(W)$ is convex with respect to W_{H+1}. This implies that starting from W^p, we can find a strictly decreasing path towards $\inf_{W_{H+1}} l(Y, W_{H+1}T_H(W_{1:H}^p))$ by fixing $W_{1:H} = W_{1:H}^p$ and moving along W_{H+1}. Moreover, since $T_H(W_{1:H}^p) \in \mathbb{R}^{d_H \times N}$ is full column rank, for any $\hat{Y} \in \mathbb{R}^{d_{H+1} \times N}$, there exists W_{H+1} such that $W_{H+1}T_H(W_{1:H}^p) = \hat{Y}$, yielding

$$\inf_{W_{H+1}} l(Y, W_{H+1}T_H(W_{1:H}^p)) = \inf_{\hat{Y}} l(Y, \hat{Y}) = \inf_W E(W). \tag{29}$$

Therefore, the constructed path is strictly decreasing towards $\inf_W E(W)$.

Now we prove by contraposition that $E(W)$ is a weakly global function. Assume in contrast that there exists a bad strict local minimum of $E(W)$ in the sense of sets, denoted by $\mathcal{W}$. Note by Definition 2, $\mathcal{W}$ is a compact set. Let $\mathcal{W}_\delta = \{W' \mid \inf_{W\in\mathcal{W}} \|W' - W\|_2 \leq \delta\}$, then there exists $\delta > 0$ such that for all $W \in \mathcal{W}$ and $W' \in \mathcal{W}_\delta \setminus \mathcal{W}$, $E(W) < E(W')$. Denote $\partial\mathcal{W}_\delta$ as the boundary of $\mathcal{W}_\delta$. Note that both $\mathcal{W}_\delta$ and $\partial\mathcal{W}_\delta$ are closed, there exists $W^* \in \partial\mathcal{W}_\delta$ such that $E(W^*) = \inf_{W'\in\partial\mathcal{W}_\delta} E(W')$. Moreover, $E(W^*) = \sup_{W\in\mathcal{W}} E(W) + \varepsilon$ for some $\varepsilon > 0$.

Consider an arbitrary point $W^o \in \mathcal{W}$. Since $E(W)$ is a continuous function, there exists $\delta > \delta_0 > 0$ such that for any $W' \in B(W^o, \delta_0)$, $|E(W') - E(W^o)| < \varepsilon/2$. According to the first conclusion, we can find $W^p \in B(W^o, \delta_0)$ such that there exists a strictly decreasing path from W^p to $\inf_W E(W)$. Since $\mathcal{W}$ is a bad local minimum, $\inf_{W\in\mathcal{W}_\delta} E(W) > \inf_W E(W)$. Therefore, the above strictly decreasing path starting from W^p must pass through the boundary $\partial\mathcal{W}_\delta$. However, $E(W^p) < E(W^o) + \varepsilon/2 < \sup_{W\in\mathcal{W}} E(W) + \varepsilon = E(W^*) = \inf_{W'\in\partial\mathcal{W}_\delta} E(W')$. This implies that the considered path can never be strictly decreasing, leading to a contradiction. Therefore, we conclude that there is no bad strict local minima in the sense of sets, and therefore $E(W)$ is a weakly global function. □

I Proof of Theorem 5

First, we show that for any initial weight W^o, we can perturb it to a point W^p, and starting from the perturbed point W^p, there exists a strictly decreasing path towards the global infimum.

Denote Θ as the space of W and let $W_{1:h} = (W_1, W_2, \cdots, W_h)$ denotes the weights for the first h hidden layers. We consider the following set:

$$\Omega = \{W | \text{rank}(T_{h_0}) = N, \quad \text{rank}(W_h) = d_h, \quad \forall h_0 < h \leq H+1\}. \tag{30}$$

[4] To be more specific, if the global infimum is achievable (i.e., global minimum), then the decreasing path ends exactly at some point with empirical loss $\inf_W E(W)$. Otherwise, the empirical loss along the path converges to $\inf_W E(W)$

Since σ_h satisfies Assumption 3 for any $1 \leq h \leq h_0$, according to Lemma 5, all $W_{1:h_0}$'s that entail a non-full-rank T_{h_0} only constitute a zero-measure set. That is, all $W_{1:h_0}$'s such that T_{h_0} is $\mathrm{rank}(N)$ constitute a dense set. Further, note that full-rank matrices are dense. For all $h_0 < h \leq H+1$, since $d_h \leq d_{h-1}$ from Assumption B1, $\{W_h | \mathrm{rank}(W_h) = d_h\}$ is dense in $\mathbb{R}^{d_h \times d_{h-1}}$. From the above analysis, we conclude that Ω is dense in Θ. Therefore, for any initial weight W^o and an arbitrarily small $\delta > 0$, there exists $W^p = (W_1^p, W_2^p, \cdots, W_{H+1}^p) \in B(W^o, \delta)$ such that $W^p \in \Omega$.

In what follows, we show that if $E(W^p) > \inf_W (E(W))$, there exists a continuous path starting from W^p, and the empirical loss along which strictly decreases to $\inf_W E(W)$.

Case 1: There exists $\hat{Y}' \in \mathbb{R}^{d_{H+1} \times N}$ such that $l(Y, \hat{Y}') = \inf_{\hat{Y}} l(Y, \hat{Y})$.

Denote T_h^p as the output of the h-th hidden layer at weight W^p, and $\hat{Y}^p = W_{H+1}^p T_H^p$ as the network output at weight W^p. Since the loss function $l(Y, \hat{Y})$ is convex with respect to $\hat{Y}$, it is also continuous with respect to $\hat{Y}$. Further, we have

$$l(Y, \hat{Y}^p) = E(W^p) > \inf_W E(W) \geq \inf_{\hat{Y}} l(Y, \hat{Y}), \tag{31}$$

and hence $\hat{Y}^p$ is not a global minimum of $l(Y, \hat{Y}^p)$. Then, there exists a continuous path $\hat{Y}(\lambda) : [0,1] \to \mathbb{R}$ such that $\hat{Y}(0) = \hat{Y}^p$, $l(Y, \hat{Y}(1)) = \inf_{\hat{Y}} l(Y, \hat{Y})$, and $l(Y, Y(\lambda))$ is strictly decreasing with respect to λ, i.e.,

$$l(Y, Y(\lambda_1)) > l(Y, Y(\lambda_2)), \quad \forall \lambda_1 < \lambda_2, \quad \lambda_1, \lambda_2 \in [0,1]. \tag{32}$$

By Assumption 4, for each $h_0 < h \leq H$, σ_h has a continuous inverse $\sigma_h^{-1} : \mathbb{R} \to \mathbb{R}$. Also, since W_h^p is of full row rank for each $h_0 < h \leq H$, it has a right inverse $(W_h^p)^\dagger$ such that $W_h^p (W_h^p)^\dagger = \mathbf{I}$. Further, as $T_{h_0}^p$ has full column rank, it has a left inverse $\left(T_{h_0}^p\right)^\dagger$ such that $\left(T_{h_0}^p\right)^\dagger \left(T_{h_0}^p\right) = \mathbf{I}$.

We construct $W(\lambda) : [0,1] \to \Theta$ as follows

$$W(\lambda) = (W_1^p, W_2^p, \cdots, W_{h_0}^p, W_{h_0+1}(\lambda), W_{h_0+2}^p, \cdots, W_{H+1}^p), \quad \lambda \in [0,1] \tag{33}$$

where $W_h(\lambda)$ is defined recursively as follows

$$W_{h_0+1}(\lambda) = \left[\sigma_{h_0+1}^{-1}\left(T_{h_0+1}(\lambda)\right) - \sigma_{h_0+1}^{-1}\left(T_{h_0+1}^p\right)\right]\left(T_{h_0}^p\right)^\dagger + W_{h_0+1}^p \tag{34a}$$

$$T_h(\lambda) = \left(W_{h+1}^p\right)^\dagger \left(\sigma_{h+1}^{-1}\left(T_{h+1}(\lambda)\right) - \sigma_{h+1}^{-1}\left(T_{h+1}^p\right)\right) + T_h^p, \qquad h = h_0+1, h_0+2, \cdots, H-1 \tag{34b}$$

$$T_H(\lambda) = \left(W_{H+1}^p\right)^\dagger \left(\hat{Y}(\lambda) - \hat{Y}^p\right) + T_H^p. \tag{34c}$$

For the constructed $W(\lambda)$, we verify the following three facts.

(1) $W(\lambda)$ is a continuous path.

In fact, since $\hat{Y}(\lambda)$ is continuous, σ_h^{-1} is a continuous function for all $h_0 < h \leq H$, therefore each $T_h(\lambda)$ is continuous with respect λ for all $h_0 < h \leq H$. Thus, $W(\lambda)$ is also continuous with respect to λ.

(2) $W(0) = W^p$, and $W(1)$ is a global minimum of $E(W)$.

Note that $\hat{Y}(0) = \hat{Y}^p$. From equation 34, we have

$$T_H(0) = \left(W_{H+1}^p\right)^\dagger \mathbf{0} + T_H^p = T_H^p \tag{35a}$$

$$T_h(0) = \left(W_{h+1}^p\right)^\dagger \mathbf{0} + T_h^p = T_h^p, \quad h = h_0+1, h_0+2, \cdots, H \tag{35b}$$

$$W_{h_0+1}(0) = \mathbf{0}\left(T_{h_0}^p\right)^\dagger + W_{h_0+1}^p = W_{h_0+1}^p. \tag{35c}$$

Notice that $W(\lambda)$ is identical to W^p except the weights to the (h_0+1)-th layer. Thus, $W(0) = W^p$. Now consider the output of each hidden layer at weight $W(\lambda)$, denoted by

T^λ_h. For $W(\lambda)$ and W^p, the weights to the first h_0 hidden layers are the same, and hence $T^\lambda_{h_0} = T^p_{h_0}$. From equation 34, we have

$$\begin{aligned}
T^\lambda_{h_0+1} &= \sigma_{h_0+1}\left(W_{h_0+1}(\lambda)T^\lambda_{h_0}\right) = \sigma_{h_0+1}\left(W_{h_0+1}(\lambda)T^p_{h_0}\right) && (36a)\\
&= \sigma_{h_0+1}\left(\sigma^{-1}_{h_0+1}\left(T_{h_0+1}(\lambda)\right) - \sigma^{-1}_{h_0+1}\left(T^p_{h_0+1}\right) + W^p_{h_0+1}T^p_{h_0}\right) && (36b)\\
&= \sigma_{h_0+1}\left(\sigma^{-1}_{h_0+1}\left(T_{h_0+1}(\lambda)\right) - \sigma^{-1}_{h_0+1}\left(\sigma_{h_0+1}\left(W^p_{h_0+1}T^p_{h_0}\right)\right) + W^p_{h_0+1}T^p_{h_0}\right) && (36c)\\
&= \sigma_{h_0+1}\left(\sigma^{-1}_{h_0+1}\left(T_{h_0+1}(\lambda)\right)\right) && (36d)\\
&= T_{h_0+1}(\lambda) && (36e)
\end{aligned}$$

For $h_0 + 1 < h \leq H$, if $T^\lambda_{h-1} = T_{h-1}(\lambda)$, we have

$$\begin{aligned}
T^\lambda_h &= \sigma_h\left(W^p_h T^\lambda_{h-1}\right) = \sigma_h\left(W^p_h T_{h-1}(\lambda)\right) && (37a)\\
&= \sigma_h\left(W^p_h\left(W^p_h\right)^\dagger\left(\sigma_h^{-1}\left(T_h(\lambda)\right) - \sigma_h^{-1}\left(T^p_h\right)\right) + W^p_h T^p_{h-1}\right) && (37b)\\
&= \sigma_h\left(\sigma_h^{-1}\left(T_h(\lambda)\right) - \sigma_h^{-1}\left(\sigma_h\left(W^p_h T^p_{h-1}\right)\right) + W^p_h T^p_{h-1}\right) && (37c)\\
&= \sigma_h\left(\sigma_h^{-1}\left(T_h(\lambda)\right)\right) && (37d)\\
&= T_h(\lambda). && (37e)
\end{aligned}$$

And similarly, for the network output at $W(\lambda)$, denoted by $\hat{Y}^\lambda$ we have

$$\begin{aligned}
\hat{Y}^\lambda &= W^p_{H+1}T^\lambda_H = W^p_{H+1}T_H(\lambda) && (38a)\\
&= W^p_{H+1}\left(W^p_{H+1}\right)^\dagger\left(\hat{Y}(\lambda) - \hat{Y}^p\right) + W^p_{H+1}T^p_H && (38b)\\
&= \hat{Y}(\lambda) - \hat{Y}^p + W^p_{H+1}T^p_H && (38c)\\
&= \hat{Y}(\lambda). && (38d)
\end{aligned}$$

Then, the empirical loss

$$E(W(1)) = l(Y, \hat{Y}^1) = \inf_{\hat{Y}} l(Y, \hat{Y}) \leq \inf_W E(W). \tag{39}$$

Therefore we must have $E(W(1)) = \inf_W E(W)$. That is, $E(1)$ is a global minimum of $E(W)$.

(3) $E(W(\lambda))$ is strictly decreasing with respect to λ.
From (2), we have

$$E(W(\lambda)) = l(Y, \hat{Y}^\lambda) = l(Y, \hat{Y}(\lambda)). \tag{40}$$

Then for any $\lambda_1, \lambda_2 \in [0, 1]$ and $\lambda_1 < \lambda_2$, we have

$$E(W(\lambda_1)) = l(Y, \hat{Y}(\lambda_1)) > l(Y, \hat{Y}(\lambda_2)) = E(W(\lambda_2)). \tag{41}$$

$E(W(\lambda))$ is strictly decreasing with respect to λ.

We conclude that $W(\lambda)$ starts at from W^p and is a strictly decreasing path towards the global infimum.

Case 2: There does not exist $\hat{Y}' \in \mathbb{R}^{d_{H+1}\times N}$ such that $l(Y, \hat{Y}') = \inf_{\hat{Y}} l(Y, \hat{Y})$.

Similar to Case 1, there exists a continuous path $\hat{Y}(\lambda) : [0, 1) \to \mathbb{R}$ such that $\hat{Y}(0) = \hat{Y}^p$, $\lim_{\lambda\to 1} l(Y, \hat{Y}(\lambda)) = \inf_{\hat{Y}} l(Y, \hat{Y})$, and $l(Y, Y(\lambda))$ is strictly decreasing with respect to λ. We can then construct a continuous path $W(\lambda) : [0, 1) \to \Theta$, such that $E(W(\lambda))$ is strictly decreasing, and $\lim_{\lambda\to 1} E(W(\lambda))$. Since the construction and analysis are identical to that in Case 1, we omit the details here.

Then we prove by contraposition that $E(W)$ is a weakly global function. Assume in contrast that there exists a bad strict local minimum of $E(W)$ in the sense of sets, denoted by $\mathcal{W}$. Note by Definition 2, $\mathcal{W}$ is a compact set. Let $\mathcal{W}_\delta = \{W' \mid \inf_{W\in\mathcal{W}} \|W' - W\|_2 \leq \delta\}$, then there exists $\delta > 0$ such that for all $W \in \mathcal{W}$ and $W' \in \mathcal{W}_\delta \setminus \mathcal{W}$, $E(W) < E(W')$. Denote $\partial\mathcal{W}_\delta$ as the boundary of $\mathcal{W}_\delta$. Note that both $\mathcal{W}_\delta$ and $\partial\mathcal{W}_\delta$ are closed, there exists $W^* \in \partial\mathcal{W}_\delta$ such that $E(W^*) = \inf_{W' \in \partial\mathcal{W}_\delta} E(W')$. Moreover, $E(W^*) = \sup_{W \in \mathcal{W}} E(W) + \varepsilon$ for some $\varepsilon > 0$.

Consider an arbitrary point $W^o \in \mathcal{W}$. Since $E(W)$ is a continuous function, there exists $\delta > \delta_0 > 0$ such that for any $W' \in B(W^o, \delta_0)$, $|E(W') - E(W^o)| < \varepsilon/2$. According to the first conclusion, we can find $W^p \in B(W^o, \delta_0)$ such that there exists a strictly decreasing path from W^p to $\inf_W E(W)$. Since $\mathcal{W}$ is a bad local minimum, $\inf_{W \in \mathcal{W}_\delta} E(W) > \inf_W E(W)$. Therefore, the above strictly decreasing path starting from W^p must pass through the boundary $\partial\mathcal{W}_\delta$. However, $E(W^p) < E(W^o) + \varepsilon/2 < \sup_{W \in \mathcal{W}} E(W) + \varepsilon = E(W^*) = \inf_{W' \in \partial\mathcal{W}_\delta} E(W')$. This implies that the considered path can never be strictly decreasing, leading to a contradiction. Therefore, we conclude that there is no bad strict local minima in the sense of sets, and therefore $E(W)$ is a weakly global function.

J Proof of Lemma 1

Notice that A is a Vandermonde matrix multiplied by $\sigma^{n-1}(0)$ to the n-th row, $n = 1, 2, \cdots, N$. Since $\sigma^{n-1}(0) \neq 0$ according to Assumption 3, A is a non-singular matrix.

K Proof of Lemma 3

Assume that Ω is not of zero measure. Since $\sigma(\mathbf{a}^\top \mathbf{z}) - \sigma(\mathbf{b}^\top \mathbf{z})$ is an analytic function of $\mathbf{z}$, Ω must be $\mathbb{R}^N$ according to Lemma 2. That is, $\sigma(\mathbf{a}^\top \mathbf{z}) = \sigma(\mathbf{b}^\top \mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^N$. We will show that this leads to a contradiction.

If $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$, assume $\mathbf{a} = \mathbf{0}$ without loss of generality. Since $\mathbf{b} \neq \mathbf{0}$, there exists some $\mathbf{z}_0$ such that $\mathbf{b}^\top \mathbf{z}_0 = 1$. Therefore, for any $\lambda \in \mathbb{R}$, we have

$$\sigma(\lambda) = \sigma(\mathbf{b}^\top(\lambda \mathbf{z}_0)) = \sigma(\mathbf{a}^\top(\lambda \mathbf{z}_0)) = \sigma(0). \tag{42}$$

Thus, σ is a constant function and therefore $\sigma' \equiv 0$, a contradiction to Assumption 3.

If $\mathbf{a}, \mathbf{b} \neq \mathbf{0}$, then the set of all $\mathbf{z}$ satisfying $\mathbf{a}^\top \mathbf{z} = 0$ or $\mathbf{b}^\top \mathbf{z} = 0$ is of measure zero. Since $\mathbf{a} \neq \mathbf{b}$, the set of all $\mathbf{z}$ satisfying $\mathbf{a}^\top \mathbf{z} = \mathbf{b}^\top \mathbf{z}$ is also of zero measure. Therefore, there exists some $\mathbf{z}_0$ such that $\mathbf{a}^\top \mathbf{z}_0 \neq \mathbf{0}$, $\mathbf{b}^\top \mathbf{z}_0 \neq \mathbf{0}$, and $\mathbf{a}^\top \mathbf{z}_0 \neq \mathbf{b}^\top \mathbf{z}_0$. Denote $a_0 = \mathbf{a}^\top \mathbf{z}_0$ and $b_0 = \mathbf{b}^\top \mathbf{z}_0$. Since $\Omega = \mathbb{R}^N$, we conclude that for any $\lambda \neq 0$, $\sigma(\lambda a_0) = \sigma(\mathbf{a}^\top \mathbf{z}_0) = \sigma(\mathbf{b}^\top \mathbf{z}_0) = \sigma(\lambda b_0)$. Note that $a_0 b_0 \neq 0$, $a_0 \neq b_0$, and $\sigma(\lambda a_0) = \sigma(\lambda b_0)$ for all $\lambda \neq 0$. Letting $\lambda \to 0$, we have

$$0 = \lim_{\lambda \to 0} \frac{\sigma(\lambda a_0) - \sigma(\lambda b_0)}{\lambda a_0 - \lambda b_0} = \sigma'(0) \tag{43}$$

where the second equality holds since σ is analytic. This also contradicts Assumption 3.

We conclude that for any $\mathbf{a} \neq \mathbf{b}$, Ω cannot be $\mathbb{R}^N$, and therefore must be of zero measure.

L Proof of Proposition 2

Let $\boldsymbol{w}_i^\top$ be the i-th row of W_1, $i = 1, 2, \cdots, d_1$. According to Assumption A1, we can assume without loss of generality that the first row of X has distinct entries, i.e., $x_1^{(1)}, x_1^{(2)} \cdots, x_1^{(N)}$ are distinct from each other.

Notice that $T_1 \in \mathbb{R}^{d_1 \times N}$. If $d_1 < N$, we select the first d_1 columns of T_1 and obtain a sub-matrix $\hat{T}_1 \in \mathbb{R}^{d_1 \times d_1}$. Let $\Omega_1' = \{W_1 \in \mathbb{R}^{d_1 \times d_0} \mid \text{rank}(\hat{T}_1) < d_1\}$. We can show that Ω_1' is a zero-measure set by applying a similar analysis to $\hat{T}_1$ as in the proof of Proposition 3. The only change to make is that here we calculate the partial derivatives with respect to $w_{1,1,1}$. Notice that for any $W_1 \in \Omega_1$, any $d_1 \times d_1$ sub-matrix of T_1 should be singular. Since Ω_1 is a subset of Ω_1', it should also be zero measure.

If $d_1 \geq N$, we select the first N rows of T_1 and obtain a sub-matrix $\hat{T}_1 \in \mathbb{R}^{N \times N}$. Similarly, let $\hat{W}_1 \in \mathbb{R}^{N \times d_0}$ be the first N rows of W_1. Let $\Omega_1' = \{\hat{W}_1 \in \mathbb{R}^{N \times d_0} \mid \text{rank}(\hat{T}_1) < N\}$. Following a similar analysis as in the case $d_1 < N$, Ω_1' is of measure zero in $\mathbb{R}^{N \times d_0}$. Note that for any $W_1 \in \Omega_1$, its submatrix consisting of the first N rows is in Ω_1'. Thus, Ω_1 is of measure zero in $\mathbb{R}^{d_1 \times d_0}$.

M PROOF OF LEMMA 4

The proof of Lemma 4 consists of two parts. In the first part we show that the function class specified by Assumption 3 is dense in the space of analytic functions. In the second part, following the fact that the space of analytic functions is a dense set in the space of continuous function, we prove that the function class specified by Assumption 3 is also dense in the space of continuous functions.

To prove the first part, we consider an arbitrary analytic function $g : \mathbb{R} \to \mathbb{R}$, and then construct a sequence of functions $(f_k)_{k\in\mathbb{N}}$, all satisfying Assumption 3, such that f_k converges to g uniformly.

Let

$$f_k(x) = g(x) + \frac{1}{s(k+1)}\left(\sin x + \cos x\right). \tag{44}$$

Clearly, f_k is analytic for any $k \in \mathbb{N}$ and $s \neq 0$. Further, we have

$$f_k^{(n)}(0) = g^{(n)}(0) + \frac{1}{s(k+1)}(-1)^n. \tag{45}$$

We next show that there exists $s \neq 0$ such that all f_k's satisfy Assumption 3. Consider the following two cases: (1) $g^{(n)}(0) = 0$ for all $0 \le n \le N$; and (2) $g^{(n)}(0) \neq 0$ for some $0 \le n \le N$.

Case 1: For any $s \neq 0$, since $g^{(n)}(0) = 0$, we have

$$f_k^{(n)}(0) = \frac{1}{s(k+1)}(-1)^n \neq 0 \tag{46}$$

for all $n = 0, 1, \cdots, N$. Thus, all f_k's satisfy Assumption 3.

Case 2: Since $g^{(n)}(0) \neq 0$ for at least one $n \in \{0, 1, \cdots, N\}$, we can define

$$\delta_{\min} = \min\left\{|g^{(n)}(0)| \mid 0 \le n \le N, \quad g^{(n)}(0) \neq 0\right\} \tag{47}$$

i.e., the minimum non-zero absolute value of $g^{(n)}(0)$, $n = 0, 1, \cdots, N$. Clearly, $\delta_{\min} > 0$. Letting $s = 2/\delta_{\min}$, we have

$$f_k^{(n)}(0) = g^{(n)}(0) + \frac{\delta_{\min}}{2(k+1)}(-1)^n \tag{48}$$

For $g^{(n)}(0) = 0$, we have

$$f_k^{(n)}(0) = \frac{\delta_{\min}}{2(k+1)}(-1)^n \neq 0. \tag{49}$$

For $g^{(n)}(0) \neq 0$, we have

$$\begin{aligned}
\left|f_k^{(n)}(0)\right| &= \left|g^{(n)}(0) + \frac{\delta_{\min}}{2(k+1)}(-1)^n\right| && \text{(50a)}\\
&\ge \left|g^{(n)}(0)\right| - \left|\frac{\delta_{\min}}{2(k+1)}(-1)^n\right| && \text{(50b)}\\
&\ge \delta_{\min} - \frac{\delta_{\min}}{2(k+1)} && \text{(50c)}\\
&= \frac{\delta_{\min}(2k+1)}{2(k+1)} && \text{(50d)}\\
&> 0 && \text{(50e)}
\end{aligned}$$

where equation 50c holds by the definition of $\delta_{\min}$ in equation 47. Therefore, all f_k's satisfy Assumption 3.

We now prove the uniform convergence of f_k for any $s \neq 0$. Specifically, for any $\epsilon > 0$, we have

$$\begin{aligned}
|f_k(x) - g(x)| &= \frac{1}{s(k+1)}\left|\sin x + \cos x\right| && \text{(51a)}\\
&\le \frac{\sqrt{2}}{s(k+1)} && \text{(51b)}\\
&< \epsilon && \text{(51c)}
\end{aligned}$$

for all $k > \sqrt{2}/(\epsilon s) - 1$ and $x \in \mathbb{R}$. Therefore, f_k converges uniformly to g.

We conclude that function class specified by Assumption 3 is dense in the space of analytic functions.

Now we come to the second part. By the Carleman Approximation Theorem Kaplan (1955), the space of analytic functions is dense in the space of continuous functions. That is, for any continuous function $f : \mathbb{R} \to \mathbb{R}$, there exists a sequence of analytic functions $(g_k)_{k\in\mathbb{N}}$ such that g_k converges to f uniformly. Following the idea of Cantor's diagonal argument, we can construct a sequence of functions satisfying Assumption 3, which also converges to f.

Note that each g_k is an analytic function. By the analysis in the first part, for each $k \in \mathbb{N}$, we can construct a sequence of functions $(f_j^{(k)})_{j\in\mathbb{N}}$, all satisfying Assumption 3, such that $f_j^{(k)}$ converges to g_k uniformly. Further, we can require that for each $k \in \mathbb{N}$,

$$\left|f_j^{(k)}(x) - g_k(x)\right| \leq \frac{1}{k+1}, \quad \forall x \in \mathbb{R}, \quad j \in \mathbb{N}. \tag{52}$$

In fact, if equation 52 is not satisfied, we can always delete a finite number of functions at the beginning of the sequence, so as to produce a new sequence that meets equation 52. Now considered the sequence $(f_k^{(k)})_{k\in\mathbb{N}}$. Since g_k converges to f uniformly, for any $\epsilon > 0$, there exists a $K_1 \in \mathbb{N}$ such that $|g_k(x) - f(x)| \leq \epsilon/2$ for any $k \geq K_1$ and $x \in \mathbb{R}$. Then, for any $k > \max\{K_1, 2/\epsilon - 1\}$, we have

$$\begin{aligned}\left|f_k^{(k)}(x) - f(x)\right| \leq& \left|f_k^{(k)}(x) - g_k(x)\right| + |g_k(x) - f(x)| \\ \leq& \frac{1}{k+1} + \epsilon/2 \qquad (53a) \\ \leq& \epsilon. \qquad (53b)\end{aligned}$$

Therefore, $f_k^{(k)}$ converges to f uniformly. Noting that f is an arbitrary continuous function from $\mathbb{R}$ to $\mathbb{R}$, we complete the proof.

N Proof of Proposition 4

Consider a sequence of weakly global functions f_k that converges compactly towards f. Since $S \subset \mathbb{R}^m$ and $\mathbb{R}^m$ is a compactly generated space, it follows that f is continuous. We proceed to prove that f is a weakly global function by contradiction. Suppose $S_M \subset S$ is a strict local minimum that is not global minimum. There exists $\epsilon > 0$ such that the uniform neighborhood $V := \{y \in S \mid \exists x \in S_M : \|x - y\|_2 \leq \epsilon\}$ satisfies $f(x) < f(y)$ for all $x \in S_M$ and for all $y \in V \setminus S_M$. Since f is continuous on the compact set S_M, it attains a minimal value on it, say $\inf_{S_M} f := \alpha + \inf_S f$ where $\alpha > 0$ since S_M is not a global minimum. Consider a compact set $V \subset K \subset S$ such that $\inf_K f \leq \alpha/2 + \inf_S f$. Since f is continuous on the compact set ∂V, it attains a minimal value on it, say $\inf_{\partial V} f := \beta + \inf_{S_M} f$ where $\beta > 0$ by strict optimality. Let $\gamma := \min\{\alpha/2, \beta\}$. For a sufficiently large value of k, compact convergence implies that $|f_k(y) - f(y)| \leq \gamma/3$ for all $y \in K$. Since the function f_k is compact on V, it attains a minimum, say $z' \in V$. Consider the compact set defined by $Z := \{z \in V \mid f(z) = f(z')\}$. Therefore, for any $z \in Z$,

$$\begin{aligned}f_k(z) &\leq \gamma/3 + \inf_V f \qquad (54a)\\ &\leq \beta/3 + \inf_V f \qquad (54b)\\ &< 2\beta/3 + \inf_V f \qquad (54c)\\ &\leq -\gamma/3 + \beta + \inf_V f \qquad (54d)\\ &\leq -\gamma/3 + \inf_{\partial V} f \leq \inf_{\partial V} f_k. \qquad (54e)\end{aligned}$$

Thus, $z \in \text{int}(V)$. So $Z \subseteq \text{int}(V)$. Since both Z and ∂V are compact, we have $d(\partial V, Z) > 0$. We now proceed to show by contradiction that Z is a strict local minimum of f_k. Assume that for all $\epsilon' > 0$, there exists $y' \in S \setminus Z$ satisfying $d(y', Z) \leq \epsilon'$ such that $f_k(z) \geq f_k(y')$ for some $z \in Z$. We can choose $\epsilon' < d(\partial V, Z)$ to guarantee that y' belongs to V since $Z \subseteq \text{int}(V)$. The point y' then contradicts the strict minimality of Z on V. This means that $Z \in V$ is a strict local minimum of f_k. Now, observe that for any $z \in Z$,

$$\begin{aligned}\inf_K f_k &\leq \gamma/3 + \inf_K f && \text{(55a)}\\ &\leq \gamma/3 + \alpha/2 + \inf_S f && \text{(55b)}\\ &\leq 2\alpha/3 + \inf_S f && \text{(55c)}\\ &< 5\alpha/6 + \inf_S f && \text{(55d)}\\ &\leq \alpha - \gamma/3 + \inf_S f && \text{(55e)}\\ &= -\gamma/3 + \inf_X f && \text{(55f)}\\ &= -\gamma/3 + \inf_V f && \text{(55g)}\\ &\leq \inf_V f_k \leq f_k(z). && \text{(55h)}\end{aligned}$$

Thus, Z is not a global minimum of f_k. This contradicts the fact that f_k is a weakly global function.

O Proof of Proposition 5

The proof for Proposition 5 is a natural extension of Proposition 2. Specifically, we show that Assumption A1 can be preserved for the input of every hidden layer, which allows us to prove Proposition 5 by induction.

Proof:(Proof of Proposition 5) Denote $W_{1:h} = (W_1, W_2, \cdots, W_h)$, i.e., the weights of the first h hidden layers. Define

$$\Omega_h = \{W_{1:h} \mid \operatorname{rank}(T_h) < \min\{d_h, N\}\} \tag{56}$$

$$\hat{\Omega}_h = \{W_{1:h} \mid \forall i = 1, \cdots, d_h, \ \ \exists n_i \neq n_i', \ \ s.t. \ \ (T_h)_{i,n_i} = (T_h)_{i,n_i'}\}. \tag{57}$$

Ω_h is the set of $W_{1:h}$ such that the output matrix of the h-th hidden layer is not full rank, which generalizes Ω_1 defined in Proposition 2. $\hat{\Omega}_h$ is the set of $W_{1:h}$ such that there exist identical entries in every row of T_h. That is, for any $W_{1:h} \in \hat{\Omega}_h$, the resulting T_h, if regarded as an input data matrix, violates Assumption A1.

In the following, we prove by induction that $\Phi_h \triangleq \Omega_h \cup \hat{\Omega}_h$ is of measure zero for all $1 \leq h \leq H$.

We first consider the case with $h = 1$. By Proposition 2, Ω_1 is of measure zero . Further, note that $(T_1)_{i,n} = \sigma((\mathbf{w}_1)_i^\top x^{(n)})$, where $(\mathbf{w}_1)_i^\top$ and $x^{(n)}$ are the i-th row of W_1 and the n-th column of X (i.e., the n-th training data), respectively. Noting that Assumption A1 guarantees that $x^{(1)}, x^{(2)}, \cdots, x^{(N)}$ are distinct from each other, from Lemma 3, $\hat{\Omega}_1$ is of measure zero. As a result, $\Phi_1 = \Omega_1 \cup \hat{\Omega}_1$ is also of measure zero.

Now assume that Φ_{h-1} is of measure zero. Then Φ_h can be decomposed into

$$\begin{aligned}\Phi_h = &\Big\{W_{1:h} \mid W_{1:(h-1)} \in \Phi_{h-1}, \ \ W_{1:h} \in \Omega_h \cup \hat{\Omega}_h\Big\}\\ &\cup \Big\{W_{1:h} \mid W_{1:(h-1)} \notin \Phi_{h-1}, \ \ W_{1:h} \in \Omega_h \cup \hat{\Omega}_h\Big\}\end{aligned} \tag{58}$$

By the induction hypothesis, the first component of the set union in equation 58 has zero measure in the space of $W_{1:h}$. Moreover, for $W_{1:(h-1)} \notin \hat{\Omega}_{h-1}$, the resulting T_{h-1}, if regarded as an input data matrix, satisfies Assumption A1. Noting that σ_h satisfies Assumption 3, following a similar procedure as in the case of $h = 1$, we obtain that the set of W_h satisfying $(W_{1:(h-1)}, W_h) \in \Omega_h$ has zero measure in $\mathbb{R}^{d_h \times d_{h-1}}$. This implies that the second component of the set union in equation 58 also has zero measure in the space of $W_{1:h}$. Therefore, Φ_h is of measure zero for all $1 \leq h \leq H$.

Noting that $\Omega = \Omega_H$, we complete the proof of Proposition 5 □

P Proof of Lemma 5

Let $D \subset \mathbb{R}^n$ be the range of f on S. Since S is compact and f is continuous, D is also compact. Define

$$D' = \{z \in \mathbb{R}^n \mid \exists z_0 \in D, \quad ||z - z_0||_2 \leq 1\} . \tag{59}$$

Then, D' is also compact.

Since g is continuous, its restriction on D' is uniformly continuous. That is, for any $\epsilon > 0$, there exits $\delta > 0$ such that

$$|g(z_1) - g(z_2)| \leq \frac{\epsilon}{2}, \quad \forall z_1, z_2 \in D', \quad ||z_1 - z_2||_2 \leq \delta. \tag{60}$$

Further, since f_k converges to f uniformly on S, there exists $K_1 \in \mathbb{N}$ such that

$$||f_k(x) - f(x)|| \leq \min\{1, \delta\}, \quad \forall k \geq K_1, \quad x \in S. \tag{61}$$

Note that by the definition of D', equation 61 also implies $f_k(x) \in D'$ for all $k \geq K_1$ and $x \in S$. Also, as g_k compactly converges to g, g_k uniformly converges to g on D'. Then, there exists $K_2 \in \mathbb{N}$ such that $|g_k(z) - g(z)| \leq \epsilon/2$ for all $k \geq K_2$ and $z \in D'$.

For any $k \geq \max\{K_1, K_2\}$ and $x \in S$, we have

$$\begin{aligned} |g_k(f_k(x)) - g(f(x))| &\leq |g_k(f_k(x)) - g(f_k(x))| + |g(f_k(x)) - g(f(x))| && \text{(62a)} \\ &\leq \frac{\epsilon}{2} + |g(f_k(x)) - g(f(x))| && \text{(62b)} \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} && \text{(62c)} \\ &= \epsilon && \text{(62d)} \end{aligned}$$

where equation 62b follows from the fact that $f_k(x), f(x) \in D'$, and equation 62b is due to equation 60 and equation 61.

Therefore, we conclude that $g_k \circ f_k$ converges to $g \circ f$ uniformly on S. We complete the proof.

Q Proof of Lemma 6

We denote the considered network by $\mathcal{N}$. From Lemma 4, for any $1 \leq h \leq H$, there exists a sequence of activation functions $(\sigma_{h,k})_{k \in \mathbb{N}}$, each satisfying Assumption 3, that uniformly converges to σ_h. For each $k \in \mathbb{N}$, we construct a neural network, denoted by $\mathcal{N}_k$, by replacing the activation function of the h-th hidden layer with $\sigma_{h,k}$, for all $h = 1, 2, \cdots, H$. For all $\mathcal{N}_k$, we assume the training dataset to be identical to that of $\mathcal{N}$. We also denote the output matrix of the h-th hidden layer by $T_h^{(k)}$ and the empirical loss by

$$E_k(W) = l\left(Y, W_{H+1} T_H^{(k)}\right). \tag{63}$$

From Theorem 3, E_k is a weakly global function with respect to W, $\forall k \in \mathbb{N}$.

Consider the sequence of the empirical loss functions $(E_k)_{k \in \mathbb{N}}$. In what follows, we prove that E_k compactly converges to E.

Consider an arbitrary compact subset S in the space of W. For any $W \in S$, define $\tilde{t}_{h,i,n}^{(k)}(W) = (T_h^{(k)})_{i,n}$ and $\tilde{t}_{h,i,n}(W) = (T_h)_{i,n}$ for any $k \in \mathbb{N}$, $1 \leq h \leq H$, $1 \leq i \leq d_h$, and $1 \leq n \leq N$. That is, we rewrite the output of each neuron in the hidden layers as a function of W. We prove by induction that every sequence $(\tilde{t}_{h,i,n}^{(k)})_{k \in \mathbb{N}}$ converges to $\tilde{t}_{h,i,n}$ uniformly on S.

For $h = 1$, we have

$$\tilde{t}_{h,i,n}^{(k)}(W) = \sigma_k \left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n} \right) \tag{64}$$

$$\tilde{t}_{h,i,n}(W) = \sigma \left(\sum_{j=1}^{d_0} (W_1)_{i,j} X_{j,n} \right). \tag{65}$$

Since σ_k uniformly converges to σ, $\tilde{t}^{(k)}_{1,j,n}$ also uniformly converges to $\tilde{t}_{1,j,n}$ on S for all $1 \leq j \leq d_1$, $1 \leq n \leq N$.

For $h > 1$, assume that $\tilde{t}^{(k)}_{h-1,i,n}$ uniformly converges to $\tilde{t}_{h-1,i,n}$ on S for all $1 \leq i \leq d_{h-1}$, $1 \leq n \leq N$. For the h-th layer, we have

$$\begin{aligned}\tilde{t}^{(k)}_{h,i,n}(W) &= \sigma_{h,k}\left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}\left(T^{(k)}_{h-1}\right)_{j,n}\right)\\ &= \sigma_{h,k}\left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}\tilde{t}^{(k)}_{h-1,j,n}(W)\right)\end{aligned} \tag{66}$$

$$\begin{aligned}\tilde{t}_{h,i,n}(W) &= \sigma_{h}\left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}(T_{h-1})_{j,n}\right)\\ &= \sigma_{h}\left(\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}\tilde{t}_{h-1,j,n}(W)\right).\end{aligned} \tag{67}$$

By the induction hypothesis, it is easy to show that $\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}\tilde{t}^{(k)}_{h-1,j,n}(W)$ uniformly converges to $\sum_{j=1}^{d_{h-1}} (W_h)_{i,j}\tilde{t}_{h-1,j,n}(W)$ on S. Note that $(\sigma_{h,k})_{k\in\mathbb{N}}$ converges to σ_h uniformly, and uniform convergence implies compact convergence. It directly follows from Lemma 5 that $\tilde{t}^{(k)}_{h,i,n}(W)$ converges to $\tilde{t}_{h,i,n}(W)$.

Therefore, we conclude that $\tilde{t}^{(k)}_{h,i,n}$ converges to $\tilde{t}_{h,i,n}$ uniformly on S for every $1 \leq h \leq H$, $1 \leq i \leq d_h$, and $1 \leq n \leq N$.

Now we consider the empirical loss

$$E_k(W) = l\left(Y, W_{H+1}T^{(k)}_H\right) \tag{68}$$

$$E(W) = l\left(Y, W_{H+1}T_H\right). \tag{69}$$

As every component of $T^{(k)}_H$ converges uniformly to the corresponding component of T_H, it can be shown that $W_{H+1}T^{(k)}_H$ converges uniformly to $W_{H+1}T_H$ on S. By Lemma 5, where we set both g_k and g to the loss function l, we have that E_k uniformly converges to E on S. Noting that S is an arbitrary compact subset in the space of W, the empirical loss E_k converges to E compactly on the space of W. Since $E_k(W)$ is a weakly global function for every $k \in \mathbb{N}$, by Proposition 4, $E(W)$ is also a weakly global function. We complete the proof.

R Proof of Lemma 7

Consider an arbitrary continuous and non-increasing (or non-decreasing) function $f : \mathbb{R} \to \mathbb{R}$. In this proof we assume that f is non-increasing. Note that the non-decreasing case can be proved by following the same idea, and we omit the details therein.

We construct a sequence of function $(f_k)_{k\in\mathbb{N}}$ as

$$f_k(x) = \begin{cases} -x - k - 1 + f(-k-1) & x < -k-1 \\ f(x) - \frac{x+k+1}{(k+1)^2} & -k-1 \leq x \leq k+1 \\ -x + k + 1 + f(k+1) - \frac{2}{k+1} & x > k+1 \end{cases} \tag{70}$$

First, we show that f_k is continuous. To this end, we only need to verify that f_k is left-continuous at $x = -k-1$ and right continuous at $x = k+1$. From equation 70, we have

$$\lim_{x\to(-k-1)^-} = f(-k-1) = f_k(-k-1) \tag{71a}$$

$$\lim_{x\to(k+1)^+} = f(k+1) - \frac{2}{k+1} = f_k(k+1) \tag{71b}$$

Thus, $f_k(x)$ is continuous for all $k \in \mathbb{N}$.

Second, we show that f_k is strictly decreasing and $f(\mathbb{R}) = \mathbb{R}$. On both $(-\infty, -k-1)$ and $(k+1, +\infty)$, f_k is linear with negative slope, and hence strictly decreasing. Noting that f_k is non-increasing, for any $x_1, x_2 \in [-k-1, k+1]$ and $x_1 < x_2$, we have

$$f_k(x_1) = f(x_1) - \frac{x_1 + k + 1}{(k+1)^2} \geq f(x_2) - \frac{x_1 + k + 1}{(k+1)^2} > f(x_2) - \frac{x_2 + k + 1}{(k+1)^2} = f_k(x_2). \quad (72)$$

Therefore f_k is strictly decreasing on $[-k-1, k+1]$. As f_k is continuous, we conclude that f_k is strictly decreasing on $\mathbb{R}$. Further, from equation 70 we have $\lim_{x \to -\infty} f_k(x) = +\infty$ and $\lim_{x \to +\infty} f_k(x) = -\infty$. Again, from the fact that f_k is continuous, we have $f_k(\mathbb{R}) = \mathbb{R}$.

Finally, we show that f_k converges to f compactly. Consider an arbitrary compact set $S \subset \mathbb{R}$ and $\epsilon > 0$. Since S is bounded, there exists $K \in \mathbb{N}$ such that $S \subset [-K-1, K+1]$. For any $k > \max\{K, \epsilon/2 - 1\}$ and $x \in S$, we have $x \in [-k-1, k+1]$, and

$$|f_k(x) - f(x)| = \left| f(x) - \frac{x + k + 1}{(k+1)^2} - f(x) \right| = \frac{|x + k + 1|}{(k+1)^2} \leq \frac{2}{k+1} < \epsilon. \quad (73)$$

Thus, f_k converges to f uniformly on S. As S is an arbitrary compact set on $\mathbb{R}$, f_k converges to f compactly on $\mathbb{R}$.

We complete the proof.