# COMPLEX NEURAL NETWORKS HAVE NO SPURIOUS LOCAL MINIMA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Most non-linear neural networks are known to have poor local minima (Yun et al., 2019) and it is shown that training a neural network is NP-hard (Blum & Rivest, 1988). A line of work has studied the global optimality of neural networks in various settings but unfortunately all previous networks without spurious local minima are linear networks or networks with unrealistic assumptions. In this work we demonstrate for the first time that a non-linear neural network can have no poor local minima under no assumptions.

Recently, a number of papers considered complex-valued neural networks (CVNNs) in various settings and suggest that CVNNs have competitive or even preferable behaviour compared to real-valued networks. Unfortunately, there is currently no theoretical analysis on the optimization of complex-valued networks, given that complex functions usually have a disparate optimization landscape. This is the first work towards analysing the optimization landscape of CVNNs. We prove a surprising result that no spurious local minima exist for one hidden layer complex-valued neural networks with quadratic activation. Since CVNNs can have real-valued datasets and there are no assumptions, our results are applicable to practical networks. Along the way, we develop a novel set of tools and techniques for analyzing the optimization of CVNNs, which may be useful in other contexts. Lastly, we prove spurious local minima exist for CVNNs with non-analytic $\mathbb{C}$ReLU activation.

## 1 INTRODUCTION

Neural networks have seen great success empirically, inspiring a lot of work theoretically, but most of them are real-valued networks. Thus, exploring complex-valued neural networks(CVNNs) is an interesting potential way to find better architectures. A number of papers considered CVNNs and suggested that CVNNs have richer representational capacity, faster learning ability, better generalization, and noise-robust memory mechanisms compared to real-valued networks (Trabelsi et al., 2018; Hirose & Yoshida, 2012; Arjovsky et al., 2016; Danihelka et al., 2016; Wisdom et al., 2016). This naturally leads one to inquire about the optimization landscape of CVNNs, in comparison to that of real-valued neural networks

The minimum modulus principle (MMP) is a fundamental result in complex analysis which can be roughly viewed as a statement that "analytic functions have no spurious local minima" with respect to the modulus. Below is a graph that compares a same function, $x\sin(x)$, on different fields $\mathbb{R}$ and $\mathbb{C}$. Since complex numbers are not ordered, the y-axis on the right graph is the square of the modulus of the range. We can see from Figure 1 that the left graph has many local minima while the right graph only has global minima. Although most CVNNs are not analytic, the MMP still provides an insight for us that there might be a potential for CVNNs to have a better optimization landscape due to the properties of complex numbers. Note that fully-connected multi-layer perceptrons with analytic activation functions are analytic with respect to either the input or each weight matrix. Therefore we conjecture that all CVNNs with analytic activations have a superior landscape. It should be noted that generally the loss function of a CVNN is not analytic because the loss is calculated by summing modulus. Non-analytic functions not only lose the MMP but also are non-differentiable. Thus, analyzing the optimization landscape of CVNNs is not trivial and can have no connection with the MMP. The main technique we apply in this paper is called Wirtinger calculus that we will describe later.

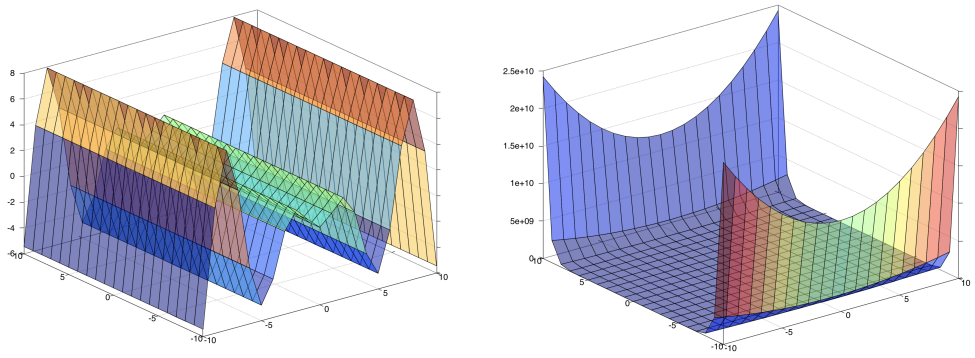

Figure 1: Left: $x \sin(x)$ vs. $x$ where $x \in \mathbb{R}$. Right: $||z \sin(z)||^2$ vs. $z$ where $z \in \mathbb{C}$. We plot the left one in three dimensions for consistency.

### 1.1 Related work

#### 1.1.1 Complex-valued neural networks

The study of complex-valued neural networks can be dated back to the last century (Benvenuto & Piazza, 1992; Little et al., 1990; Georgiou & Koutsougeras, 1992; Nitta, 2002) in the signal processing community. For some well-known deep learning architecture like CNN, RNN, and GAN, there are complex-valued counterparts (Arjovsky et al., 2016; Danihelka et al., 2016; Wisdom et al., 2016; Minin, 2012; Goodfellow et al., 2014; Guberman, 2016; Wolter & Yao, 2018; Dedmari et al., 2018; Sun et al., 2019). More recently, CVNNs started to gain attention in deep learning community. Trabelsi et al. (2018) proposed an extensive framework for complex-valued neural networks and demonstrated that they have a competitive performance compared to real-valued networks. Complex-valued RNNs were shown to have richer representational capacity, faster learning ability, and better memory mechanisms (Arjovsky et al., 2016; Danihelka et al., 2016; Wisdom et al., 2016). Hirose & Yoshida (2012) showed that CVNNs have better generalization characteristics. A connection between CVNNs and privacy protection was also explored (Xiang et al., 2020). In addition, a large number of papers have studied the application of CVNNs to fields such as quantum, medicine, geoscience, audio, image, NLP, signal processing, and more (Grant et al., 2018; Dedmari et al., 2018; Jingkun Gao & Li, 2019; Choi et al., 2019; Tay et al., 2018; Gaudet & Maida, 2018; Pande et al., 2008). We can see from the above that CVNNs are not just a theoretical curiosity, and have seen wide application in real-world problems.

#### 1.1.2 Optimization landscape of neural networks

Since the loss function of neural networks is non-convex, analyzing the optimization landscape of the loss is always hard. Given that poor local minima exist in common neural networks like over-parametrized ReLU networks (Yun et al., 2019), people try to prove poor local minima do not exist in other settings. Linear neural networks, for example, were proved to have no spurious local minima although having a non-convex loss (Baldi & Hornik., 1989; Baldi & Lu, 2012; Kawaguchi, 2016). More recently, there are results of "no spurious local minima" on more networks such as shallow quadratic networks and shallow ReLU networks (Wu et al., 2018; Soltanolkotabi et al., 2019; Ghorbani et al., 2019). Unfortunately, all prior works make unrealistic assumptions of one form or another, which we summarize and describe in Table 1. The assumptions we list in the table below indicate unrealistic ones. For the results on linear networks, it is assumed that the covariance of the training data is full rank, but this is not unrealistic.

**Assumption A.** A1p-m and A5u-m in (Kawaguchi, 2016).

**Assumption B.** Two hidden units. Weight vectors are unit-normed and orthogonal.

**Assumption C.** The weight vector $v \in \mathbb{R}^k$ connecting the hidden layer and the output node must contain at least $d$ positive entries and $d$ negative entries where $k \geq 2d$.

| Reference | Model | Linearity | Assumptions |
|---|---|---|---|
| (Baldi & Hornik., 1989) | Shallow linear networks | Linear | **None** |
| (Baldi & Lu, 2012) | Shallow complex linear networks | Linear | **None** |
| (Kawaguchi, 2016) | Deep linear networks | Linear | **None** |
| (Kawaguchi, 2016) | Deep ReLU networks | **Non-linear** | Asm. A |
| (Wu et al., 2018) | Shallow ReLU networks | **Non-linear** | Asm. B |
| (Soltanolkotabi et al., 2019) | Shallow quadratic networks | **Non-linear** | Asm. C |
| (Ghorbani et al., 2019) | Shallow quadratic networks | **Non-linear** | Asm. D |
| **Ours** | Shallow complex quadratic networks | **Non-linear** | **None** |

Table 1: Neural Networks Without Spurious Local Minima

**Assumption D.** Over-parametrization and feature vectors being Gaussian.

A5u-m is unrealistic as suggested in (Kawaguchi, 2016; Choromanska et al., 2015), where they assumed independence between hidden nodes. Assumption B and C are restrictive, and are unlikely to arise in real-world settings. Assumption C set the second weight vector to have at least d positive weights and at least d negative weights. There is no way to guarantee such setting during and after training, especially with algorithm like backpropagation. As for Assumption D, it is unlikely that the feature vectors will be Gaussian in real-world problems. We can see all previous analysis on non-linear neural networks have extremely unrealistic assumptions. Another work done in Kawaguchi & Kaelbling (2020) and Liang et al. (2018) suggested that adding one neuron can distinguish all spurious local minima, but local minima still exist with infinity norm. Our work has no assumption as in all the above. It is, therefore, the first non-linear neural network that has no poor local minima under no unrealistic assumption. Linearity is important because linear models can only fit data that are linearly separable. However, with enough parameters non-linear neural nets can approximate any function by universal approximation theorem (Cybenko.; Barron., 1993).

### 1.2 Our contribution

We prove that one hidden layer CVNNs with quadratic activation have no spurious local minima. The proof is built on (Soltanolkotabi et al., 2019), and it turns out that their assumptions can be avoided in the complex-valued setting. Theorem 1 states our result formally.

**Theorem 1.** *Assume the dataset is $\{x_i, y_i\}$ for $i = 1, 2, \ldots, n$ with $x_i \in \mathbb{C}^d$ and $y_i \in \mathbb{C}$. The training model we consider is one hidden layer complex-valued neural networks with quadratic activation. It is in the form of*

$$\boldsymbol{x} \mapsto \boldsymbol{v}^T \psi(\boldsymbol{W}\boldsymbol{x})$$

*with $\psi(z) = \psi((z_1, \ldots, z_d)) = (z_1^2, \ldots, z_d^2)$, $\boldsymbol{W} \in \mathbb{C}^{k \times d}, \boldsymbol{v} \in \mathbb{C}^k$ with $v_i \neq 0$, and $k \geq d$. Then the training loss as a function of the weights $\boldsymbol{W}$*

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}) &= \frac{1}{2n} \sum_{i=1}^{n} \parallel y_i - \boldsymbol{v}^T \psi(\boldsymbol{W}\boldsymbol{x}_i) \parallel^2 \\ &= \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{v}^T \psi(\boldsymbol{W}\boldsymbol{x}_i))^* (y_i - \boldsymbol{v}^T \psi(\boldsymbol{W}\boldsymbol{x}_i))\end{aligned}$$

*has no spurious local minima, i.e. all local minima are global.*

We require only the mild assumptions that $\mathbf{v}$ has non-zero entries and $k \geq d$. If an entry of zero is needed for achieving global minimum, we can make the corresponding row of $\mathbf{W}$ to be zero and have the same loss as being a global minimum. For the same reason, note that for any non-zero weights pair $(\widetilde{\mathbf{W}}, \widetilde{\mathbf{v}})$ that achieves the global minimum can be rescaled to have the same loss with any $\mathbf{v}$ we want. Thus, a global optimum with respect to $\widetilde{\mathbf{W}}$ is also a global optimum with respect to $(\widetilde{\mathbf{W}}, \widetilde{\mathbf{v}})$. $k \geq d$ implies the number of hidden nodes is larger than the number of input nodes, which is very common in over-parametrized deep learning era.

To prove Theorem 1 we make use of the semi-definite property of the Wirtinger hessian matrix. We will show that a local minimum $\mathbf{W}$ must satisfy the expression in Lemma 3 which makes it to be a global minimum. The first half of the proof is similar to the proof of Theorem 2.1 in (Soltanolkotabi et al., 2019). We derive a nice and simple expression of the hessian multiplied by an arbitrary direction $\mathbf{U}$. Wirtinger calculus is used to extend their steps to the complex case. The second half of the proof turns out to be very different because we avoided their main assumption that $\mathbf{v}$ having at least $d$ positive entries and at least $d$ negative entries, which is unrealistic. By using the property of complex numbers, we can show that a local minimum $\mathbf{W}$ must satisfy the expression in Lemma 3 by contradiction. A full proof can be found in Section 3.

We also explore the optimization landscape of $\mathbb{C}$ReLU activated CVNNs. See the definition of $\mathbb{C}$ReLU in Appendix A.3.

**Theorem 2.** *Assume the dataset is $\{x_i, y_i\}$ for $i = 1, 2, \ldots, n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The training model we consider is one hidden layer complex-valued neural networks with $\mathbb{C}$ReLU activation. It is in the form of*

$$\boldsymbol{x} \mapsto \boldsymbol{v}^T \mathbb{C}\text{ReLU}(\boldsymbol{W}\boldsymbol{x})$$

*with $\boldsymbol{W} \in \mathbb{C}^{k\times d}$, and $\boldsymbol{v} \in \mathbb{C}^k$. Suppose the dataset is real-valued, $x_i$'s are distinct, the hidden layer has a width of at least 2, and cannot be fitted linearly, then the loss function*

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n}\sum_{i=1}^{n} \parallel y_i - \boldsymbol{v}^T \mathbb{C}\text{ReLU}(\boldsymbol{W}\boldsymbol{x}) \parallel^2$$

*has infinitely many spurious local minima.*

Unlike quadratic functions, $\mathbb{C}$ReLU is not analytic. Therefore, CVNNs activated by $\mathbb{C}$ReLU having poor local minima is not beyond our expectation. The proof is built on (Yun et al., 2019) with some modifications to fit in the complex case. Firstly we construct a point that is as good as the global minimum of a linear model. Then we prove that point is a local minimum. The rest of the proof is the same as in (Yun et al., 2019). They showed that there exists a strictly better point to prove that the local minimum is spurious. The proof is provided in Appendix A.3 due to space constraints.

## 2 PRELIMINARIES

In this section we provide some introduction to complex analysis and Wirtinger calculus. See more definitions and lemmas in the appendix.

### 2.1 NOTATIONS AND USEFUL IDENTITIES

Let $\mathbb{R}$ denote the real field and $\mathbb{C}$ denote the complex field. Since both of them are fields, they share many common properties. Notice that $\mathbb{R} \subseteq \mathbb{C}$ and $\mathbb{R}^{m\times n} \subseteq \mathbb{C}^{m\times n}$. Let $z = z_1 + iz_2 \in \mathbb{C}$, we use $||z|| = \sqrt{z_1^2 + z_2^2} \in \mathbb{R}$ to denote its modulus and $z^* = z_1 - iz_2$ to denote its conjugate. $\mathcal{R}(\mathbf{z})$ and $\mathcal{I}(\mathbf{z})$ are used to denote the real and imaginary part of a complex vector $\mathbf{z} \in \mathbb{C}^n$. For $\mathbf{M} \in \mathbb{C}^{m\times n}$, $\mathbf{M}^T$ and $\mathbf{M}^*$ are transpose and conjugate transpose of $\mathbf{M}$, $\mathbf{M}^C$ denotes the matrix whose entries are conjugates of entries in $\mathbf{M}$, Null$(\mathbf{M}) := \{\mathbf{v} \mid \mathbf{M}\mathbf{v} = 0\}$ denotes the null space, and vec$(\mathbf{M})$ denotes the vectorization of $\mathbf{M}$. For $z \in \mathbb{C}$ and $\mathbf{M} \in \mathbb{C}^{m\times n}$, we have $z^* = z^C$ and $\mathbf{M}^* = (\mathbf{M}^C)^T$.

### 2.2 COMPLEX FUNCTIONS

A complex function $f : \mathbb{C} \mapsto \mathbb{C}$ is given by $f(z) = u(z) + iv(z)$. We can also think of $f$ as $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ where $f(x, y) = (u(x, y), v(x, y))$. A complex-valued multivariate function $f : \mathbb{C}^n \mapsto \mathbb{C}$ is given by $f(\mathbf{z}) = u(\mathbf{z}) + iv(\mathbf{z})$ where $u(\mathbf{z}), v(\mathbf{z}) \in \mathbb{R}$. The function we will consider most in this paper is a real-valued function with matrix input $f : \mathbb{C}^{m\times n} \mapsto \mathbb{R}$.

A complex function is analytic if it is differentiable at every point and the point's neighbourhood in the domain. An analytic function must satisfy the Cauchy Riemann equations (CRE). See Appendix A.1.1 for the definitions of differentiable, analytic, CRE, and more. We mention that a non-constant real-valued complex function does not satisfy the CRE and is not analytic.

Recall that the loss function in Theorem 1 is

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^{n} \parallel y_i - \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i) \parallel^2$$

which is not analytic. However, $y_i - \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i)$ is analytic for each $i$, which indicates their derivatives are well defined.

### 2.3 WIRTINGER CALCULUS

Since $\mathcal{L}(\mathbf{W})$ is not differentiable in the traditional sense, we require a new way of calculating the complex gradient. Many non-differentiable complex functions are in fact differentiable in the real sense if we treat $\mathbb{C}^n$ as $\mathbb{R}^{2n}$. $\mathcal{L}(\mathbf{W})$ is one of them. Wirtinger calculus, also known as the $\mathbb{CR}$-Calculus, is a neat way of deriving the derivatives. For a differentiable complex function, Wirtinger derivatives are the same as the traditional derivatives. By using Wirtinger calculus we not only have defined derivatives that reflect a function's gradient, but also have meaningful results on the first and second derivatives. Critical points, positive semi-definite Hessian, and Taylor expansions all have their counterparts in Wirtinger calculus. We provide few important exposition of Wirtinger calculus here. More explanations are provided in the appendix and a systematic introduction can be found in (Kreutz-Delgado, 2005) and (Bouboulis, 2010).

Consider the complex-valued function $f : \mathbb{C}^n \mapsto \mathbb{C}$, $f(\mathbf{z}) = u(\mathbf{x}, \mathbf{y}) + iv(\mathbf{x}, \mathbf{y})$. The Wirtinger derivative and the conjugate Wirtinger derivative are defined to be

$$\frac{\partial f}{\partial \mathbf{z}} := \left[\frac{\partial f}{\partial z_1}, \ldots, \frac{\partial f}{\partial z_n}\right], \ \frac{\partial f}{\partial \mathbf{z}^C} := \left[\frac{\partial f}{\partial z_1^*}, \ldots, \frac{\partial f}{\partial z_n^*}\right]$$

where

$$\frac{\partial f}{\partial z_j} := \frac{1}{2}\left(\frac{\partial f}{\partial x_j} - i\frac{\partial f}{\partial y_j}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x_j} + \frac{\partial v}{\partial y_j}\right) + \frac{i}{2}\left(\frac{\partial v}{\partial x_j} - \frac{\partial u}{\partial y_j}\right),$$
$$\frac{\partial f}{\partial z_j^*} := \frac{1}{2}\left(\frac{\partial f}{\partial x_j} + i\frac{\partial f}{\partial y_j}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x_j} - \frac{\partial v}{\partial y_j}\right) + \frac{i}{2}\left(\frac{\partial v}{\partial x_j} + \frac{\partial u}{\partial y_j}\right).$$

Note that the Wirtinger derivative is well defined as long as the real functions $u$ and $v$ are differentiable with respect to $\mathbf{x}$ and $\mathbf{y}$. In our case, the loss function $\mathcal{L}(\mathbf{W})$ has well-defined Wirtinger derivative.

See the definition of conjugate-complex derivative and conjugate-complex differentiable in the appendix. We now have the following lemma which follows directly from the definitions.

**Lemma 1.** *If $f$ is complex differentiable, then its Wirtinger derivative is the same as the normal derivative, while the conjugate Wirtinger derivative is equal to zero.*

$$\frac{\partial f}{\partial \mathbf{z}} = f^{'}, \ \frac{\partial f}{\partial \mathbf{z}^C} = \mathbf{0}.$$

*Similarly, if $f$ is conjugate-complex differentiable, then its conjugate Wirtinger derivative is equal to the normal conjugate-complex derivative, while the Wirtinger derivative is equal to zero.*

$$\frac{\partial f}{\partial \mathbf{z}^C} = f_*^{'}, \ \frac{\partial f}{\partial \mathbf{z}} = \mathbf{0}.$$

Wirtinger derivative share many properties as normal derivatives like linearity, product rule, and chain rule. We have more explanations in the appendix.

Lastly we provide expressions for Wirtinger gradient, Wirtinger Hessian, and the second order Taylor's expansion formula.

$$\widetilde{\nabla} f(\mathbf{z}) = \left[\frac{\partial f}{\partial \mathbf{z}}, \frac{\partial f}{\partial \mathbf{z}^C}\right]^*$$

$$\widetilde{\nabla}^2 f(\mathbf{z}) = \begin{pmatrix} \frac{\partial}{\partial \mathbf{z}}(\frac{\partial f}{\partial \mathbf{z}})^* & \frac{\partial}{\partial \mathbf{z}^C}(\frac{\partial f}{\partial \mathbf{z}})^* \\ \frac{\partial}{\partial \mathbf{z}}(\frac{\partial f}{\partial \mathbf{z}^C})^* & \frac{\partial}{\partial \mathbf{z}^C}(\frac{\partial f}{\partial \mathbf{z}^C})^* \end{pmatrix}$$

$$f(\mathbf{z}+\mathbf{h}) = f(\mathbf{z}) + (\widetilde{\nabla} f(\mathbf{z}))^* \cdot \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^C \end{pmatrix} + \frac{1}{2}(\mathbf{h}^*, \mathbf{h}^T) \cdot \widetilde{\nabla}^2 f(\mathbf{z}) \cdot \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^C \end{pmatrix} + o(||\mathbf{h}||^2)$$

A point $\mathbf{z}$ is called a critical point of $f$ if and only if $\widetilde{\nabla} f(\mathbf{z}) = \mathbf{0}$. Since the loss function we will be analyzing is real-valued, as in the standard setting, if $\mathbf{W}$ is a local minimum of $\mathcal{L}(\mathbf{W})$, then the Wirtinger's Hessian of $\mathcal{L}(\mathbf{W})$ is positive semi-definite. The formal statement is provided in Section 3 and the proof can be found in the appendix.

## 3 CVNNs HAVE BETTER OPTIMIZATION LANDSCAPE

We prove Theorem 1 in this section step by step. The framework and techniques used here can provide insights for future work in analysing the optimization landscape of complex networks.

### 3.1 DERIVATIVE CALCULATIONS

Firstly we observe that $\mathcal{L}(\mathbf{W})$ is a function which maps complex input to real output, i.e. $\mathcal{L}(\mathbf{W}) : \mathbb{C}^{k\times d} \mapsto \mathbb{R}$, and it is not differentiable because conjugate functions do not satisfy the CRE (Cauchy-Riemann Equation). Now let $\mathcal{L}(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^{n} \mathcal{L}_i(\mathbf{W})^* \mathcal{L}_i(\mathbf{W})$ where $\mathcal{L}_i(\mathbf{W}) : \mathbb{C}^{k\times d} \mapsto \mathbb{C}$ given by $\mathcal{L}_i(\mathbf{W}) = y_i - \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i)$. Observe that $\mathcal{L}_i(\mathbf{W})$ is complex differentiable in the traditional sense and thus $\mathcal{L}_i(\mathbf{W})$ has well-defined first and second derivatives. For a fixed $i$, we let $\mathcal{G}_i(\mathbf{W}) = \mathcal{L}_i(\mathbf{W})^* \mathcal{L}_i(\mathbf{W})$. We now show how to calculate the derivatives of $\mathcal{L}_i(\mathbf{W})$ and $\mathcal{G}_i(\mathbf{W})$. The derivatives of $\mathcal{L}(\mathbf{W})$ follow easily by linearity.

#### 3.1.1 DERIVATIVE CALCULATIONS OF $\mathcal{L}_i(\mathbf{W})$

Now we have

$$\mathcal{L}_i(\mathbf{W}) = \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i) - y_i$$

which is complex differentiable. The first derivative with respect to the $q$-th row of $\mathbf{W}$ is denoted by $\nabla_{\mathbf{w}_q}\mathcal{L}_i(\mathbf{W})$ and we have

$$\nabla_{\mathbf{w}_q}\mathcal{L}_i(\mathbf{W}) = v_q\psi^{'}(\langle \mathbf{w}_q, \mathbf{x}_i\rangle)\mathbf{x}_i$$
$$\nabla_{\mathbf{W}}\mathcal{L}_i(\mathbf{W}) = \mathbf{D}_{\mathbf{v}}\psi^{'}(\mathbf{W}\mathbf{x}_i)\mathbf{x}_i^T$$

where $\mathbf{D}_{\mathbf{v}} = \text{diag}(v_1, \ldots, v_k)$. And the second derivatives

$$\frac{\partial^2}{\partial \mathbf{w}_p^2}\mathcal{L}_i(\mathbf{W}) = v_p\psi^{''}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)\mathbf{x}_i\mathbf{x}_i^T$$

$$\frac{\partial^2}{\partial \mathbf{w}_p \mathbf{w}_q}\mathcal{L}_i(\mathbf{W}) = 0$$

for $p \neq q$.

#### 3.1.2 DERIVATIVE CALCULATIONS OF $\mathcal{G}_i(\mathbf{W})$

By the product rule of Wirtinger calculus, we have
$\nabla_{\mathbf{W}}\mathcal{G}_i(\mathbf{W}) = \nabla_{\mathbf{W}}(\mathcal{L}_i^*\mathcal{L}_i)(\mathbf{W}) = \nabla_{\mathbf{W}}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) + \nabla_{\mathbf{W}}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W}) = \nabla_{\mathbf{W}}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W})$.
$\nabla_{\mathbf{W}}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) = 0$ since $\mathcal{L}_i^*$ is conjugate-complex differentiable. Similarly,
$\nabla_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W}) = \nabla_{\mathbf{W}^C}(\mathcal{L}_i^*\mathcal{L}_i)(\mathbf{W}) = \nabla_{\mathbf{W}^C}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) + \nabla_{\mathbf{W}^C}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W}) = \nabla_{\mathbf{W}^C}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W})$.
Based on that we have the second derivatives

$$\nabla^2_{\mathbf{W}}\mathcal{G}_i(\mathbf{W}) = \nabla^2_{\mathbf{W}}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W})$$
$$\nabla^2_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W}) = \nabla^2_{\mathbf{W}^C}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W})$$
$$\nabla_{\mathbf{w}_p}\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W}) = v_p^*v_p\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)^*\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)\mathbf{x}_i^C\mathbf{x}_i^T$$
$$\nabla_{\mathbf{w}_q}\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W}) = v_p^*v_q\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)^*\psi^{'}(\langle \mathbf{w}_q, \mathbf{x}_i\rangle)\mathbf{x}_i^C\mathbf{x}_i^T$$
$$\nabla_{\mathbf{w}_p^C}\nabla_{\mathbf{w}_p}\mathcal{G}_i(\mathbf{W}) = v_p^*v_p\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)^*\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)\mathbf{x}_i\mathbf{x}_i^*$$
$$\nabla_{\mathbf{w}_q^C}\nabla_{\mathbf{w}_p}\mathcal{G}_i(\mathbf{W}) = v_pv_q^*\psi^{'}(\langle \mathbf{w}_p, \mathbf{x}_i\rangle)\psi^{'}(\langle \mathbf{w}_q, \mathbf{x}_i\rangle)^*\mathbf{x}_i\mathbf{x}_i^*.$$

Notice that $\nabla_{\mathbf{W}}\nabla_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W})$ and $\nabla_{\mathbf{W}^C}\nabla_{\mathbf{W}}\mathcal{G}_i(\mathbf{W})$ are $kd \times kd$ matrices.

### 3.2 TECHNICAL LEMMAS

We provide some important lemmas before proving the theorem. The proofs can be found in the appendix.

**Lemma 2.** If $\widetilde{\mathbf{W}}$ is a local minimum of $\mathcal{L}(\widetilde{\mathbf{W}})$,

$$0 \leq (\mathbf{h}^*, \mathbf{h}^T) \cdot \widetilde{\nabla}^2 \mathcal{L}(\widetilde{\mathbf{W}}) \cdot \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^C \end{pmatrix} \in \mathbb{R}$$

where

$$\widetilde{\nabla}^2 \mathcal{L}(\widetilde{\mathbf{W}}) = \begin{pmatrix} \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{L}(\widetilde{\mathbf{W}}) & \nabla^2_{\mathbf{W}^C} \mathcal{L}(\widetilde{\mathbf{W}}) \\ \nabla^2_{\mathbf{W}} \mathcal{L}(\widetilde{\mathbf{W}}) & \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{L}(\widetilde{\mathbf{W}}) \end{pmatrix}$$

for all $\mathbf{h} \in \mathbb{C}^{kd}$.

*Proof.* See Appendix.

**Lemma 3.** (Extension of Lemma 6.1 in (Soltanolkotabi et al., 2019) to the complex case.) Any point $\widetilde{\mathbf{W}} \in \mathbb{C}^{k \times d}$ obeying

$$\frac{1}{2n} \sum_{i=1}^{n} (\mathbf{x}_i^T \widetilde{\mathbf{W}}^T \text{diag}(\mathbf{v}) \widetilde{\mathbf{W}} \mathbf{x}_i - y_i)^* \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{2n} \sum_{i=1}^{n} \mathcal{L}_i^*(\widetilde{\mathbf{W}}) \mathbf{x}_i \mathbf{x}_i^T = 0$$

is a global optimum of the loss function

$$\begin{aligned} \mathcal{L}(\mathbf{W}) &= \frac{1}{2n} \sum_{i=1}^{n} \parallel \mathbf{x}_i^T \mathbf{W}^T \text{diag}(\mathbf{v}) \mathbf{W} \mathbf{x}_i - y_i \parallel^2 \\ &= \frac{1}{2n} \sum_{i=1}^{n} \parallel y_i - \mathbf{v}^T \psi(\mathbf{W} \mathbf{x}_i) \parallel^2 . \end{aligned}$$

*Proof.* See Appendix.

### 3.3 PROOF OF THEOREM 1

Let $\mathbf{W} \in \mathbb{C}^{k \times d}$ be a local minimum. Let $\mathbf{U} \in \mathbb{C}^{k \times d}$ be an arbitrary direction and $\mathbf{h} = \text{vec}(\mathbf{U})$. We define

$$\begin{aligned} \mathcal{H} &= \frac{1}{2} (\mathbf{h}^*, \mathbf{h}^T) \cdot \widetilde{\nabla}^2 \mathcal{L}(\mathbf{W}) \cdot \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^C \end{pmatrix} \\ &= \frac{1}{2} (\text{vec}(\mathbf{U})^*, \text{vec}(\mathbf{U})^T) \begin{pmatrix} \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{L}(\mathbf{W}) & \nabla^2_{\mathbf{W}^C} \mathcal{L}(\mathbf{W}) \\ \nabla^2_{\mathbf{W}} \mathcal{L}(\mathbf{W}) & \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \end{pmatrix} \begin{pmatrix} \text{vec}(\mathbf{U}) \\ \text{vec}(\mathbf{U})^C \end{pmatrix} \end{aligned}$$

By linearity,

$$\mathcal{H} = \frac{1}{2} (\text{vec}(\mathbf{U})^*, \text{vec}(\mathbf{U})^T) \begin{pmatrix} \frac{1}{2n} \sum_{i=1}^{n} \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) & \frac{1}{2n} \sum_{i=1}^{n} \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \\ \frac{1}{2n} \sum_{i=1}^{n} \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) & \frac{1}{2n} \sum_{i=1}^{n} \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \end{pmatrix} \begin{pmatrix} \text{vec}(\mathbf{U}) \\ \text{vec}(\mathbf{U})^C \end{pmatrix}.$$

For each term we have

$$\begin{aligned} &(\text{vec}(\mathbf{U})^*, \text{vec}(\mathbf{U})^T) \begin{pmatrix} \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) & \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \\ \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) & \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \end{pmatrix} \begin{pmatrix} \text{vec}(\mathbf{U}) \\ \text{vec}(\mathbf{U})^C \end{pmatrix} \\ &= 2\mathcal{R}(\text{vec}(\mathbf{U})^T \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}) + \text{vec}(\mathbf{U})^* \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U})). \end{aligned}$$

Now we consider two cases. Case 1 is for $\text{rank}(\mathbf{D_v W}) = d$ and case 2 is for $\text{rank}(\mathbf{D_v W}) < d$. For the first case, since $k \geq d$ and $\text{rank}(\mathbf{D_v W}) = d$, $\mathbf{D_v W}$ has a left inverse $\mathbf{K} \in \mathbb{C}^{d \times k}$ such that $\mathbf{K D_v W} = \mathbf{I}$. Notice that by $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = 0$ and $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = \frac{1}{2n} \sum_{i=1}^{n} \nabla_{\mathbf{W}} \mathcal{L}_i(\mathbf{W}) \mathcal{L}_i^*(\mathbf{W}) = \mathbf{D_v W} (\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i^*(\mathbf{W}) \mathbf{x}_i \mathbf{x}_i^T)$, we have

$$\mathbf{D_v W} (\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i^*(\mathbf{W}) \mathbf{x}_i \mathbf{x}_i^T) = 0.$$

Multiplying both sides by $\mathbf{K}$ we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T=0,$$

which concludes the proof by Lemma 3.

For the second case, we can let $\mathbf{U}=\mathbf{a}\mathbf{b}^T$ with $\mathbf{a}\in\mathbb{C}^k$ and $\mathbf{D_v a}\in\text{Null}(\mathbf{W}^T)$. $\mathbf{b}\in\mathbb{C}^d$ is an arbitrary vector. We now show that $\text{vec}(\mathbf{U})^*\nabla_{\mathbf{W}}\nabla_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W})\text{vec}(\mathbf{U})=0$. Recall that

$$\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W})=\nabla_{\mathbf{w}_p^C}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W})=v_p^*\psi^{'}(\langle\mathbf{w}_p,\mathbf{x}_i\rangle)^*\mathbf{x}_i^C\mathcal{L}_i(\mathbf{W})$$
$$\nabla_{\mathbf{w}_p}\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W})=v_p^*v_p\psi^{'}(\langle\mathbf{w}_p,\mathbf{x}_i\rangle)^*\psi^{'}(\langle\mathbf{w}_p,\mathbf{x}_i\rangle)\mathbf{x}_i^C\mathbf{x}_i^T=\parallel v_p\psi^{'}(\langle\mathbf{w}_p,\mathbf{x}_i\rangle)\parallel^2\mathbf{x}_i^C\mathbf{x}_i^T$$
$$\nabla_{\mathbf{w}_q}\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W})=v_p^*v_q\psi^{'}(\langle\mathbf{w}_p,\mathbf{x}_i\rangle)^*\psi^{'}(\langle\mathbf{w}_q,\mathbf{x}_i\rangle)\mathbf{x}_i^C\mathbf{x}_i^T.$$

Therefore we can treat $\nabla_{\mathbf{W}}\nabla_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W})$ as a $k\times k$ matrix with each entry being a $d\times d$ matrix. Now by some algebra we have

$$\text{vec}(\mathbf{U})^*\nabla_{\mathbf{W}}\nabla_{\mathbf{W}^C}\mathcal{G}_i(\mathbf{W})\text{vec}(\mathbf{U})=\parallel\mathbf{x}_i^T\mathbf{W}^T\mathbf{D_v}\mathbf{U}\mathbf{x}_i\parallel^2$$

where

$$\mathbf{x}_i^T\mathbf{W}^T\mathbf{D_v}\mathbf{U}\mathbf{x}_i=\mathbf{x}_i^T\mathbf{W}^T\mathbf{D_v}\mathbf{a}\mathbf{b}^T\mathbf{x}_i=0.$$

Now by linearity and Lemma 2 we have

$$\mathcal{H}=\frac{1}{2n}\mathcal{R}(\text{vec}(\mathbf{U})^T\sum_{i=1}^{n}\nabla_{\mathbf{W}}^2\mathcal{G}_i(\mathbf{W})\text{vec}(\mathbf{U}))\geq 0$$

where

$$\begin{aligned}\sum_{i=1}^{n}\text{vec}(\mathbf{U})^T\nabla_{\mathbf{W}}^2\mathcal{G}_i(\mathbf{W})\text{vec}(\mathbf{U})&=\sum_{i=1}^{n}\text{vec}(\mathbf{U})^T\nabla_{\mathbf{W}}^2\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W})\text{vec}(\mathbf{U})\\&=2\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})(\mathbf{x}_i^T\mathbf{U}^T\mathbf{D_v}\mathbf{U}\mathbf{x}_i)\\&=2(\mathbf{a}^T\mathbf{D_v}\mathbf{a})\mathbf{b}^T\left(\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{b}.\end{aligned}$$

We argue that we can assume $(\mathbf{a}^T\mathbf{D_v}\mathbf{a})\neq 0$ here. The reason is the following. Since $\mathbf{a}\neq\mathbf{0}$, there is an entry $a_i\neq 0$. Suppose $\mathbf{a}^T\mathbf{D_v}\mathbf{a}=0$, we can multiply $v_i$ by 4 and multiply the the $i'$th row of $\mathbf{W}$ by $\frac{1}{2}$. Now $\mathbf{a}^T\mathbf{D}_{\mathbf{v}_{new}}\mathbf{a}=3v_ia_i^2\neq 0$. Note that the two weight matrices $\mathbf{W}$ and $\mathbf{W}_{new}$ have the same null space. By lemma 2 the old matrix $\mathbf{W}$ together with $\mathbf{v}$ is a global minimum if and only if the new matrix $\mathbf{W}_{new}$ together with $\mathbf{v}_{new}$ is a global minimum, because their corresponding $\mathbf{M}=\mathbf{W}^T\text{diag}(\mathbf{v})\mathbf{W}$ is the same. Therefore, proving $\mathbf{W}_{new}$ is a global minimum of $\mathcal{L}_i$ is equivalent to proving $\mathbf{W}$ is a global minimum. Thus, we can assume $(\mathbf{a}^T\mathbf{D_v}\mathbf{a})\neq 0$ without loss of generality.

Let $2(\mathbf{a}^T\mathbf{D_v}\mathbf{a})=a_1+ia_2\in\mathbb{C}$ and $\mathbf{b}^T\left(\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{b}=b_1+ib_2\in\mathbb{C}$. We now prove $\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T=0$ by contradiction. Since $\mathcal{H}=\frac{1}{2n}\mathcal{R}((a_1+a_2i)\cdot(b_1+b_2i))=\frac{1}{2n}(a_1b_1-a_2b_2)\geq 0$, we prove if $\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T\neq 0$ then $\mathcal{H}<0$ for some $(b_1,b_2)$. Since for a fixed pair $(a_1,a_2)$ we can make $a_1b_1-a_2b_2$ a negative number simply by setting the signs of $(b_1,b_2)$ according to the signs of $(a_1,a_2)$. For example, if $a_1>0$ and $a_2<0$ then $a_1b_1-a_2b_2<0$ for $(b_1,b_2)$ that $b_1<0$ and $b_2<0$. Now let

$$\mathcal{M}=\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T\neq 0,\mathcal{M}\in\mathbb{C}^{d\times d}.$$

Let $\mathcal{M}_{i,j}$ denotes the entry on the $i$'th row and the $j$'th column of $\mathcal{M}$.

We prove that we can have any sign on $b_1$ and $b_2$, which implies $\mathcal{M}$ must be zero. Suppose there exists a $i\in[d]$ such that $\mathcal{M}_{i,i}\neq 0$, then we let $\mathbf{b}=(0,\ldots,\beta_i,\ldots,0)^T$ where $\beta_i$ can be any complex number we want. Therefore, $b_1+ib_2=\mathcal{M}_{i,i}\cdot\beta_i^2$ can have any sign. Suppose $\mathcal{M}_{i,i}=0$ for all $i\in[d]$ and $\mathcal{M}_{i,j}\neq 0$ for some $(i,j)$, then we let $\mathbf{b}=(0,\ldots,\beta_i,\ldots,0,\ldots,\beta_j,\ldots,0)^T$. Now $b_1+ib_2=2\mathcal{M}_{i,j}\cdot\beta_i\cdot\beta_j$ which can have any sign.

Therefore, $\mathbf{W}$ is a global minimum by Lemma 3. □

## 4 DISCUSSION AND FUTURE WORK

We studied the optimization of complex-valued neural networks for the first time. We proved spurious local minima do not exist for a CVNN with analytic activation. The properties of complex numbers endow CVNNs with better optimization landscape compared to real-valued networks. Our result can serve as a strong reason for using complex networks. We showed that spurious local minima exist for a CVNN with non-analytic activation. Thus, a promising future research direction will be investigating the optimization landscape of CVNNs with analytic activations, for example, tanh. It will also be interesting to extend our result to deep networks. Therefore, complex-valued neural networks have the potential to be deep non-linear neural networks with practical activations having no spurious local minima. Such result would be invaluable in modern deep learning era because existence of poor local minima is one of the most intractable problems in deep learning due to the non-convexity in optimization of neural networks.

## REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. 2016.

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

Pierre Baldi and Zhiqin Lu. Complex-valued autoencoders. *Neural Networks*, 33:136–147, 2012.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

N. Benvenuto and F. Piazza. On the complex backpropagation algorithm. *IEEE Transactions on Signal Processing*, 40(4):967–969, 1992.

Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT '88, pp. 9–18, 1988.

Pantelis Bouboulis. Wirtinger's calculus in general hilbert spaces. 2010.

Hyeong-Seok Choi, Janghyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex U-net. In *International Conference on Learning Representations*, 2019.

Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. volume 40 of *Proceedings of Machine Learning Research*, pp. 1756–1760, 2015.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, (4):303–314.

Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, and Alex Graves. Associative long short-term memory. Proceedings of Machine Learning Research, pp. 1986–1994, 2016.

Muneer Ahmad Dedmari, Sailesh Conjeti, Santiago Estrada, Phillip Ehses, Tony Stöcker, and Martin Reuter. Complex fully convolutional neural networks for MR image reconstruction. In *Machine Learning for Medical Image Reconstruction*, pp. 30–38, 2018.

Chase Gaudet and Anthony Maida. Deep quaternion networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2018.

G. M. Georgiou and C. Koutsougeras. Complex domain backpropagation. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 39(5):330–334, 1992.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems 32*, pp. 9111–9121. 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. 2014.

E. Grant, Marcello Benedetti, Shuxiang Cao, A. Hallam, J. Lockhart, V. Stojevic, Andrew G. Green, and S. Severini. Hierarchical quantum classifiers. *npj Quantum Information*, 4:1–8, 2018.

Nitzan Guberman. On complex valued convolutional neural networks. *ArXiv*, abs/1602.09046, 2016.

Akira Hirose and Shotaro Yoshida. Generalization characteristics of complex-valued feedforward neural networks in relation to signal coherence. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):541–551, 2012.

Yuliang Qin Hongqiang Wang Jingkun Gao, Bin Deng and Xiang Li. Enhanced radar imaging using a complex-valued convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 16(1):35–39, 2019.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pp. 586–594. 2016.

Kenji Kawaguchi and Leslie Kaelbling. Elimination of all bad local minima in deep learning. volume 108 of *Proceedings of Machine Learning Research*, pp. 853–863, 2020.

Ken Kreutz-Delgado. The complex gradient operator and the CR-calculus. 2005.

Shiyu Liang, Ruoyu Sun, Jason D. Lee, and R. Srikant. Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems 31*, pp. 4350–4360. 2018.

Gordon R. Little, Steven C. Gustafson, and Robert A. Senn. Generalization of the backpropagation neural network learning algorithm to permit complex weights. *Applied Optics*, 29(11):1591–1592, 1990.

Alexey Minin. Complex valued recurrent neural network: From architecture to training. *Journal of Signal and Information Processing*, 03:192–197, 2012.

Tohru Nitta. On the critical points of the complex-valued neural network. In *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, volume 3, pp. 1099–1103 vol.3, 2002.

Anupama Pande, Ashok Kumar Thakur, and Swapnoneel Roy. Complex-valued neural network in signal processing: A study on the effectiveness of complex valued generalized mean neuron model. *International Journal of Electrical and Computer Engineering*, 2(1):39 – 44, 2008.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inf. Theor.*, 65(2): 742–769, 2019.

Qigong Sun, Xiufang Li, Lingling Li, Xu Liu, Fang Liu, and Licheng Jiao. Semi-supervised complex-valued GAN for polarimetric SAR image classification. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3245–3248, 2019.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. Hermitian co-attention networks for text matching in asymmetrical domains. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 4425–4431, 2018.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018.

Scott Wisdom, Thomas Powers, John R. Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4887–4895, 2016.

Moritz Wolter and Angela Yao. Complex gated recurrent neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 10536–10546. 2018.

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

Liyao Xiang, Hao Zhang, Haotian Ma, Yifan Zhang, Jie Ren, and Quanshi Zhang. Interpretable complex-valued neural networks for privacy protection. In *International Conference on Learning Representations*, 2020.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019.

# A APPENDIX

## A.1 SUPPLEMENTS OF SECTION 2

### A.1.1 MORE ON COMPLEX ANALYSIS

We provide some basic definitions of univariate complex functions. The generalization of multivariate functions is the same as in the real case.

Let $f : \mathbb{C} \mapsto \mathbb{C}$ given by $f(z) = u(z) + iv(z)$ where $z = x + iy$.

**Definition 1.** Suppose that $f$ is defined on some open neighbourhood of $z_0$. Then, the derivative of $f$ at $z_0$ is given by

$$f^{'}(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

where $\Delta z = \Delta x + i\Delta y$, provided this limit exists. Such an $f$ is said to be differentiable at $z_0$.

**Definition 2** (Analytic functions). A complex function $f(z)$ is called analytic at the point $z_0$ if it is differentiable at $z_0$ and in a neighbourhood of $z_0$.

Some examples of analytic functions include all polynomials, trigonometric functions, and exponential functions.

**Definition 3** (Cauchy-Riemann equations). If $f^{'}(z)$ exists, the partials of $u$ and $v$ exist at $(x, y)$ and satisfy the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x}(x, y) = \frac{\partial v}{\partial y}(x, y) \text{ and } \frac{\partial u}{\partial y}(x, y) = -\frac{\partial v}{\partial x}(x, y).$$

**Theorem 3** (Necessary conditions for differentiability). Suppose that $f$ is differentiable at $z$. Then the Cauchy-Riemann equations hold at $z$ and $f^{'}(z) = \frac{\partial u}{\partial x}(x, y) + i\frac{\partial v}{\partial x}(x, y) = \frac{\partial v}{\partial y}(x, y) - i\frac{\partial u}{\partial y}(x, y)$.

**Theorem 4** (Sufficient conditions for differentiability). Suppose $f(z)$ is defined throughout some open neighbourhood $U$ of the point $z_0 = x_0 + iy_0$, and suppose that $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ exist everywhere in $U$. Then, if $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are continuous at $(x_0, y_0)$ and satisfy the Cauchy-Riemann equations at $(x_0, y_0)$, then $f$ is differentiable at $z_0$ and $f^{'}(z) = \frac{\partial u}{\partial x}(x, y) + i\frac{\partial v}{\partial x}(x, y) = \frac{\partial v}{\partial y}(x, y) - i\frac{\partial u}{\partial y}(x, y)$.

Let $z^*$ denotes the conjugate of $z$ and $|z|$ denotes the modulus of $z$. We list some properties of complex numbers. For all $z, y \in \mathbb{C}$, we have the following.

$$|z^*| = |z|$$

$$zz^* = |z|^2$$

$$z^{-1} = \frac{z^*}{|z|^2} \text{ if } z \neq 0$$

$$\mathcal{R}(z) = \frac{z + z^*}{2} \text{ and } \mathcal{I}(z) = \frac{z - z^*}{2i}$$

$$|zy| = |z||y|$$

$$|z^n| = |z|^n$$

### A.1.2 MORE ON WIRTINGER CALCULUS

We provide more details on Wirtinger calculus that we use in proving theorem 1.

Let $f : \mathbb{C}^n \mapsto \mathbb{C}$ given by $f(\mathbf{z}) = u(\mathbf{z}) + iv(\mathbf{z})$ where $\mathbf{z} = \mathbf{x} + i\mathbf{y}$.

**Definition 4.** Conjugate Cauchy Riemann conditions (CCRC).

$$\frac{\partial u}{\partial \mathbf{x}} = -\frac{\partial v}{\partial \mathbf{y}} \text{ and } \frac{\partial u}{\partial \mathbf{y}} = \frac{\partial v}{\partial \mathbf{x}}$$

**Proposition 1.** If $f$ is differentiable in the real sense at $(\mathbf{x}, \mathbf{y})$ and the CCRC hold, then $f$ is conjugate-complex differentiable.

**Proposition 2.** If $f$ is conjugate-complex differentiable at $\mathbf{z}$ then $u$ and $v$ are differentiable in the real sense and they satisfy the conjugate Cauchy Riemann conditions.

**Proposition 3.** If $f$ is differentiable in the real sense, then

$$(\frac{\partial f}{\partial \mathbf{z}})^C = \frac{\partial f^C}{\partial \mathbf{z}^C} \text{ and } (\frac{\partial f}{\partial \mathbf{z}^C})^C = \frac{\partial f^C}{\partial \mathbf{z}}.$$

**Proposition 4** (Linearity). If $f, g$ are differentiable in the real sense and $\alpha, \beta \in \mathbb{C}$, then

$$\frac{\partial(\alpha f + \beta g)}{\partial \mathbf{z}} = \alpha \frac{\partial f}{\partial \mathbf{z}} + \beta \frac{\partial g}{\partial \mathbf{z}},$$
$$\frac{\partial(\alpha f + \beta g)}{\partial \mathbf{z}^C} = \alpha \frac{\partial f}{\partial \mathbf{z}^C} + \beta \frac{\partial g}{\partial \mathbf{z}^C}.$$

**Proposition 5** (Product Rule). If $f, g$ are differentiable in the real sense, then

$$\frac{\partial(f \cdot g)}{\partial \mathbf{z}} = \frac{\partial f}{\partial \mathbf{z}} g + \frac{\partial g}{\partial \mathbf{z}} f,$$
$$\frac{\partial(f \cdot g)}{\partial \mathbf{z}^C} = \frac{\partial f}{\partial \mathbf{z}^C} g + \frac{\partial g}{\partial \mathbf{z}^C} f.$$

**Proposition 6** (Chain Rule). If $f, g$ are differentiable in the real sense, then

$$\frac{\partial(f \circ g)}{\partial \mathbf{z}} = \frac{\partial f}{\partial \mathbf{z}}(g) \frac{\partial g}{\partial \mathbf{z}} + \frac{\partial f}{\partial \mathbf{z}^C}(f) \frac{\partial g^C}{\partial \mathbf{z}},$$
$$\frac{\partial(f \circ g)}{\partial \mathbf{z}^C} = \frac{\partial f}{\partial \mathbf{z}}(g) \frac{\partial g}{\partial \mathbf{z}^C} + \frac{\partial f}{\partial \mathbf{z}^C}(f) \frac{\partial g^C}{\partial \mathbf{z}^C}.$$

### A.2 Omitted proofs

#### A.2.1 Proof of Lemma 2

We first prove that it is a real value. By linearity it is sufficient to show

$$(\mathbf{h}^*, \mathbf{h}^T) \cdot \widetilde{\nabla}^2 \mathcal{G}_i(\mathbf{W}) \cdot \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^C \end{pmatrix} \in \mathbb{R}.$$

Let $\mathbf{h} = \text{vec}(\mathbf{U})$ be an arbitrary direction. Since

$$(\nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}))^C = \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W})$$

and

$$(\nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}))^C = \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}),$$

we have

$$(\text{vec}(\mathbf{U})^T \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U})^C)^C = \text{vec}(\mathbf{U})^* \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U})$$

and

$$(\text{vec}(\mathbf{U})^T \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}))^C = \text{vec}(\mathbf{U})^* \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}^C).$$

Thus,

$$\begin{aligned}
&(\text{vec}(\mathbf{U})^*, \text{vec}(\mathbf{U})^T) \begin{pmatrix} \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) & \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \\ \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) & \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \end{pmatrix} \begin{pmatrix} \text{vec}(\mathbf{U}) \\ \text{vec}(\mathbf{U})^C \end{pmatrix} \\
&= (\text{vec}(\mathbf{U})^T \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}) + \text{vec}(\mathbf{U})^T \nabla_{\mathbf{W}^C} \nabla_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}^C) \\
&+ \text{vec}(\mathbf{U})^* \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}) + \text{vec}(\mathbf{U})^* \nabla^2_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}^C)) \\
&= 2\mathcal{R}(\text{vec}(\mathbf{U})^T \nabla^2_{\mathbf{W}} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U}) + \text{vec}(\mathbf{U})^* \nabla_{\mathbf{W}} \nabla_{\mathbf{W}^C} \mathcal{G}_i(\mathbf{W}) \text{vec}(\mathbf{U})) \in \mathbb{R}
\end{aligned}$$

Now suppose $\widetilde{\mathbf{W}}$ is a local minimum, the second order expansion at $\widetilde{\mathbf{W}}$ is

$$\mathcal{L}(\widetilde{\mathbf{W}}+\mathbf{U}) = \mathcal{L}(\widetilde{\mathbf{W}}) + \widetilde{\nabla}\mathcal{L}(\widetilde{\mathbf{W}}) \cdot \begin{pmatrix}\mathbf{h}\\ \mathbf{h}\end{pmatrix} + \frac{1}{2}(\mathbf{h},\mathbf{h}^*) \cdot \widetilde{\nabla}^2\mathcal{L}(\widetilde{\mathbf{W}}) \cdot \begin{pmatrix}\mathbf{h}\\ \mathbf{h}^*\end{pmatrix} + o(||\mathbf{h}||^2).$$

Since $\widetilde{\mathbf{W}}$ is a local minimum, the gradient is zero. When $||\mathbf{h}||$ is small enough,

$$\frac{1}{2}(\mathbf{h},\mathbf{h}^*) \cdot \widetilde{\nabla}^2\mathcal{L}(\widetilde{\mathbf{W}}) \cdot \begin{pmatrix}\mathbf{h}\\ \mathbf{h}^*\end{pmatrix} = \mathcal{L}(\widetilde{\mathbf{W}}+\mathbf{U}) - \mathcal{L}(\widetilde{\mathbf{W}}) \geq 0.$$

□

### A.2.2 Proof of Lemma 3

Let $\mathbf{M} = \mathbf{W}^T \text{diag}(\mathbf{v})\mathbf{W}$. Then loss function becomes $\mathcal{L}(\mathbf{M}) = \frac{1}{2n}\sum_{i=1}^n \parallel \mathbf{x}_i^T\mathbf{M}\mathbf{x}_i - y_i \parallel^2$. By some algebra, we write

$$\mathcal{L}(\mathbf{M}) = \frac{1}{2n}\sum_{i=1}^n (\mathbf{x}_i^*\mathbf{M}^*\mathbf{x}_i^{T*}\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i - 2\mathcal{R}(y_i^*\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i) + \parallel y_i \parallel^2).$$

Notice that $\mathbf{x}_i^*\mathbf{M}^*\mathbf{x}_i^{T*}\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i = \parallel \mathbf{x}_i^T\mathbf{M}\mathbf{x}_i \parallel^2$. From the expression we can see $\mathcal{L}(\mathbf{M})$ is convex in $\mathbf{M}$ because $\mathcal{R}(y_i^*\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i)$ and $\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i$ are linear with respect to $\mathbf{M}$.

Now by Wirtinger calculus and the convexity,

$$\widetilde{\mathbf{M}} \text{ being a global minimum of } \mathcal{L}(\mathbf{M}) \Leftrightarrow \frac{1}{2n}\sum_{i=1}^n (\mathbf{x}_i^T\widetilde{\mathbf{M}}\mathbf{x}_i - y_i)^*\mathbf{x}_i\mathbf{x}_i^T = 0.$$

The rest follows the proof of Lemma 6.1 in (Soltanolkotabi et al., 2019).

□

### A.3 Existence of spurious local minima in CVNNs with $\mathbb{C}$ReLU

We prove spurious local minima exist for shallow CVNNs with $\mathbb{C}$ReLU. We define $\mathbb{C}$ReLU first.

**Definition 5** ($\mathbb{C}$ReLU). $\forall z \in \mathbb{C}$,

$$\mathbb{C}\text{ReLU}(z) = \text{ReLU}(\mathcal{R}(z)) + i\text{ReLU}(\mathcal{I}(z))$$

where

$$\text{ReLU}(x) = \max(0, x)$$

for $x \in \mathbb{R}$. Thus $\forall \mathbf{z} \in \mathbb{C}^n$, $\mathbb{C}\text{ReLU}(\mathbf{z})$ means $(\mathbb{C}\text{ReLU}(z_1), \ldots, \mathbb{C}\text{ReLU}(z_n))$.

As in (Yun et al., 2019), we make the activation in the proof to be more general.

**Definition 6.** $\forall z \in \mathbb{C}$,

$$h(z) = \hat{h}(\mathcal{R}(z)) + i\hat{h}(\mathcal{I}(z))$$

where

$$\hat{h}(x) = \max(0, s_+x) + \min(0, s_-x)$$

for $x \in \mathbb{R}$. We let $s_+ > 0$, $s_- \geq 0$ and $s_+ \neq s_-$. Note that $\mathbb{C}$ReLU is a member of this class.

We now prove Theorem 2. The proof is built on (Yun et al., 2019). The first step is to construct a local minimum that is as good as the linear solution, which is essentially the same as in (Yun et al., 2019) with some modifications. Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ denotes the input data where each column corresponds to a single data point. Without loss of generality, we define the augmented $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$ to be $[\mathbf{X}^T\ \mathbf{1}_n]^T$ and $[\mathbf{W}\ \mathbf{b}]$, and the loss function to be $\mathcal{L}(\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2) = \frac{1}{2n} \parallel \mathbf{W}_2 h(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1) + \mathbf{b}_2\mathbf{1}_n^T - \mathbf{y} \parallel_F^2$. Note that the expression of loss function here is more general than the expression we stated in Theorem 2. Now we define a linear least square loss $\ell(\mathbf{W}) = \frac{1}{2n} \parallel \mathbf{W}\tilde{\mathbf{X}} - \mathbf{y} \parallel_F^2$ and its minimizer $\bar{\mathbf{W}} \in \mathbb{C}^{1\times(d+1)}$. By Wirtinger's derivative we have $\nabla\ell(\bar{\mathbf{W}}) = (\mathbf{W}\tilde{\mathbf{X}} - \mathbf{y})^*\tilde{\mathbf{X}}^T = 0$. Let $\bar{\mathbf{y}} = \bar{\mathbf{W}}\tilde{\mathbf{X}}$ where $\bar{\mathbf{y}}_i = \bar{\mathbf{W}}[\mathbf{x}_i^T\ 1]^T$.

Let $\eta \in \mathbb{C}$ satisfies that $\mathcal{R}(\eta) = \min\{-1, 2\min_i \mathcal{R}(\bar{y}_i)\}$ and $\mathcal{I}(\eta) = \min\{-1, 2\min_i \mathcal{I}(\bar{y}_i)\}$. Let $[\mathbf{M}]_{[d]}$ denotes the first $d$ components of $\mathbf{M}$ and $[\mathbf{M}]_{d+1}$ denotes the last component. As in (Yun et al., 2019) we define parameters

$$\hat{\mathbf{W}}_1 = \alpha \begin{bmatrix} [\bar{\mathbf{W}}]_{[d]} \\ \mathbf{0}_{(k-1)\times d} \end{bmatrix}, \hat{\mathbf{b}}_1 = \alpha \begin{bmatrix} [\bar{\mathbf{W}}]_{d+1} - \eta \\ -\eta \mathbf{1}_{d-1} \end{bmatrix}, \hat{\mathbf{W}}_2 = \begin{bmatrix} \frac{1}{\alpha s_+} & \mathbf{0}_{(k-1)\times d} \end{bmatrix}, \hat{\mathbf{b}}_2 = \eta,$$

and it can be checked that $\mathcal{L}(\hat{\mathbf{W}}_1, \hat{\mathbf{b}}_1, \hat{\mathbf{W}}_2, \hat{\mathbf{b}}_2) = \frac{1}{2n} \parallel \bar{\mathbf{y}} - \mathbf{y} \parallel_F^2 = \ell(\bar{\mathbf{W}})$. We now prove that it is a local minimum of $\mathcal{L}$. We slightly permute the parameters and prove their risk is always larger. Let the permuted parameters be $\hat{\mathbf{W}}_1 + \boldsymbol{\epsilon}_1, \hat{\mathbf{b}}_1 + \boldsymbol{\delta}_1, \hat{\mathbf{W}}_2 + \boldsymbol{\epsilon}_2$, and $\hat{\mathbf{b}}_2 + \boldsymbol{\delta}_2$. Note that

$$(\bar{\mathbf{W}}\tilde{\mathbf{X}} - \mathbf{y})^* \tilde{\mathbf{X}}^T = (\bar{\mathbf{y}} - \mathbf{y})^* [\mathbf{X}^T \ \mathbf{1}_n] = 0,$$

which means $(\bar{\mathbf{y}} - \mathbf{y})^* \mathbf{X}^T = 0$ and $(\bar{\mathbf{y}} - \mathbf{y})^* \mathbf{1}_n = 0$. We mention that for a matrix $\mathbf{M} \in \mathbb{C}^{k\times n}$ its Frobenius norm can be written as

$$\parallel \mathbf{M} \parallel_F^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} |m_{ij}|^2 = \text{trace}(\mathbf{M}^* \mathbf{M}).$$

Therefore, for small enough $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\delta}_1$ such that $(\hat{\mathbf{W}}_1 + \boldsymbol{\epsilon}_1)\mathbf{x}_i + \hat{\mathbf{b}}_1 + \boldsymbol{\delta}_1 > 0$ for all $i$,

$$\begin{aligned}
&\mathcal{L}(\hat{\mathbf{W}}_1 + \boldsymbol{\epsilon}_1, \hat{\mathbf{b}}_1 + \boldsymbol{\delta}_1, \hat{\mathbf{W}}_2 + \boldsymbol{\epsilon}_2, \hat{\mathbf{b}}_2 + \boldsymbol{\delta}_2) \\
&= \frac{1}{2n} \parallel \bar{\mathbf{y}} - \mathbf{y} + s_+(\hat{\mathbf{W}}_2\boldsymbol{\epsilon}_1 + \hat{\mathbf{W}}_1\boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_2)\mathbf{X} + s_+(\hat{\mathbf{W}}_2\boldsymbol{\delta}_1 + \boldsymbol{\epsilon}_2\hat{\mathbf{b}}_1 + \boldsymbol{\epsilon}_2\boldsymbol{\delta}_1)\mathbf{1}_n^T \parallel_F^2 \\
&= \frac{1}{2n} \parallel \bar{\mathbf{y}} - \mathbf{y} \parallel_F^2 + \frac{1}{2n} \parallel s_+(\hat{\mathbf{W}}_2\boldsymbol{\epsilon}_1 + \hat{\mathbf{W}}_1\boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_2)\mathbf{X} + s_+(\hat{\mathbf{W}}_2\boldsymbol{\delta}_1 + \boldsymbol{\epsilon}_2\hat{\mathbf{b}}_1 + \boldsymbol{\epsilon}_2\boldsymbol{\delta}_1)\mathbf{1}_n^T \parallel_F^2 \\
&\geq \mathcal{L}(\hat{\mathbf{W}}_1, \hat{\mathbf{b}}_1, \hat{\mathbf{W}}_2, \hat{\mathbf{b}}_2).
\end{aligned}$$

Since the inequality holds for any $\alpha > 0$, we showed that there are infinitely many local minima. The rest of the proof follows the step 2 of the proof of Theorem 1 in (Yun et al., 2019). While our networks are complex-valued, we assumed the dataset are real-valued in Theorem 2. Besides, since Step 2 is to construct a point strictly better than the local minimum constructed before, the rest part of the proof is exactly the same as in there. There is no need to extend their proof to the complex-valued case. For readers who are not familiar with their proof, we provide a sketch proof. They defined a set $\mathcal{J} := \{j \in [m-1] | \sum_{i\leq j}(\bar{y}_i - y_i) \neq 0, \bar{y}_j < \bar{y}_{j+1}\}$ based on the $y$ labels of the dataset. There are two cases, $\mathcal{J} \neq \emptyset$ and $\mathcal{J} = \emptyset$, whose main ideas are essentially the same.

For the case of $\mathcal{J} \neq \emptyset$, by careful choices of weights $\tilde{\mathbf{W}}_1, \tilde{\mathbf{b}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{b}}_2$, and parameters $\beta$, $\gamma$, they showed the loss of the new weights are strictly better than that of the old weights. Let $\bar{y}_i = [\tilde{\mathbf{W}}]_{[d_x]}\mathbf{x}_i + [\tilde{\mathbf{W}}]_{d_x+1}$ and $\mathcal{L}_0(\bar{\mathbf{W}})$ denotes the loss of the previous local minimum. It was shown that

$$\begin{aligned}
&\mathcal{L}(\tilde{\mathbf{W}}_1, \tilde{\mathbf{b}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{b}}_2) \\
&= \frac{1}{2} \sum_{i\leq j_0} (\bar{y}_i - \frac{s_+ - s_-}{s_+ + s_-}\gamma - y_i)^2 + \frac{1}{2} \sum_{i>j_0} (\bar{y}_i + \frac{s_+ - s_-}{s_+ + s_-}\gamma - y_i)^2 \\
&= \mathcal{L}_0(\bar{\mathbf{W}}) - 2 \left[ \sum_{i\leq j_0} (\bar{y}_i - y_i) \right] \frac{s_+ - s_-}{s_+ + s_-}\gamma + O(\gamma^2)
\end{aligned}$$

which concludes the previous weight is spurious local minimum.

Similarly, for the case of $\mathcal{J} = \emptyset$, by careful choices of weights $\tilde{\mathbf{W}}_1, \tilde{\mathbf{b}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{b}}_2$, and parameters $\alpha$, $\beta$, $\gamma$, it was shown that

$$
\begin{aligned}
&\mathcal{L}(\tilde{\mathbf{W}}_1, \tilde{\mathbf{b}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{b}}_2) \\
&= \frac{1}{2} \sum_{i \leq j_1} (\bar{y}_i - \alpha \mathbf{u}^T \mathbf{x}_i - \frac{s_+ - s_-}{s_+ + s_-}\gamma - y_i)^2 + \frac{1}{2} \sum_{i > j_1} (\bar{y}_i - \alpha \mathbf{u}^T \mathbf{x}_i + \frac{s_+ - s_-}{s_+ + s_-}\gamma - y_i)^2 \\
&= \frac{1}{2} \sum_{i=1}^{m} (\bar{y}_i - \alpha \mathbf{u}^T \mathbf{x}_i - y_i)^2 + \left[ \sum_{i > j_1} (\bar{y}_i - \alpha \mathbf{u}^T \mathbf{x}_i - y_i) - \sum_{i \leq j_1} (\bar{y}_i - \alpha \mathbf{u}^T \mathbf{x}_i - y_i) \right] \frac{s_+ - s_-}{s_+ + s_-}\gamma + O(\gamma^2) \\
&= \mathcal{L}_0(\bar{\mathbf{W}}) - \alpha \left[ \sum_{i=1}^{m} (\bar{y}_i - y_i) \mathbf{x}_i^T \right] \mathbf{u} + O(\alpha^2) + \left[ \sum_{i > j_1} (\bar{y}_i - y_i) - \sum_{i \leq j_1} (\bar{y}_i - y_i) \right] \frac{s_+ - s_-}{s_+ + s_-}\gamma + O(\alpha\gamma) + O(\gamma^2)
\end{aligned}
$$

which concludes the previous weight is spurious local minimum. □