OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Neural Networks with Complex-Valued Weights Have No Spurious Local Minima

Xingtu Liu RLTHEORY@OUTLOOK.COM
Simon Fraser University

Abstract

We study the benefits of complex-valued weights for neural networks. We prove that shallow complex neural networks with quadratic activations have no spurious local minima. In contrast, shallow real neural networks with quadratic activations have infinitely many spurious local minima under the same conditions. In addition, we provide specific examples to demonstrate that complex-valued weights turn poor local minima into saddle points.

1. Introduction

A major challenge in deep learning is to avoid gradient descent from getting trapped in suboptimal local minima. Thus, understanding the optimization landscape of neural networks has become a significant area of research. It has been shown that almost all non-linear real-valued neural networks have poor local minima, which includes neural networks with *quadratic*, *tanh*, *sigmoid*, *arctan*, ReLU, ELU, and SELU activations [21]. Even under the over-parametrization condition, real-valued neural networks have strict poor local minima [19]. Only linear neural networks have been proven to have no spurious local minima [10], however, they are rarely used in practice because of their limited expressive power. Universal approximation theorem only holds for *non-linear* neural networks.

In this paper, we prove a simple but surprising result:

> *All local minima in shallow complex neural networks with quadratic activations are global minima.*

This result is unanticipated for the following reasons.

- Under the exact same conditions (i.e., shallow neural networks with quadratic activations), real-valued networks have infinitely many spurious local minima.
- No unrealistic assumptions are made, and not even mild over-parameterization (i.e., $k > n$, where k is the number of hidden nodes and n is the number of samples) is required .

We only require that $k \geq d$, where d is the dimension of each sample. This implies that the number of hidden nodes is at least equal to the *dimension of each sample*, which is a common setting. In contrast, the assumption of mild over-parametrization requires the number of hidden nodes to be larger than the *number of input samples*, i.e. $k \geq n$. Moreover, the standard over-parametrization setting, for example in neural tangent kernel (NTK), the assumption of $k \geq O(n^4)$ was made.

In addition, we provide some specific examples of poor local minima in the real-valued case and demonstrate how they become saddle points in the complex-valued neural networks (CVNNs).

Experimental results have shown that complex-valued weights are more likely to escape poor local minima in neural networks with polynomial activations [1, 5]. However, this work is the first to theoretically demonstrate that the loss surface of nonlinear complex-valued neural networks is indeed superior to that of their real-valued counterparts. The minimum modulus principle in complex analysis tells us complex analytic functions have superior landscapes. Therefore, we conjecture that the "no spurious local minima" result also holds true for deeper complex networks and complex networks with other analytic activations such as *tanh* and *sigmoid*. Along the way, we develop a novel set of tools and techniques for analyzing the optimization landscape of CVNNs, which may be useful in other contexts.

1.1. Related Work

The analysis of optimization landscape of neural networks began with linear neural networks. Baldi et al. [2, 3] proved the "no spurious local minima" result for both real-valued and complex-valued shallow linear neural networks. This was later translated to deep linear neural networks by Kawaguchi [10]. Work on the optimization of linear networks from other perspectives can be found elsewhere [9, 12, 14, 22]. Although optimizing linear networks being a non-convex problem have a nice landscape, the representation power of linear network is limited as they can only fit linearly separable data.

For non-linear networks, all demonstrations of "no spurious local minima" required unrealistic assumptions or an extremely restricted parameter space. For instance, independence between weights was assumed for deep ReLU networks [10]. For shallow ReLU networks, the result in [20] only holds for two hidden units networks and weight vectors must to be unit-normed and orthogonal. For shallow quadratic networks, [8] assumed Gaussian feature vectors; while [18] assumed the weight vector connecting the hidden layer and the output node must contain at least d positive entries and d negative entries. Without these unrealistic assumptions, all common non-linear real-valued networks were shown to have poor local minima [21].

The optimization landscape of neural networks under over-parametrization has also been studied [13, 15, 16]. Unfortunately, it has been shown that neural networks with over-parametrization can still have strict poor local minima [17, 19], and utilizing the set of measure zero only eliminated non-strict poor local minima, i.e. sub-optimal basins.

Some attempts from the algorithm dynamics have been made, showing that gradient descent can converge to global minima under certain conditions [6, 7]. However, the networks in the analysis are ultra-wide.

2. Preliminaries

This section provides an introduction to complex analysis and Wirtinger calculus. More definitions and lemmas are provided in the appendix.

2.1. Notations

Let $\mathbb{R}$ and $\mathbb{C}$ denote the real and complex fields respectively. They share many common properties as number fields. Note 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 the transpose and conjugate transpose of $\mathbf{M}$, $\mathbf{M}^C$ denotes the matrix whose entries are conjugates of entries in $\mathbf{M}$, $\text{Null}(\mathbf{M}) := \{\mathbf{v} \mid \mathbf{Mv} = 0\}$ denotes the null space, and $\text{vec}(\mathbf{M})$ 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 most prominent function used in this work is a real-valued function with complex-valued matrix input $f : \mathbb{C}^{m\times n} \mapsto \mathbb{R}$.

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

Recall that the loss function

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

is not analytic. However, $y_i - \mathbf{v}^T \psi(\mathbf{W}\mathbf{x}_i)$ is analytic for each i, which indicates that 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 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, provides a neat way of deriving the derivatives. For a differentiable complex function, Wirtinger derivatives are the same as the traditional derivatives. Using Wirtinger calculus provides not only defined derivatives that reflect a function's gradient, but also have meaningful results for the first and second derivatives. Note that critical points, positive semi-definite Hessian, and Taylor expansions all have their counterparts in Wirtinger calculus. Due to space constraints, an extended introduction of Wirtinger calculus can be found in Appendix. A comprehensive introduction can be found in [11] and [4].

3. Main Results

We analyze shallow neural networks with quadratic activation functions over the fields $\mathbb{R}$ and $\mathbb{C}$, with the network structure and loss function defined as below.

Definition 1 *Define the dataset $\{(\boldsymbol{x}_i, y_i)\}$ for $i = 1, 2, \ldots, n$ with $\boldsymbol{x}_i \in \mathbb{F}^d$ and $y_i \in \mathbb{F}$. Consider one hidden layer complex-valued neural networks with the quadratic activation ψ in the form of*

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

where $\boldsymbol{W} \in \mathbb{F}^{k\times d}, \boldsymbol{v} \in \mathbb{F}^k$, $v_i \neq 0$, *and* $k \geq d$. *The training loss as a function of the weights* ($\boldsymbol{W}$, $\boldsymbol{v}$) *is defined as*

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

Noted that $\|y_i - \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i)\|^2 = (y_i - \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i))^2$ when $\mathbb{F} = \mathbb{R}$. We prove that one hidden layer CVNNs with quadratic activations have no spurious local minima.

Theorem 2 *Assume the dataset, loss function, and training model are defined as in Definition 1 with* $\mathbb{F} = \mathbb{C}$. *Then the training loss as a function of the weight* ($\mathbf{W}$, $\mathbf{v}$) *has no spurious local minima, i.e. all local minima are global.*

Proof See Appendix B. ■

Lemma 3 shows that, under the same setting, Theorem 2 does not hold when $\mathbb{F} = \mathbb{R}$, i.e. in real-valued networks poor local minima exist. This provides evidence for a substantial difference between real-valued and complex-valued networks.

Lemma 3 (Theorem 2, Corollary 3 in [21]) *Let the loss function* $\mathcal{L}(\mathbf{W}, \mathbf{v})$ *and the network structure be defined as in Definition 1. Consider the dataset*

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & \frac{1}{2}\\ 0 & 1 & \frac{1}{2}\end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix}y_1 & y_2 & y_3\end{bmatrix} = \begin{bmatrix}0 & 0 & 1\end{bmatrix},$$

and the weight

$$\bar{\mathbf{W}} = \begin{bmatrix}\bar{\boldsymbol{w}}_1\\ \bar{\boldsymbol{w}}_2\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \bar{\mathbf{v}} = \begin{bmatrix}\bar{v}_1 & \bar{v}_2\end{bmatrix} = \begin{bmatrix}\frac{1}{6} & \frac{1}{6}\end{bmatrix}.$$

When $\mathbb{F} = \mathbb{R}$ *the weight is a poor local minimum of the loss function.*

4. Complex-Valued Weights Turn Local Minima into Saddle Points

In this section, we provide a concrete example of how complex-valued weights turn local minima into saddle points. Let the weight

$$(\bar{\mathbf{W}}, \bar{\mathbf{v}}) = \left(\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \begin{bmatrix}\frac{1}{6} & \frac{1}{6}\end{bmatrix}\right)$$

be a local minimum of the real-valued network for the given dataset as defined in the previous section. Now we analyze $\left(\bar{\mathbf{W}}, \bar{\mathbf{v}}\right)$ in two different networks and in two ways. On one hand, we show that the Hessians at the point are different in real networks and in complex networks. On the other hand, we prove that there is a point with strictly less loss in an arbitrarily small neighborhood of $(\bar{\mathbf{W}}, \bar{\mathbf{v}})$ in the complex network.

Lemma 4 *Assume the dataset, loss function, and training model are defined as above. At* $\bar{\mathbf{W}}$, *the Hessian has no negative eigenvalue when* $\mathbb{F} = \mathbb{R}$ *and has both positive and negative eigenvalues when* $\mathbb{F} = \mathbb{C}$.

Proof See Appendix C. ■

By analyzing a specific critical point, Lemma 4 shows us that complex-valued weights can turn poor local minima into saddle points. As an alternative way to illustrate the insight, we provide Lemma 5. The proofs are provided in the appendix.

Lemma 5 *Assume the dataset, loss function, and training model are defined as above. When* $\mathbb{F} = \mathbb{C}$*, within an arbitrarily small neighborhood of* $(\bar{\mathbf{W}}, \bar{\mathbf{v}})$ *there is a point* $(\hat{\mathbf{W}}, \hat{\mathbf{v}})$ *such that* $\mathcal{L}(\hat{\mathbf{W}}, \hat{\mathbf{v}}) < \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{v}})$.

Proof See Appendix D. ■

Now, we provide proof sketches for Lemma 4 and Lemma 5. In the proof of Lemma 4, we derive the Hessian matrices at $\bar{\mathbf{W}}$ for real-valued and complex-valued networks and calculate their eigenvalues. We denote the Hessian matrix by $\mathcal{H}^{\mathbb{R}}_{\bar{\mathbf{W}}}$ for the real-valued case and $\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}}$ for the complex-valued case. Note that $\mathbb{C}^n$ can be treated as $\mathbb{R}^{2n}$ in some sense. Therefore $\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}}$ is four times the size of $\mathcal{H}^{\mathbb{R}}_{\bar{\mathbf{W}}}$. By some derivative calculations and substituting the weight and dataset given in Lemma 3 into the expression, we have

$$\mathcal{H}^{\mathbb{R}}_{\bar{\mathbf{W}}} = \begin{bmatrix} \frac{7}{108} & \frac{-1}{108} & \frac{5}{108} & \frac{1}{108} \\ \frac{-1}{108} & \frac{7}{108} & \frac{1}{108} & \frac{5}{108} \\ \frac{5}{108} & \frac{1}{108} & \frac{7}{108} & \frac{-1}{108} \\ \frac{1}{108} & \frac{5}{108} & \frac{-1}{108} & \frac{7}{108} \end{bmatrix}$$

which has no negative eigenvalues, and

$$\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}} = \begin{bmatrix} \mathcal{H}_1 & \mathcal{H}_1 & \mathcal{H}_2 & \mathcal{H}_3 \\ \mathcal{H}_1 & \mathcal{H}_1 & \mathcal{H}_3 & \mathcal{H}_2 \\ \mathcal{H}_2 & \mathcal{H}_3 & \mathcal{H}_1 & \mathcal{H}_1 \\ \mathcal{H}_3 & \mathcal{H}_2 & \mathcal{H}_1 & \mathcal{H}_1 \end{bmatrix}$$

where

$$\mathcal{H}_1 = \begin{bmatrix} \frac{5}{216} & \frac{1}{216} \\ \frac{1}{216} & \frac{5}{216} \end{bmatrix}, \mathcal{H}_2 = \begin{bmatrix} \frac{1}{108} & -\frac{1}{108} \\ -\frac{1}{108} & \frac{1}{108} \end{bmatrix}$$

and $\mathcal{H}_3$ is a zero matrix. It can be verified that $\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}}$ has both negative and positive eigenvalues. This illustrates that complex-valued weights turn local minima into saddle points. As an alternative proof, in the proof of Lemma 5, we make the following permutation on $\bar{\mathbf{W}}$,

$$\hat{\mathbf{W}} = \begin{bmatrix} 1 - \frac{1}{10^N} & 1 + \frac{i}{10^N} \\ 1 - \frac{1}{10^N} & 1 + \frac{i}{10^N} \end{bmatrix}$$

for an arbitrarily large $N \in \mathbb{N}^+$, and by simple calculations we can prove that $\mathcal{L}(\hat{\mathbf{W}}, \bar{\mathbf{v}}) < \frac{1}{9} = \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{v}})$.

References

[1] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1908–1916, 2014.

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

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

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

[5] Sebastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks, 2020.

[6] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019.

[7] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2019.

[8] 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*, pages 9111–9121. 2019.

[9] Moritz Hardt and Tengyu Ma. Identity matters in deep learning, 2018.

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

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

[12] Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2902–2907, 2018.

[13] Dawei Li, Tian Ding, and Ruoyu Sun. On the benefit of width for neural networks: Disappearance of bad basins, 2021.

[14] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima, 2017.

[15] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2603–2612, 2017.

[16] Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2019.

[17] Arsalan Sharifnassab, Saber Salehkaleybar, and S. Jamaloddin Golestani. Bounds on over-parameterization for guaranteed existence of descent paths in shallow relu networks. In *International Conference on Learning Representations*, 2020.

[18] 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), 2019.

[19] Ruoyu Sun, Dawei Li, Shiyu Liang, Tian Ding, and R Srikant. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.

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

[21] 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.

[22] Li Zhang. Depth creates no more spurious local minima, 2020.

Appendix A. Supporting Lemmas

Lemma 6 *If $\widetilde{\mathbf{W}}$ is a local minimum of $\mathcal{L}(\widetilde{\mathbf{W}})$,*

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

where

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

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

Proof 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. As in the standard proof, when $||\mathbf{h}||$ is small enough,

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

■

Lemma 7 *Any point* $\widetilde{\mathbf{W}} \in \mathbb{C}^{k\times d}$ *obeying*

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

is a global optimum of the loss function

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

Proof 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,

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

Note that $\widetilde{\mathbf{M}} = \widetilde{\mathbf{W}}^T\text{diag}(\mathbf{v})\widetilde{\mathbf{W}}$ for some $\widetilde{\mathbf{W}}$. Thus, for any arbitrary $\mathbf{M} \in \mathbb{C}^{d\times d}$ we have

$$\mathcal{L}(\mathbf{M}) \geq \mathcal{L}(\widetilde{\mathbf{M}})$$

which implies that for any $\mathbf{W} \in \mathbb{C}^{k\times d}$

$$\mathcal{L}(\mathbf{W}) \geq \mathcal{L}(\widetilde{\mathbf{W}}).$$

■

Appendix B. Proof of Theorem 2

Step 1: Derivative calculations. First, we demonstrate how to calculate the derivative of $\mathcal{L}(\mathbf{W})$. First, 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 Cauchy-Riemann Equation. By letting $\mathcal{L}(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i(\mathbf{W})^*\mathcal{L}_i(\mathbf{W})$ where $\mathcal{L}_i(\mathbf{W}) : \mathbb{C}^{k\times d} \mapsto \mathbb{C}$ given by $\mathcal{L}_i(\mathbf{W}) = \mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i) - y_i$, $\mathcal{L}_i(\mathbf{W})$ is complex differentiable in the traditional sense. Thus, $\mathcal{L}_i(\mathbf{W})$ has well-defined first and second derivatives. For a fixed i, we let $\mathcal{G}_i(\mathbf{W}) = \mathcal{L}_i(\mathbf{W})^*\mathcal{L}_i(\mathbf{W})$. Now, we 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. Denote $\nabla_{\mathbf{w}_q}\mathcal{L}_i(\mathbf{W})$ the first derivative with respect to the q-th row of $\mathbf{W}$, and we derive

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

where $\mathbf{D}_{\mathbf{v}} = \text{diag}(v_1, \ldots, v_k)$. The second derivative can be expressed as

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

and

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

for $p \neq q$. Next, we derive the derivative of $\mathcal{G}_i(\mathbf{W})$. By the product rule of Wirtinger calculus, we have

$$\nabla_{\mathbf{W}}\mathcal{G}_i(\mathbf{W}) = \nabla_{\mathbf{W}}(\mathcal{L}_i^*\mathcal{L}_i)(\mathbf{W}) = \nabla_{\mathbf{W}}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) + \nabla_{\mathbf{W}}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W}) = \nabla_{\mathbf{W}}\mathcal{L}_i(\mathbf{W})\mathcal{L}_i^*(\mathbf{W}).$$

Note that $\nabla_{\mathbf{W}}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) = 0$ 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 the above equalities, we obtain the second derivatives

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

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.

Step 2: W being a local minimum implies (W, v) being a local minimum. We have both weights optimized, i.e. under the condition of both $\mathbf{W}$ and $\mathbf{v}$ being local minima. However, it is sufficient to utilize only the fact that $\mathbf{W}$ is a local minimum. In this proof $\mathbf{v}$ can be any vectors with non-zero entries including the local minima. A key observation is that given $v_i \neq 0$, a global minimum of $\mathcal{L}$ w.r.t $\mathbf{W}$ is a global minimum w.r.t. ($\mathbf{W}$, $\mathbf{v}$). The reason is that for any current v_i and the "targeted" v_i^*, $\mathcal{L}$ has the same loss with either $(\mathbf{w}_i, v_i^*)$ or $\left(\sqrt{\frac{v_i^*}{v_i}}\cdot\mathbf{w}_i, v_i\right)$ on the i^{th} row. Consider a toy example, $\psi(\sqrt{2}i\cdot[1,2])\cdot 3 = \psi([1,2])\cdot -6$, where $v_i = 3$, $v_i^* = -6$, and $\sqrt{\frac{v_i^*}{v_i}} = \sqrt{2}i$. Therefore, proving any local minimum of $\mathcal{L}$ w.r.t $\mathbf{W}$ is a global minimum is sufficient to prove Theorem 2.

Step 3: Simplifying $\mathcal{H}$. 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 7. 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

$$\nabla_{\mathbf{w}_p^C}\mathcal{G}_i(\mathbf{W}) = \nabla_{\mathbf{w}_p^C}\mathcal{L}_i^*(\mathbf{W})\mathcal{L}_i(\mathbf{W}) = v_p^*\psi'(\langle\mathbf{w}_p, \mathbf{x}_i\rangle)^*\mathbf{x}_i^C\mathcal{L}_i(\mathbf{W})$$

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

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

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

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

where

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

Now by linearity and Lemma 6 we have

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

where

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

We argue that we can assume $(\mathbf{a}^T\mathbf{D}_{\mathbf{v}}\mathbf{a}) \neq 0$ here. The reason is the following. Since $\mathbf{a} \neq \mathbf{0}$, there is an entry $a_i \neq 0$. Suppose $\mathbf{a}^T\mathbf{D}_{\mathbf{v}}\mathbf{a} = 0$, we can multiply v_i by $\frac{1}{4}$ and multiply the the i'th row of $\mathbf{W}$ by 2. Now $\mathbf{a}^T\mathbf{D}_{\mathbf{v}_{new}}\mathbf{a} = -\frac{3}{4}v_i a_i^2 \neq 0$. Note that the two weight matrices $\mathbf{W}$ and $\mathbf{W}_{new}$ have the same null space and $\mathbf{W}_{new}$ is also a local minimum. By Lemma 7 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 $\mathcal{L}$ is equivalent to proving $\mathbf{W}$ is a global minimum. Thus, we can assume $(\mathbf{a}^T\mathbf{D}_{\mathbf{v}}\mathbf{a}) \neq 0$ without loss of generality.

Step 4: Proving $\sum_{i=1}^{n} \mathcal{L}_i^*(\mathbf{W})\mathbf{x}_i\mathbf{x}_i^T = 0$. Let $2(\mathbf{a}^T\mathbf{D}_{\mathbf{v}}\mathbf{a}) = a_1 + ia_2 \in \mathbb{C}$ 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 $\mathcal{H} = \frac{1}{2n}\mathcal{R}((a_1 + a_2 i) \cdot (b_1 + b_2 i)) = \frac{1}{2n}(a_1b_1 - a_2b_2) \geq 0$, 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) with $b_1 < 0$ 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 $\mathcal{M}$. Now, we show 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 β_i can be any complex number. 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. Thus, 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). Therefore, $\mathbf{W}$ is a global minimum by Lemma 7, concluding the proof.

Appendix C. Proof of Lemma 4

Recall that

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

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \end{bmatrix},$$
$$\boldsymbol{Y} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix},$$
$$\bar{\mathbf{W}} = \begin{bmatrix} \bar{\boldsymbol{w}}_1 \\ \bar{\boldsymbol{w}}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$
$$\bar{\mathbf{v}} = \begin{bmatrix} \bar{v}_1 & \bar{v}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} \end{bmatrix}.$$

It is to be noticed that the Hessian at $\bar{\mathbf{W}}$ is

$$\mathcal{H}^{\mathbb{R}}_{\bar{\mathbf{W}}} = \begin{bmatrix} \frac{\partial^2}{\partial \bar{\mathbf{w}}_1^2}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \frac{\partial^2}{\partial \bar{\mathbf{w}}_1 \bar{\mathbf{w}}_2}\mathcal{L}(\mathbf{W}, \mathbf{v}) \\ \frac{\partial^2}{\partial \bar{\mathbf{w}}_2 \bar{\mathbf{w}}_1}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \frac{\partial^2}{\partial \bar{\mathbf{w}}_2^2}\mathcal{L}(\mathbf{W}, \mathbf{v}) \end{bmatrix}$$

where

$$\begin{aligned} \frac{\partial^2}{\partial \bar{\mathbf{w}}_1^2}\mathcal{L}(\mathbf{W}, \mathbf{v}) &= \frac{\bar{v}_1}{n}\sum_{i=1}^{n}(\bar{\mathbf{v}}^T\psi(\bar{\mathbf{W}}\mathbf{x}_i) - y_i)\psi''(\bar{\mathbf{w}}_1\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{\bar{v}_1^2}{n}\sum_{i=1}^{n}(\psi'(\bar{\mathbf{w}}_1\mathbf{x}_i))^2\mathbf{x}_i\mathbf{x}_i^T \\ &= \begin{bmatrix} \frac{7}{108} & \frac{-1}{108} \\ \frac{-1}{108} & \frac{7}{108} \end{bmatrix} \\ &= \frac{\bar{v}_2}{n}\sum_{i=1}^{n}(\bar{\mathbf{v}}^T\psi(\bar{\mathbf{W}}\mathbf{x}_i) - y_i)\psi''(\bar{\mathbf{w}}_2\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{\bar{v}_2^2}{n}\sum_{i=1}^{n}(\psi'(\bar{\mathbf{w}}_2\mathbf{x}_i))^2\mathbf{x}_i\mathbf{x}_i^T \\ &= \frac{\partial^2}{\partial \bar{\mathbf{w}}_2^2}\mathcal{L}(\mathbf{W}, \mathbf{v}), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2}{\partial \bar{\mathbf{w}}_1 \bar{\mathbf{w}}_2}\mathcal{L}(\mathbf{W}, \mathbf{v}) &= \frac{\bar{v}_1\bar{v}_2}{n}\sum_{i=1}^{n}\psi'(\bar{\mathbf{w}}_1\mathbf{x}_i)\psi'(\bar{\mathbf{w}}_2\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &= \begin{bmatrix} \frac{5}{108} & \frac{1}{108} \\ \frac{1}{108} & \frac{5}{108} \end{bmatrix} \\ &= \frac{\bar{v}_2\bar{v}_1}{n}\sum_{i=1}^{n}\psi'(\bar{\mathbf{w}}_2\mathbf{x}_i)\psi'(\bar{\mathbf{w}}_1\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &= \frac{\partial^2}{\partial \bar{\mathbf{w}}_2 \bar{\mathbf{w}}_1}\mathcal{L}(\mathbf{W}, \mathbf{v}). \end{aligned}$$

It can be verified easily that $\mathcal{H}^{\mathbb{R}}_{\bar{\mathbf{W}}}$ has no negative eigenvalue. Now we analyze the Wirtinger Hessians at $\bar{\mathbf{W}}$, namely the Hessian at the complex-valued setting. Recall that

$$\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}} = \begin{pmatrix} \nabla_{\bar{\mathbf{W}}}\nabla_{\bar{\mathbf{W}}^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \nabla^2_{\bar{\mathbf{W}}^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) \\ \nabla^2_{\bar{\mathbf{W}}}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \nabla_{\bar{\mathbf{W}}^C}\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\mathbf{W}, \mathbf{v}) \end{pmatrix}$$

and we calculate it term by term. Firstly we notice that $\nabla_{\bar{\mathbf{W}}}\nabla_{\bar{\mathbf{W}}^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) = \nabla_{\bar{\mathbf{W}}^C}\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\mathbf{W}, \mathbf{v})$ and $\nabla^2_{\bar{\mathbf{W}}^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) = \nabla^2_{\bar{\mathbf{W}}}\mathcal{L}(\mathbf{W}, \mathbf{v})$ because the weights and data are real-valued. Now we have

$$\nabla_{\bar{\mathbf{W}}}\nabla_{\bar{\mathbf{W}}^C}\mathcal{L}(\mathbf{W}) = \begin{bmatrix} \nabla_{\bar{\mathbf{w}}_1}\nabla_{\bar{\mathbf{w}}_1^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \nabla_{\bar{\mathbf{w}}_1}\nabla_{\bar{\mathbf{w}}_2^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) \\ \nabla_{\bar{\mathbf{w}}_2}\nabla_{\bar{\mathbf{w}}_1^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) & \nabla_{\bar{\mathbf{w}}_2}\nabla_{\bar{\mathbf{w}}_2^C}\mathcal{L}(\mathbf{W}, \mathbf{v}) \end{bmatrix}$$

and we observe that

$$\begin{aligned}&\nabla_{\bar{\mathbf{w}}_1}\nabla_{\bar{\mathbf{w}}_1^C}\mathcal{L}(\mathbf{W},\mathbf{v})=\nabla_{\bar{\mathbf{w}}_1}\nabla_{\bar{\mathbf{w}}_2^C}\mathcal{L}(\mathbf{W},\mathbf{v})=\nabla_{\bar{\mathbf{w}}_2}\nabla_{\bar{\mathbf{w}}_1^C}\mathcal{L}(\mathbf{W},\mathbf{v})=\nabla_{\bar{\mathbf{w}}_2}\nabla_{\bar{\mathbf{w}}_2^C}\mathcal{L}(\mathbf{W},\mathbf{v})\\&=\frac{1}{2n}\sum_{i=1}^{n}v_1^*v_1\psi^{'}(\langle\mathbf{w}_1,\mathbf{x}_i\rangle)^*\psi^{'}(\langle\mathbf{w}_1,\mathbf{x}_i\rangle)\mathbf{x}_i^C\mathbf{x}_i^T\\&=\begin{bmatrix}\frac{5}{216}&\frac{1}{216}\\\frac{1}{216}&\frac{5}{216}\end{bmatrix}.\end{aligned}$$

The other two are slightly different under Wirtinger calculus. By Wirtinger calculus, we have Wirtinger Hessian

$$\nabla^2_{\bar{\mathbf{W}}}\mathcal{L}(\mathbf{W})=\begin{bmatrix}\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_1^2}\mathcal{L}_i(\mathbf{W},\mathbf{v})&\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_1\bar{\mathbf{w}}_2}\mathcal{L}_i(\mathbf{W},\mathbf{v})\\\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_2\bar{\mathbf{w}}_1}\mathcal{L}_i(\mathbf{W},\mathbf{v})&\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_2^2}\mathcal{L}_i(\mathbf{W},\mathbf{v})\end{bmatrix}.$$

We also have

$$\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_1\bar{\mathbf{w}}_2}\mathcal{L}_i(\mathbf{W},\mathbf{v})=\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_2\bar{\mathbf{w}}_1}\mathcal{L}_i(\mathbf{W},\mathbf{v})=0$$

and

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_1^2}\mathcal{L}_i(\mathbf{W},\mathbf{v})\\&=\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{v}^T\psi(\mathbf{W}\mathbf{x}_i)-y_i)v_1\psi^{''}(\langle\mathbf{w}_1,\mathbf{x}_i\rangle)\mathbf{x}_i\mathbf{x}_i^T\\&=\begin{bmatrix}\frac{1}{108}&-\frac{1}{108}\\-\frac{1}{108}&\frac{1}{108}\end{bmatrix}\\&=\frac{1}{2n}\sum_{i=1}^{n}\mathcal{L}_i^*(\mathbf{W},\mathbf{v})\frac{\partial^2}{\partial\bar{\mathbf{w}}_2^2}\mathcal{L}_i(\mathbf{W},\mathbf{v}).\end{aligned}$$

It can be verified that $\mathcal{H}^{\mathbb{C}}_{\bar{\mathbf{W}}}$ has 1 negative eigenvalues.

Appendix D. Proof of Lemma 5

To select $(\hat{\mathbf{W}},\hat{\mathbf{v}})$ in an arbitrarily small neightbor of $(\bar{\mathbf{W}},\bar{\mathbf{v}})$, we permute $\bar{\mathbf{W}}$ only and let $\hat{\mathbf{v}}=\bar{\mathbf{v}}$. For an arbitrarily large $N\in\mathbb{N}^+$, we let

$$\hat{\mathbf{W}}=\begin{bmatrix}1-\frac{1}{10^N}&1+\frac{i}{10^N}\\1-\frac{1}{10^N}&1+\frac{i}{10^N}\end{bmatrix}.$$

Firstly, we notice that $\mathcal{L}(\bar{\mathbf{W}},\bar{\mathbf{v}})=\frac{1}{9}$. Therefore, it is enough to show that $\mathcal{L}(\hat{\mathbf{W}},\hat{\mathbf{v}})<\frac{1}{9}$ for an arbitrarily large N. By simple calculations, we have

$$\hat{\mathbf{W}}\mathbf{X}=\begin{bmatrix}1-\frac{1}{10^N}&1+\frac{i}{10^N}&1-\frac{1}{2\cdot10^N}+\frac{i}{2\cdot10^N}\\1-\frac{1}{10^N}&1+\frac{i}{10^N}&1-\frac{1}{2\cdot10^N}+\frac{i}{2\cdot10^N}\end{bmatrix},$$

$$\psi(\hat{\mathbf{W}}\mathbf{X}) = \begin{bmatrix} \frac{(10^N-1)^2}{10^{2N}} & \frac{(10^N+i)^2}{10^{2N}} & \frac{(2\cdot 10^N-1+i)^2}{4\cdot 10^{2N}} \\ \frac{(10^N-1)^2}{10^{2N}} & \frac{(10^N+i)^2}{10^{2N}} & \frac{(2\cdot 10^N-1+i)^2}{4\cdot 10^{2N}} \end{bmatrix},$$

$$\hat{\mathbf{v}}\psi(\hat{\mathbf{W}}\mathbf{X}) = \begin{bmatrix} \frac{(10^N-1)^2}{3\cdot 10^{2N}} & \frac{(10^N+i)^2}{3\cdot 10^{2N}} & \frac{(2\cdot 10^N-1+i)^2}{12\cdot 10^{2N}} \end{bmatrix}.$$

Therefore,

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{W}}, \hat{\mathbf{v}}) &= \frac{1}{6} \cdot \left(\frac{(10^N-1)^4}{9\cdot 10^{4N}} + \left\| \frac{(10^N+i)^2}{3\cdot 10^{2N}} \right\|^2 + \left\| 1 - \frac{(2\cdot 10^N - 1 + i)^2}{12\cdot 10^{2N}} \right\|^2 \right) \\
&= \frac{1}{6} \cdot \left(\frac{(10^N-1)^4}{9\cdot 10^{4N}} + \frac{(10^{2N}-1)^2}{9\cdot 10^{4N}} + \frac{4\cdot 10^{2N}}{9\cdot 10^{4N}} + \frac{(8\cdot 10^{2N} - 4\cdot 10^N)^2}{144\cdot 10^{4N}} + \frac{4\cdot(2\cdot 10^N-1)^2}{144\cdot 10^{4N}} \right) \\
&= \frac{1}{6} \cdot \left(\frac{96\cdot 10^{4N} - 128\cdot 10^{3N} + 160\cdot 10^{2N} - 80\cdot 10^N + 36}{144\cdot 10^{4N}} \right) \\
&< \frac{1}{9}
\end{aligned}$$

for all $N \in \mathbb{N}^+$.

Appendix E. Additional Preliminaries (Supplement to Section 2)

E.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 8 *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 9 (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 10 (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 11 (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 12 (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 recall some properties of complex numbers. For all $z, y \in \mathbb{C}$, we have $|z^*| = |z|, zz^* = |z|^2, z^{-1} = \frac{z^*}{|z|^2}$ if $z \neq 0, \mathcal{R}(z) = \frac{z+z^*}{2}, \mathcal{I}(z) = \frac{z-z^*}{2i}, |zy| = |z||y|$, and $|z^n| = |z|^n$.

E.2. More on Wirtinger calculus

We provide few important exposition of Wirtinger calculus here. More explanations can be found in [11] and [4]. 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.

We now have the following lemma which follows directly from the definitions.

Lemma 13 *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.*

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

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.

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

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.

Lastly we state some important propositions.

Definition 14 (Conjugate Cauchy Riemann conditions (CCRC))

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

Proposition 15 *If f is differentiable in the real sense at $(\boldsymbol{x}, \boldsymbol{y})$ and the CCRC hold, then f is conjugate-complex differentiable.*

Proposition 16 *If f is conjugate-complex differentiable at $\boldsymbol{z}$ then u and v are differentiable in the real sense and they satisfy the conjugate Cauchy Riemann conditions.*

Proposition 17 *If f is differentiable in the real sense, then*

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

Wirtinger derivative share many properties as normal derivatives like linearity, product rule, and chain rule.

Proposition 18 (Linearity) *If f, g are differentiable in the real sense and $\alpha, \beta \in \mathbb{C}$, then*

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

Proposition 19 (Product Rule) *If f, g are differentiable in the real sense, then*

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

Proposition 20 (Chain Rule) *If f, g are differentiable in the real sense, then*

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