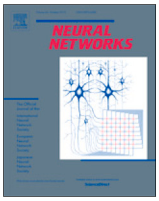
Neural Networks 151 (2022) 80–93

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Towards understanding theoretical advantages of complex-reaction networks

Shao-Qun Zhang, Wei Gao, Zhi-Hua Zhou [*]

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China

ARTICLE INFO

Article history:
Received 15 November 2021
Received in revised form 21 January 2022
Accepted 22 March 2022
Available online 29 March 2022

Keywords:
Complex-valued neural network
Complex-reaction network
Approximation
Convergence
Critical point set

ABSTRACT

Complex-valued neural networks have attracted increasing attention in recent years, while it remains open on the advantages of complex-valued neural networks in comparison with real-valued networks. This work takes one step on this direction by introducing the *complex-reaction network* with fully-connected feed-forward architecture. We prove the universal approximation property for complex-reaction networks, and show that a class of radial functions can be approximated by a complex-reaction network using the polynomial number of parameters, whereas real-valued networks need at least exponential parameters to reach the same approximation level. For empirical risk minimization, we study the landscape and convergence of complex gradient descents. Our theoretical result shows that the critical point set of complex-reaction networks is a proper subset of that of real-valued networks, which may show some insights on finding the optimal solutions more easily for complex-reaction networks.

1. Introduction

Deep neural networks have become a mainstream model of deep learning (LeCun, Bengio, & Hinton, 2015) during the past decades, mostly working with real-valued neural networks. For example, great progresses have been made for real-valued neural networks in many real applications such as computer vision (Krizhevsky, Sutskever, & Hinton, 2012), speech recognition (Graves, Mohamed, & Hinton, 2013; Sutskever, Vinyals, & Le, 2014), machine translation (Bahdanau, Cho, & Bengio, 2014), etc. Theoretical studies have also attracted much attention on the deep understanding of real-valued neural networks, including universal approximation (Barron, 1994; Funahashi, 1989; Hornik, 1991; Kidger & Lyons, 2020; Leshno, Lin, Pinkus, & Schocken, 1993; Lu, Pu, Wang, Hu, & Wang, 2017; Sun, Chen, Wang, Liu, & Liu, 2016), optimization dynamics (Allen-Zhu, Li, & Song, 2019; Dauphin et al., 2014; Du, Lee, Li, Wang, & Zhai, 2019; Poggio, Banburski, & Liao, 2020), generalization (Hardt, Recht, & Singer, 2016; Zhang, Bengio, Hardt, Recht, & Vinyals, 2017), neural tangent kernel (Arora, Du, Hu, Li, & Wang, 2019; Du, Zhai, Poczos, & Singh, 2018; Jacot, Gabriel, & Hongler, 2018), etc.

Recent years have also witnessed an increasing interest on complex-valued neural networks. Hirose and Yoshida (2012) introduced the complex-valued neural networks with amplitude–phase-type activation function, and showed better generalization than real-valued networks in the fitting interpolation of temporal signals. Tygert et al. (2016) simulated the complex-valued convolution operations from the perspective of wavelets. Danihelka, Wayne, Uria, Kalchbrenner, and Graves (2016) presented faster learning by adding some complex-valued modules to the recurrent architecture. Some studies showed that complex-valued neural networks can surpass their real-valued contenders in some applications, such as vision (Koenderink, van Doorn, & Gegenfurtner, 2021; Oyallon & Mallat, 2015; Virtue, Stella, & Lustig, 2017; Worrall, Garbin, Turmukhambetov, & Brostow, 2017), NLP (Trouillon, Welbl, Riedel, Gaussier, & Bouchard, 2016), signal processing (Adali, Schreier, & Scharf, 2011; Hirose & Yoshida, 2011, 2012), MRI fingerprinting (Virtue et al., 2017), time series forecasting (Burkard, Zimmermann, & Schwarzer, 2021; Wolter & Yao, 2018; Zhang & Zhou, 2021), etc.

From the theoretical perspective, Voigtlaender (2020) took an important step on the universal approximation of shallow and deep complex-valued networks, and similar to real-valued networks, complex-valued neural networks could achieve universal approximation with exponential depth or width (Arena, Fortuna, Re, & Xibilia, 1993, 1995; Voigtlaender, 2020). Several researchers made efforts on optimization dynamics, e.g., all critical points are proven to be saddle points, which are generated from the hierarchical structure of complex-valued networks (Nitta, 2002, 2013), and Adali et al. (2011) showed that complex-valued networks have no bad local minima as for fitting low-degree polynomials. Despite promising theoretical progress, the advantages or killer areas of complex-valued neural networks are not yet

[*] Corresponding author.
E-mail addresses: zhangsq@lamda.nju.edu.cn (S.-Q. Zhang), gaow@lamda.nju.edu.cn (W. Gao), zhouzh@lamda.nju.edu.cn (Z.-H. Zhou).

https://doi.org/10.1016/j.neunet.2022.03.024
0893-6080/© 2022 Elsevier Ltd. All rights reserved.

well understood theoretically compared to the existing network models.

This work presents theoretical understandings on the advantages of complex-valued networks in comparison with real-valued ones. We investigate a practical complex-valued neural network with fully-connected feed-forward architecture from the perspectives of approximation and optimization dynamics. The main theoretical results can be summarized as follows:

- For approximation, we show that the complex-reaction network has universal approximation property in Theorem 1.
- We prove that a kind of radial functions can be approximated by a complex-reaction network with a polynomial number of parameters, whereas the real-valued network cannot arrive at the same approximation level even with exponential ($\mathcal{O}(C_1(2d+1)e^{C_1(2d)})$) parameters for some constant $C > 0$, where $2d$ denotes the input dimension. This conclusion is shown in Theorem 2.
- For optimization dynamics, we consider the empirical risk minimization based on the standard gradient descent algorithm, and provide a corresponding convergence analysis in Theorem 3.
- We prove that the critical point set of complex-reaction networks is a proper subset of that of real-valued networks in Theorem 4, which may shed some insights on finding optimal solutions more easily for complex-reaction networks.

The rest of this paper is organized as follows. Section 2 introduces the preliminaries and notations. Section 3 presents the approximation analysis of complex-reaction networks. Section 4 studies the optimization dynamics of complex-reaction networks. Section 5 discusses with future issues. Section 6 concludes this work.

2. Preliminaries

We start our work by introducing the *Complex-Reaction Network* with fully-connected feed-forward architecture. Let $z = z_1 + z_2\mathrm{i}$ be a complex number where $\mathrm{i} = \sqrt{-1}$ and $z_1, z_2 \in \mathbb{R}$. We denote by $\bar{z} = z_1 - z_2\mathrm{i}$ and $|z|^2 = z_1^2 + z_2^2$. Let $[\cdot]_R$ and $[\cdot]_I$ denote the operators on the extraction of real and imaginary parts from a complex-valued formation, respectively, for examples, $[z_1 + z_2\mathrm{i}]_R = z_1$ and $[z_1 + z_2\mathrm{i}]_I = z_2$.

Generally, we consider the real-valued data including instances and labels, as the works of Hirose and Yoshida (2012), Trabelsi et al. (2018), Wolter and Yao (2018). For complex-valued formation, we enable the first d feature maps to represent the real components and the remaining d to record the imaginary ones, which has been implemented by Trabelsi et al. (2018). Hence, the basic building block of a complex-reaction network can be formalized as

$$\tau : \mathbb{C}^d \to \mathbb{C}, \quad \boldsymbol{z} \mapsto \sigma_{cr}(\boldsymbol{w}^\top \boldsymbol{z})$$

where $\boldsymbol{w} \in \mathbb{C}^d$ denotes the connection weights and σ_{cr} is a complex-valued activation function. In this work, we employ the zReLU function (Trabelsi et al., 2018; Zhang & Zhou, 2021) as activation σ_{cr}

$$\mathrm{zReLU}(z) = \begin{cases} z, & \text{if} \quad \theta_z \in [0, \pi/2] \cup [\pi, 3\pi/2], \\ 0, & \text{otherwise}. \end{cases}$$

Thus, for $\alpha \in \mathbb{R}^+$, we have the following complex-homogeneity property

$$\sigma_{cr}(z) = \frac{\partial \sigma_{cr}(z)}{\partial z} z \quad \text{and} \quad \sigma_{cr}(\alpha z) = \alpha \sigma_{cr}(z).$$

We also employ a pure linear connection as the final layer and extract the real part as the outputs. Thus, we have established the *Complex-Reaction Network*, denoted by $f_{CR} : \mathbb{C}^d \to \mathbb{R}$.

Notice that a fully-connected complex-reaction network with one-hidden layer has $l(d+m)$ complex-valued connection weights, which is equivalent to $2l(d+m)$ real-valued connection weights, where l and m denote the number of neurons in hidden and output layers, respectively. Notice that we focus on the number of "real-valued" parameters when one mentions the number of parameters in complex-reaction networks. Besides, provided homogeneous activation functions, for $a, b \in \mathbb{C}$, we have

$$\begin{cases} \left[\dfrac{\partial f_{CR}(a \cdot b)}{\partial b}\right]_R = \dfrac{\partial [f_{CR}(a \cdot b)]_R}{\partial [b]_R} = \dfrac{\partial [f_{CR}(a \cdot b)]_I}{\partial [b]_I}, \\ \left[\dfrac{\partial f_{CR}(a \cdot b)}{\partial b}\right]_I = \dfrac{\partial [f_{CR}(a \cdot b)]_I}{\partial [b]_R} = -\dfrac{\partial [f_{CR}(a \cdot b)]_R}{\partial [b]_I}. \end{cases} \tag{1}$$

Finally, we introduce the some notations. Let $[N] = \{1, 2, \ldots, N\}$ be the set for an integer $N > 0$. Given a function $g(n)$, we denote by $h_1(n) = \Theta(g(n))$ if there exist positive constants c_1, c_2 and n_0 such that $c_1 g(n) \leq h_1(n) \leq c_2 g(n)$ for every $n \geq n_0$; we also denote by $h_2(n) = \mathcal{O}(g(n))$ if there exist positive constants c and n_0 such that $h_2(n) \leq c g(n)$ for every $n \geq n_0$.

For $\boldsymbol{w} \in \mathbb{C}^n$ and $\mathbf{W} \in \mathbb{C}^{n \times m}$, we denote by

$$\|\boldsymbol{w}\|_2 \overset{\text{def}}{=} \left(\sum_{i=1}^n |w_i|^2\right)^{1/2} \quad \text{and} \quad \|\mathbf{W}\|_2 \overset{\text{def}}{=} \left(\sum_{i=1}^n \sum_{j=1}^m |w_{ij}|^2\right)^{1/2}$$

the 'entry-wise' vector norm and matrix norm, respectively.

We say that f is a radial function if $f(\boldsymbol{x}) = f(\boldsymbol{x}')$ for $\|\boldsymbol{x}\| = \|\boldsymbol{x}'\|$. Let $\phi^2(\boldsymbol{x})$ be the density function of some probability measure μ, which satisfies

$$\int_{\boldsymbol{x} \in \mathbb{R}^{2d}} \phi^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{x} \in \mathcal{B}} 1 \, \mathrm{d}\boldsymbol{x} = 1, \tag{2}$$

where $\mathcal{B}$ is a unit ball. Then, for continuous functions f, g, we have the following equalities under Fourier transform

$$\|\widehat{f\phi} - \widehat{g\phi}\|_{L_2(\mu)} = \|f\phi - g\phi\|_{L_2(\mu)}, \tag{3}$$

and

$$\widehat{f\phi} = \hat{f} * \hat{\phi}, \quad \text{for} \quad \text{the convolution operator} * .$$

3. Approximation

In this section, we first present the universal approximation for the complex-reaction network, which assures its legitimacy, and then, show the approximation complexity advantages of complex-reaction networks over the real-valued ones.

We now present the universal approximation for complex-reaction networks as follows.

Theorem 1. *For zReLU activation function, the complex-reaction network with one-hidden layer $f_{CR} : \mathbb{C}^d \to \mathbb{R}$ is a universal approximator for any continuous function $f : \mathbb{R}^{2d} \to \mathbb{R}$, where $\mathbb{C} \cong \mathbb{R}^2$.*

Theorem 1 shows that complex-reaction networks have the universal approximation property, similar to that of the real-valued networks. This theorem can be easily derived from (Voigtlaender, 2020, Theorem 1.3), where a shallow complex-valued neural network has the universal approximation property, and we omit the detailed proof of Theorem 1.

Next, we present the approximation complexity theorem for complex-reaction networks as follows.

Theorem 2. *For zReLU activation function, there exist a probability measure μ and a radial function $f : \mathbb{R}^{2d} \to \mathbb{R}$ such that*

(i) *for any $\delta > 0$, there is a one-hidden-layer complex-reaction network $f_{CR} : \mathbb{C}^d \to \mathbb{R}$ with $\mathcal{O}((d+1)(2d)^{3.75})$ parameters such that the followings hold:*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}(f_{CR}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 < \delta,$$

(ii) *there exists a constant $\delta > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}\,(f_R(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \geq \delta,$$

for any one-hidden-layer real-valued network $f_R : \mathbb{R}^{2d} \to \mathbb{R}$ with exponential parameter number $\mathcal{O}(C_1(2d+1)e^{C_1(2d)})$. Here, C_1 is a constant independent to d.

This theorem shows that a kind of radial functions can be approximated by one-hidden-layer complex-reaction networks of polynomial parameters, whereas such functions cannot be approximated by real-valued networks with exponential ($O(C_1(2d+1)e^{C_1(2d)})$) parameters.

This proof idea can be summarized as follows. It is observed that the zReLU function σ_{cr} comprises the radius (i.e., norm) and phase (i.e., angle). Thus, there exist some linear connections (including rotation transformations) such that the combination of some complex-reaction neurons is invariant to rotations. In other words, complex-reaction networks can easily and well approximate some radial functions, see Lemma 2, since the radial function is invariant to rotations, and is dependent on the input norm. On the other hand, provided the probability measure μ and corresponding density function ϕ, defined as Eq. (2), it is observed that the composition $f\phi$ is radial as well as some radial function f. Further, we conjecture that $\widehat{f\phi}$ is also a radial function according to Eq. (3) with $g \equiv 0$. In contrast, the distribution of the composition $f_R\phi$ under Fourier transform, corresponding to a one-hidden-layer real-valued network f_R, is supported on a finite collection of lines $\{\boldsymbol{w}_i^\top \boldsymbol{x}\}$, i.e.,

$$\mathrm{Supp}(\widehat{f_R\phi}) \subseteq \left\{ \boldsymbol{x} \;\middle|\; \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq 1, \boldsymbol{x}' \in \bigcup_{i=1}^{k} \mathrm{span}\{\boldsymbol{w}_i\} \right\}.$$

Notice that the $\mathrm{Supp}(\widehat{f_R\phi})$ is sparse in the Fourier space unless k is an exponential. Thus, a one-hidden-layer real-valued network within polynomial parameters cannot achieve arbitrarily approximation for radial functions.

Based on the ideas above, it is sufficient to provide an approximation guarantee between the complex-reaction network f_{CR} and target function f, that is,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mu}(f_{CR}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 &= \int (f_{CR}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \phi^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \|(f_{CR} - f)\phi\|_{L_2}^2(\mu) = \|\widehat{f\phi} - \widehat{g\phi}\|_{L_2(\mu)}^2 \leq \delta.\end{aligned}$$

Motivated from Eldan and Shamir (2016), we consider the radial function

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x}) \quad \text{with} \quad f_i(\boldsymbol{x}) = \mathbf{1}\{\|\boldsymbol{x}\| \in \Omega_i\}, \tag{4}$$

where $\epsilon_i \in \{-1, +1\}$, N is a polynomial function of d, and Ω_i's are disjoint intervals of width $\mathcal{O}(1/N)$ on values in the range $\Theta(\sqrt{2d})$. Next, we formally begin our proof of Theorem 2 with some useful lemmas.

Lemma 1. *Let $f : [-R, R] \to \mathbb{R}$ be an L-Lipschitz function for constant $R > 0$. For any $\delta > 0$, $C_r \geq 1$, and $n_r \leq 2C_rLR/\delta$, there exists a real-valued network with one-hidden layer f_R s.t.*

$$\sup_{x\in\mathbb{R}^{2d}} |f(x) - f_R(x)| \leq \delta.$$

Specifically, $f_R : \mathbb{R} \to \mathbb{R}$ can be given by

$$f_R(x) = \sum_{i=1}^{n_r} \alpha_i\, \sigma_r(\beta_i x - b_i) - a,$$

for ReLU or general sigmoidal activation σ_r and real-valued parameters a, $\{\alpha_i, \beta_i, b_i\}_{i=1}^{n_r}$.

Lemma 1 shows that a one-dimensional L-Lipschitz function can be approximated by real-valued networks of one-hidden layer with general sigmoidal or ReLU activations. The detailed proof of Lemma 1 is given in Appendix A.2.

We now present a crucial lemma for complex-valued networks from Lemma 1 as follows.

Lemma 2. *Let $g : [r, R] \to \mathbb{R}$ be an L-Lipschitz function for $r \leq R$. For any $\delta > 0$, $C_{cr} \geq 1$, and $n_{cr} \leq 2C_{cr}dR^2L/(\sqrt{r}\delta)$, there exists a complex-reaction network with one-hidden layer f_{CR} s.t.*

$$\sup_{\boldsymbol{x}\in\mathbb{C}^d} |g(\|\boldsymbol{x}\|) - f_{CR}(\boldsymbol{x})| \leq \delta.$$

Specifically, $f_{CR} : \mathbb{C}^d \to \mathbb{R}$ can be given by

$$f_{CR}(\boldsymbol{x}) = \left[\sum_{i=1}^{n_{cr}} v_i\, \sigma_{cr}(\boldsymbol{w}_i^T \boldsymbol{x} - b_i) - a \right]_R,$$

for zReLU function σ_{cr} and complex-valued parameters a, $\{\boldsymbol{w}_i, v_i, b_i\}_{i=1}^{n_{cr}}$.

Proof. Let $f' : \mathbb{C}^d \to \mathbb{R}$ be a radial function with $f'(x) = |x|$. For any $\delta > 0$ and $d \geq 1$, we have

$$\sup_{x\in\mathbb{C}} \left| f'(x) - |\sigma_{cr}(wx - b)| \right| \leq \delta/2. \tag{5}$$

We further introduce a new function $g' : \mathbb{R} \to \mathbb{R}$ as follows.

$$g'(s) = \sum_{i=1}^{n'} \alpha'_i \sigma'(s) - a'_i,$$

where σ is the ReLU function and $\alpha'_i, a'_i \in \mathbb{R}$. For Lipschitz continuous function $\sqrt[r]{\cdot}$ and from Lemma 1, we have

$$\sup_{s\in[r^k,R^k]} \left| g(\sqrt[k]{s}) - g'(s) \right| \leq \delta/2, \tag{6}$$

when $n' \leq C'L(R^k - r^k)/(\sqrt[k]{r}\delta)$ for some constant $C' > 0$ and integer $k \geq 2$. Given complex-reaction network

$$f_{CR}(\boldsymbol{x}) = \left[\sum_{i=1}^{n_{cr}} v_i\, \sigma_{cr}(\boldsymbol{w}_i^\top \boldsymbol{x} - b_i) - a \right]_R,$$

we have

$$\begin{aligned}\left| g'(s) - f_{CR}(\boldsymbol{x}) \right| &= \left| g'(s) - f'(\boldsymbol{x}) \right| \\ &\quad + \left| f'(\boldsymbol{x}) - \left[\sum_{i=1}^{n_{cr}} v_i\, \sigma_{cr}(\boldsymbol{w}_i^\top \boldsymbol{x} - b_i) - a \right]_R \right|,\end{aligned} \tag{7}$$

where

$$f'(\boldsymbol{x}) = \sum_{i=1}^{n'_{cr}} v'_i \, \left| \sigma_{cr}(\boldsymbol{w}_i'^\top \boldsymbol{x} - b'_i) \right| - a'.$$

in which $\{\boldsymbol{w}'_i, b'_i\}$ and $\{v'_i\}$, a' denote another collection of complex-valued and real-valued parameters, respectively. The first term of Eq. (7) can be bounded $\delta/4$ from Lemma 1 for any $s \in [r^k, R^k]$. The second term is at most $\delta/4$ when $n_{cr} \geq n'_{cr}$ from Eq. (5). This follows that

$$\left| g'(s) - f_{CR}(\boldsymbol{x}) \right| \leq \delta/2. \tag{8}$$

Combining with Eqs. (6) and (8), we have

$$|g(\|\boldsymbol{x}\|) - f_{CR}(\boldsymbol{x})| \leq \left|g(\sqrt[k]{s}) - g'(s)\right| + \left|g'(s) - f_{CR}(\boldsymbol{x})\right| \leq \delta,$$

where $\boldsymbol{x} \in \mathbb{R}^{2d} \cong \mathbb{C}^d$ and $s \in [r^k, R^k]$. We finally obtain

$$n_{cr} \leq 2C_{cr}(R^k - r^k)dL/(\sqrt[k]{r}\delta),$$

provided $n_{cr} \leq 2dn'$ and $C' \leq C_{cr}$. We complete the proof by setting $k = 2$ in the above upper bound. □

Lemma 3. *For $2d > C_2 > 0$, let $f : \mathbb{R}^{2d} \to \mathbb{R}$ is an L-Lipschitz radial function supported on the set*

$$\mathcal{S} = \{\boldsymbol{x} : 0 < C_2\sqrt{2d} \leq \|\boldsymbol{x}\| \leq 2C_2\sqrt{2d}\}.$$

For any $\delta > 0$, there exists a complex-reaction network f_{CR} of one-hidden layer with width at most $2C_{cr}(C_2)^{3/2}L(2d)^{7/4}/\delta$ such that

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{2d}\cong\mathbb{C}^d} |f(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})| < \delta.$$

Lemma 3 shows that the radial functions can be approximated by complex-reaction networks with polynomial parameters, which is proved as follows.

Proof. Let $r = C_2\sqrt{2d}$, $R = 2C_2\sqrt{2d}$, and $d \geq 1$, then we have $r \geq 1$, which satisfies the condition of Lemma 2. Invoke Lemma 2 to construct the complex-reaction networks and define $\delta' \leq \delta/2d$. Then for any L-Lipschitz radial function $f : \mathbb{R}^{2d} \to \mathbb{R}$ supported on $\mathcal{S}$, we have

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{2d}\cong\mathbb{C}^d} |f(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})| \leq \delta',$$

where the width of the hidden layer is bounded by

$$n_{cr} \leq \frac{2C_{cr}(C_2)^{3/2}dL}{\delta}(2d)^{3/4} \leq \frac{2C_{cr}(C_2)^{3/2}L}{\delta}(2d)^{7/4}.$$

This completes the proof. □

Lemma 4. *Let $f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x})$ be defined by Eq. (4). For any $\epsilon_i \in \{-1, +1\}$ $(i \in [N])$, there exists a Lipschitz function $g : \mathcal{S} \to [-1, +1]$ such that*

$$\int_{\mathbb{R}^{2d}} (g(\boldsymbol{x}) - f(\boldsymbol{x}))^2\, \phi^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \leq \frac{3}{(C_2)^2\sqrt{2d}}.$$

Lemma 4 shows that any non-Lipschitz function $f(\boldsymbol{x})$ can be approximated and bounded by a Lipschitz function with density ϕ^2, which is proved in Appendix A.3.

So far, the part (i) of Theorem 1 can be summarized as follows.

Proposition 1. *Let f be the radial function described by Eq. (4). For $C_2, C_3 > 0$ with $d > C_2$, any $\delta > 0$, and any choice of $\epsilon_i \in \{-1, +1\}$ $(i \in [N])$, there exists a complex-reaction network f_{CR} of one-hidden layer with range in $[-2, +2]$ and width at most $C_3C_{cr}(2d)^{3.75}$, such that*

$$\|f(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})\|_{L_2(\mu)} \leq \frac{\sqrt{3}}{C_2(2d)^{1/4}} + \delta.$$

Lemma 5. *For positive constants $C_1, C_2, C_3, \rho, \alpha$ with $2d > C_2$ and $\alpha > C_2$, we define*

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x}) \quad \text{and} \quad f_R(\boldsymbol{x}) = \sum_{i=1}^{n_r} \tilde{f}_i(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle),$$

where $N \geq 4C_2\alpha^{3/2}d^2$, $n_r \leq C_1e^{2C_1d}$, and $\tilde{f}_i : \mathbb{R} \to \mathbb{R}$ are measurable functions with $|f_i(x)| \leq C_3(1 + |x|^\rho)$. For any $\delta > 0$, there exists a group of $\epsilon_i \in \{-1, +1\}$ $(i \in [N])$ such that

$$\|f(\boldsymbol{x}) - f_R(\boldsymbol{x})\| \geq \delta/\alpha.$$

Lemma 5 shows that some radial function cannot be approximated by real-valued networks with exponential $(\mathcal{O}(C_1e^{C_1(2d)}))$ neurons, beyond which (ii) of Theorem 1 holds. The detailed proof can be accessed in Appendix A.4.

Proof of Theorem 2. Let $f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x})$ be defined by Eq. (4) and $N \geq 4C_2^{5/2}d^2$. According to Lemma 4, there exists a Lipschitz function h with range $[-1, +1]$ such that

$$\|h(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L_2(\mu)} \leq \frac{\sqrt{3}}{C_2(2d)^{1/4}}.$$

Based on Lemmas 2 and 3, any Lipschitz radial function supported on $\mathcal{S}$ can be approximated by a complex-reaction network f_{CR} with one-hidden layer of width at most $C_3C_{cr}(2d)^{3.75}$, where C_3 is a constant relative to C_2 and δ. This means that,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{2d}} |h(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})| \leq \delta.$$

Thus, we have

$$\|h - f_{CR}\|_{L_2(\mu)} \leq \delta.$$

Hence, the range of f_{CR} is in $[-1 - \delta, +1 + \delta] \subseteq [-2, +2]$. In summary, we have

$$\begin{aligned}\|f(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})\|_{L_2(\mu)} &\leq \|f(\boldsymbol{x}) - h(\boldsymbol{x})\|_{L_2(\mu)} + \|h(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})\|_{L_2(\mu)} \\ &\leq \frac{\sqrt{3}}{C_2(2d)^{1/4}} + \delta.\end{aligned}$$

This implies that given constants $2d > C_2 > 0$ and $C_3 > 0$, for any $\delta > 0$ and $\epsilon_i \in \{-1, +1\}$ $(i \in [N])$, the target radial function f can be approximated by a complex-reaction network f_{CR} of one-hidden layer with range in $[-2, +2]$ and width at most $C_3C_{cr}(2d)^{3.75}$, that is,

$$\|f_{CR} - f\|_{L_2(\mu)} \leq \frac{\sqrt{3}}{C_2(2d)^{1/4}} + \delta < \delta_1.$$

According to Lemmas 1 and 5, there are some groups of $\epsilon_i \in \{-1, +1\}$ $(i \in [N])$ such that

$$\|f_R - f\|_{L_2(\mu)} \geq \delta_1,$$

for any real-valued network f_R of one-hidden layer with width at most $C_1C_re^{2C_1d}$. The real-valued and complex-reaction networks have the number of parameters:

$$\begin{cases} N_r = (2d) \times n_r + n_r \leq C_1C_r(2d + 1)e^{C_1(2d)}, \\ N_{cr} = 2 \times d \times n_{cr} + 2 \times n_{cr} \leq 2C_3C_{cr}(d + 1)(2d)^{3.75}, \end{cases}$$

where N_r and N_{cr} indicate the parameter numbers of the real-valued and complex-reaction networks, respectively. This completes the proof. □

4. Optimization dynamics

This section studies the optimization dynamics of complex-reaction networks, and focuses on binary classification where $y \in \{-1, +1\}$ for simplicity. Let $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ be a training dataset, and denote by $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$. We employ the zReLU activation function, and, for convenience, use $[f(\mathbf{W}; \boldsymbol{x})]_R$ to denote the complex-reaction network. We consider minimizing the empirical exponential loss, as studied in Hirose (2012)

$$L(\mathbf{W}; \mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} \exp\left(-y_n[f(\mathbf{W}; \boldsymbol{x}_n)]_R\right), \tag{9}$$

where $\mathbf{W} = (\mathbf{W}^1, \mathbf{W}^2, \ldots, \mathbf{W}^L)$ for some integer $L > 0$.

4.1. Convergence analysis

This subsection focuses on the convergence of minimizing Eq. (9) using complex gradient descents. As shown in previous works (Zhang & Mandic, 2015; Zhang & Zhou, 2021), the complex gradient descent usually follows the recursive form

$$\mathbf{W}^l(t+1) = \mathbf{W}^l(t) - \rho \frac{\partial L(\mathbf{W}(t); \mathbf{X})}{\partial \mathbf{W}^l}, \tag{10}$$

where ρ is the learning rate at the t-th epoch and $t \in \mathbb{N}^+$.

We now provide the convergence analysis for Eq. (10).

Theorem 3. *Let $\mathbf{W}(0)$ denote the initial point and the gradient $\partial L(\mathbf{W}(t); \mathbf{X})/\partial \mathbf{W}^l$ is an L-Lipschitz function with respect to $\mathbf{W}^l$. Then, provided a real-valued learning rate, i.e., $\rho \in \mathbb{R}$, if $\rho \in [0, (2L\theta)^{-1}]$, we have*

$$L(\boldsymbol{w}(t+1)) - L(\boldsymbol{w}(t)) \le 0 \quad \text{and} \quad \lim_{t\to\infty} \left\| \frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}} \right\| = 0.$$

Provided a complex-valued learning rate, i.e., $\rho \in \mathbb{C}$, if $2L\theta([\rho]_R^2 + [\rho]_I^2) \le [\rho]_R$, we have

$$L(\boldsymbol{w}(t+1)) - L(\boldsymbol{w}(t)) \le 0 \quad \text{and} \quad \lim_{t\to\infty} \left\| \frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}} \right\| = 0.$$

Theorem 3 shows the convergence of complex gradient descents with real-valued and complex-valued learning rates for minimizing the empirical exponential loss. The proof idea is similar to those of Zhang, Liu, Xu, and Zhang (2014).

Proof. Let $\boldsymbol{w}$ denote the i-th column vector of $\mathbf{W}^l$ for any i and l and abbreviate $L(\mathbf{W}(t); \mathbf{X})$ as $L(\boldsymbol{w}(t))$. It is observed that

$$\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}} = -\frac{1}{N}\sum_{n=1}^{N} y_n \exp\left(-y_n[f(\boldsymbol{w}(t))]_R\right) \times \frac{\partial [f(\boldsymbol{w}(t))]_R}{\partial \boldsymbol{w}} \quad \text{and} \quad L(\boldsymbol{w}(t)) \ge 0. \tag{11}$$

Since $\exp(\cdot)$ is a monotonically increasing function, we have

$$\frac{\partial L(\boldsymbol{w}(t))}{\partial \bar{\boldsymbol{w}}} = \frac{\partial L(\boldsymbol{w}(t))}{\partial [\boldsymbol{w}]_R} - \frac{\partial L(\boldsymbol{w}(t))}{\partial [\boldsymbol{w}]_I} \mathrm{i} = \overline{\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}}.$$

Let

$$\Delta_t = \boldsymbol{w}(t+1) - \boldsymbol{w}(t)$$

and

$$\overline{\Delta_t} = \overline{\boldsymbol{w}(t+1) - \boldsymbol{w}(t)} = \overline{\boldsymbol{w}(t+1)} - \overline{\boldsymbol{w}(t)}.$$

Thus, we have

$$\begin{aligned}
&L(\boldsymbol{w}(t+1)) - L(\boldsymbol{w}(t)) \\
&\overset{(a)}{=} \frac{1}{2}\frac{\partial L(\boldsymbol{w}(t)+\theta\Delta_t)}{\partial \boldsymbol{w}}\Delta_t + \frac{1}{2}\frac{\partial L(\boldsymbol{w}(t)+\theta\Delta_t)}{\partial \bar{\boldsymbol{w}}}\overline{\Delta_t} \\
&= \frac{1}{2}\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\Delta_t + \frac{1}{2}\left[\frac{\partial L(\boldsymbol{w}(t)+\theta\Delta_t)}{\partial \boldsymbol{w}} - \frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right]\Delta_t \\
&\quad + \frac{1}{2}\frac{\partial L(\boldsymbol{w}(t))}{\partial \bar{\boldsymbol{w}}}\overline{\Delta_t} + \frac{1}{2}\left[\frac{\partial L(\boldsymbol{w}(t)+\theta\Delta_t)}{\partial \bar{\boldsymbol{w}}} - \frac{\partial L(\boldsymbol{w}(t))}{\partial \bar{\boldsymbol{w}}}\right]\overline{\Delta_t} \\
&\le \left[\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\Delta_t\right]_R + \left\|\frac{\partial L(\boldsymbol{w}(t)+\theta\Delta_t)}{\partial \boldsymbol{w}} - \frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\| \|\Delta_t\| \\
&\le \left[\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\Delta_t\right]_R + 2L\theta \|\Delta_t\|^2 \\
&\overset{(b)}{=} \left[-\rho\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\left(\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right)^{\top}\right]_R + 2L\theta\left\|\rho\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\|^2.
\end{aligned}$$

where the equalities (a) and (b) follow from the mean value theorem and Eq. (10), respectively. Provided $\rho \in \mathbb{C}$, we have

$$L(\boldsymbol{w}(t+1)) - L(\boldsymbol{w}(t)) = \left(-[\rho]_R + 2L\theta|\rho|^2\right)\left\|\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\|^2.$$

Once $-[\rho]_R + 2L\theta|\rho|^2 \le 0$, then we have $L(\boldsymbol{w}(t+1)) - L(\boldsymbol{w}(t)) \le 0$. Further, we have

$$\begin{aligned}
L(\boldsymbol{w}(t+1)) &\le L(\boldsymbol{w}(t)) + \left(-[\rho]_R + 2L\theta|\rho|^2\right)\left\|\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\|^2 \\
&\le L(\boldsymbol{w}(0)) + \left(-[\rho_t]_R + 2L\theta|\rho|^2\right)\sum_{s=1}^{t}\left\|\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\|^2.
\end{aligned}$$

According to Eq. (11), the following holds

$$\left([\rho_t]_R - 2L\theta|\rho|^2\right)\sum_{s=1}^{t}\left\|\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\|^2 \le L(\boldsymbol{w}(0)) < \infty \quad \text{for} \quad t \to \infty,$$

which implies that

$$\lim_{t\to\infty}\left\|\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}\right\| = 0.$$

One can assess the proof of the case that $\rho \in \mathbb{R}$ from Zhang et al. (2014, Theorem 1). This completes the proof. □

4.2. Critical point sets

This subsection concerns the critical point sets of real-valued and complex-reaction networks by using standard gradient descents. For specification, we here employ the weight normalization technique, which is formulated as follows

$$\mathbf{W}_j^l = \gamma_j^l\, \mathbf{V}_j^l, \quad \text{with} \quad \gamma_j^l \in \mathbb{R}^+ \quad \text{and} \quad \|\mathbf{V}_j^l\| = 1,$$

where $\mathbf{W}_j^l$ denotes the i-th row vector of the matrix $\mathbf{W}^l$ in the l-th layer. Hence, we can limit the critical points of the optimization problem in Eq. (9) onto the unit ball.

We now present the main result as follows.

Theorem 4. *For the minimization of Eq. (9), we have*

$$\mathcal{G} \subseteq \mathcal{S}_{CR} \subseteq \mathcal{S}_R \quad \text{and} \quad \mathcal{S}_{CR} \ne \mathcal{S}_R,$$

where $\mathcal{G}$ is the optima set of Eq. (9), $\mathcal{S}_R$ and $\mathcal{S}_{CR}$ denote the critical point sets of real-valued and complex-reaction networks, respectively.

Theorem 4 shows that the critical point set of complex-reaction networks is a proper subset of that of real-valued networks, which may shed some insights on finding optimal solutions more easily for complex-reaction networks.

This proof idea can be summarized as follows. It forms a manifold for the parameters space of neural networks. Generally speaking, the complex manifold Ω_{CR} is a subset of the real manifold Ω_R, since the coordinate transformation of complex manifold satisfies the holomorphic condition. On the other hand, it is essential to look for the critical points $\boldsymbol{\theta}$ with $\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = 0$ when we solve an empirical minimization optimization using gradient descents. Therefore, the proof of Theorem 4 can be converted into the problem of finding the critical points in Ω_R yet except Ω_{CR}.

We construct a transformation to link the real and complex manifolds that correspond to the real-valued and complex-reaction networks, respectively. Finally, we find the desired points from the symplectic manifolds, which share certain characteristics with Riemannian geometry and complex geometry and link two geometric theories in some fields of mathematics.

Definition 1. Define a linear mapping $\phi : K^{p\times q} \to K^{p\times q}$,

$$\phi(\Theta; \rho, i, j) = (\ \ldots;\ \ \underbrace{\rho\boldsymbol{\theta}_j + \boldsymbol{\theta}_i}_{i}\ \ ; \ldots;\ \ \underbrace{(1-\rho)\boldsymbol{\theta}_j}_{j}; \ldots\ ),$$

where $K^{p\times q}$ is compact in $\mathbb{R}^{p\times q}$ or $\mathbb{C}^{p\times q}$, $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q) \in K^{p\times q}$, $\rho \in \mathbb{R}$ or $\mathbb{C}$, and $i, j \in [q]$.

Lemma 6. *For any $i, j \in [q]$, $\phi(\Theta)$ consists of some straight lines in an 2-dimensional affine space.*

Lemma 6 shows that the linear mapping ϕ leads to an affine space defined by $\boldsymbol{\theta}_j$, which is proved by Appendix B.1.

Definition 2. A mapping is said to be **analytic** if it is continuous and expandable in a power series around any points.

Definition 3. Let f be a function expressed by a neural network with parameter space K. An analytic mapping $\mathcal{T} \in \mathcal{C}^1(K)$ is said to be an **equioutput transformation**, if $f(\mathcal{T}(\mathbf{W}); \mathbf{X}) = f(\mathbf{W}; \mathbf{X})$ for $\mathbf{V} \in K$.

Lemma 7. *For any equioutput transformation $\mathcal{T} \in \mathcal{C}^1(K)$, there is a collection of finite mappings $\{\phi\}$, such that $\mathcal{T}$ is a composition of ϕ's.*

Lemma 8. *All equioutput transformations, generated from Lemma 7, constitute a multiplicative group $\mathbb{G}$, which is isomorphic to a direct product of Weyl groups.*

A straightforward combination of Lemmas 7 and 8 shows that the equioutput transformation is composited of finite linear mapping ϕ, and for any i, j, the equioutput transformations constitute an algebraic group, isomorphic to a direct product of Weyl groups (Warner, 1983). The detailed proof of Lemma 8 are present as follows.

Next, we provide two crucial propositions about the dynamical systems led by complex-reaction and real-valued networks with gradient descents, respectively.

Proposition 2. *For the minimization of Eq. (9) using complex-reaction networks, we have the following dynamical systems*

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{N\gamma_j^l} \displaystyle\sum_{n=1}^{N} \exp\left(-y_n[f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n [f(\mathbf{V}; \boldsymbol{x}_n)]_R, \\ \dfrac{d\mathbf{V}_j^l}{dt} = \dfrac{\eta}{N(\gamma_j^l)^2} \displaystyle\sum_{n=1}^{N} \exp\left(-y_n[f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \Delta_j, \end{cases}$$

where η is a strictly positive constant relative to γ_j^l, and

$$\Delta_j = \left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_R} - [\mathbf{V}_j^l]_R [f(\mathbf{V}; \boldsymbol{x}_n)]_R \right) + \left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_I}{\partial \left[\mathbf{V}_j^l\right]_I} - [\mathbf{V}_j^l]_I [f(\mathbf{V}; \boldsymbol{x}_n)]_R \right) \mathrm{i}.$$

Proposition 3. *Let $f_R : \mathbb{R}^{2d} \to \{-1, +1\}$ be a real-valued neural network with ReLU activation and weight normalization $\mathbf{P}_j^l = \gamma_j^l \mathbf{Q}_j^l$ where $\|\mathbf{Q}_j^l\| = 1$. the gradient descent procedure for minimizing the exponential loss coincides with the following dynamical systems*

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{N\gamma_j^l} \dfrac{1}{N} \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) y_n f_R(\mathbf{Q}; \boldsymbol{x}_n), \\ \dfrac{d\mathbf{Q}_j^l}{dt} = \dfrac{\eta}{N(\gamma_j^l)^2} \displaystyle\sum_{n=1}^{N} \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) y_n \boldsymbol{\phi}_j^l, \end{cases}$$

where η is a strictly positive constant relative to γ_j^l and

$$\boldsymbol{\phi}_j^l = \partial f_R(\mathbf{Q}; \boldsymbol{x}_n) / \partial \mathbf{Q}_j^l - \mathbf{Q}_j^l f_R(\mathbf{Q}; \boldsymbol{x}_n).$$

Propositions 2 and 3 hold from Lemmas 9 and 10 as follows.

Lemma 9. *For $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{1\times n}$ and $\mathbf{S} = \mathbf{I}_{n\times n} - \boldsymbol{v}^\top \boldsymbol{v}$ with $\boldsymbol{w} = \gamma \boldsymbol{v}$ and $\|\boldsymbol{v}\| = 1$, we have $\mathbf{S} = \mathbf{I}_{n\times n} - \boldsymbol{w}^\top \boldsymbol{w} / \|\boldsymbol{w}\|_2^2$, $\partial \boldsymbol{v} / \partial \boldsymbol{w} = \mathbf{S}/\gamma$, $\mathbf{S}\boldsymbol{w}^T = \mathbf{S}\boldsymbol{v}^T = \mathbf{0}$, and $\mathbf{S}^2 = \mathbf{S}$.*

Lemma 10 (*Weight Norms*)**.** *During the gradient descent procedure, the change rate of $\|\gamma_j^l\|$ (i.e., weight norms) is the same for each layer.*

Lemmas 9 and 10 stand up for both complex-reaction and real-valued networks. Notice that Lemma 10 shows that $(\gamma_j^l)^2 = \|\mathbf{W}_j^l\|^2$ grows at a rate independent of the row j and layer l. Thereby, using gradient descent to solve the optimization problem, there is no difference in the change rate of connection weights layer by layer. This result also holds for real-valued networks. The detailed proofs of Lemmas 9 and 10 are presented by Appendix C.1 and C.2, respectively.

Proof of Proposition 2. We study the minimization optimization problem of complex-reaction networks as follows.

$$\min_{\mathbf{V}_j^l, \gamma_j^l} L(\mathbf{W}; \mathbf{X}) = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right),$$
$$s.t.\quad \mathbf{W}_j^l = \gamma_j^l\, \mathbf{V}_j^l, \quad \|\mathbf{V}_i^l\| = 1.$$

Solving this problem by the standard gradient descent, the optimization procedure concerning $\mathbf{W}^l$ induces the following dynamical system:

$$\frac{d\mathbf{W}^l}{dt} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \frac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \mathbf{W}^l}.$$

Similarly, the dynamical system that corresponds to the gradient descent procedure with weight normalization is

$$\begin{cases} \dfrac{d\gamma^l}{dt} = \displaystyle\sum_{n=1}^{N} \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \dfrac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \gamma^l}, \\ \dfrac{d\mathbf{V}^l}{dt} = \displaystyle\sum_{n=1}^{N} \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \dfrac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \mathbf{V}^l}. \end{cases}$$

Let $\boldsymbol{w}_j = ([\mathbf{W}_j^l]_R, -[\mathbf{W}_j^l]_I)$, $\boldsymbol{v}_j = ([\mathbf{V}_j^l]_R, -[\mathbf{V}_j^l]_I)$, and $\mathbf{S}_j = \mathbf{I} - \boldsymbol{v}_j^\top \boldsymbol{v}_j$. It is observed that if $[\mathbf{W}_j^l]_R$ has d elements, $\boldsymbol{w}_j$ is a $2d$-dimensional row vector. From Lemma 9, one has

$$\mathbf{S}_j = \mathbf{I} - \boldsymbol{w}_j^\top \boldsymbol{w}_j / \|\boldsymbol{w}_j\|_2^2 \quad \text{and} \quad \mathbf{S}_j \boldsymbol{w}_j^\top = \mathbf{S}_j \boldsymbol{v}_j^\top = \mathbf{0}.$$

Thus, we have

$$\frac{d\gamma_j^l}{dt} = \frac{d\|\boldsymbol{w}_j\|}{dt} = \boldsymbol{v}_j \left(\frac{d\boldsymbol{w}_j}{dt}\right)^\top \quad \text{and} \quad \frac{d\boldsymbol{v}_j}{dt} = \frac{\mathbf{S}_j}{\gamma_j^l} \left(\frac{d\boldsymbol{w}_j}{dt}\right)^\top.$$

So the dynamical systems concerning $\mathbf{V}^l$ and γ_j^l become

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{N\gamma_j^l} \displaystyle\sum_{n=1}^{N} y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \boldsymbol{v}_j \boldsymbol{u}_j^\top, \\ \dfrac{d\boldsymbol{v}_j}{dt} = \dfrac{\eta}{N(\gamma^l)^2} \displaystyle\sum_{n=1}^{N} y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \boldsymbol{\psi}_j, \end{cases}$$

where $\boldsymbol{\psi}_j = \left(\boldsymbol{u}_j - \boldsymbol{v}_j [f(\mathbf{V}; \boldsymbol{x}_n)]_R\right)$ and

$$\boldsymbol{u}_j = \left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_R}, -\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_I} \right).$$

Due to $\boldsymbol{v}_j\boldsymbol{u}_j^\top = [f(\mathbf{V};\boldsymbol{x}_n)]_R$, we have

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{N\gamma_j^l}\displaystyle\sum_{n=1}^N y_n \exp\left(-y_n[f(\mathbf{W};\boldsymbol{x}_n)]_R\right)[f(\mathbf{V};\boldsymbol{x}_n)]_R, \\ \dfrac{d\boldsymbol{v}_j}{dt} = \dfrac{\eta}{N(\gamma_j^l)^2}\displaystyle\sum_{n=1}^N y_n \exp\left(-y_n[f(\mathbf{W};\boldsymbol{x}_n)]_R\right)\boldsymbol{\psi}_j. \end{cases}$$

We complete this proof by multiplying the matrix $\mathbf{I}_{\mathbb{C}} = (\mathbf{I}_{d\times d}, -\mathbf{I}_{d\times d}\mathrm{i})^\top$ with the formulas above. □

Based on the aforementioned results, we now present the crucial lemmas for proving Theorem 4 as follows.

Lemma 11. *Let* $\mathbf{V}$ *be the complex-valued parameters of a complex-reaction network. If* $\mathbf{V}$ *is a critical point of* $L(\mathbf{V})$*, then we have*

(i) $\partial L(\mathcal{T}(\mathbf{V}))/\partial\mathbf{V} = 0$ *for any equioutput transformation* $\mathcal{T}$ *with* $\rho \in \mathbb{R}$*;*
(ii) $\partial L(\mathcal{T}(\mathbf{V}))/\partial\mathbf{V} \neq 0$ *for any equioutput transformation* $\mathcal{T}$ *with* $\rho \in \mathbb{C}$*.*

Proof. From Lemma 7, any transformation is composited of ϕ's. Thus, it suffices to prove that the conclusions above hold upon the minor equioutput transformation from the basic theorem of algebra. Here, we consider building the following minor equioutput transformation.

Let $\Theta \in \mathbb{C}^{p\times q}$ and $\Lambda = (\alpha_{sk}) \in \mathbb{C}^{q\times r}$ denote the connection weights of adjacent layers in a complex-reaction network, respectively. For any $i,j \in [q]$, $s \in [r]$, and $\rho,\rho' \in \mathbb{C}$, we abbreviate $\phi(\Theta;\rho,i,j)$, $\rho\boldsymbol{\theta}_j + \boldsymbol{\theta}_i$, $(1-\rho)\boldsymbol{\theta}_j$, $\rho'\alpha_{sj} + \boldsymbol{\theta}_{si}$, and $(1-\rho')\alpha_{sj}$ as $\tilde{\Theta}$, $\tilde{\boldsymbol{\theta}}_i$, $\tilde{\boldsymbol{\theta}}_j$, $\tilde{\alpha}_{si}$, and $\tilde{\alpha}_{sj}$, respectively. Hence, the output of the original network

$$h_s(\boldsymbol{z}) = \boldsymbol{\alpha}_s\sigma_{cr}\left(\Theta\boldsymbol{z}\right),$$

where $\boldsymbol{\alpha}_s$ denotes the s-th row vector of Λ and $\boldsymbol{z} \in \mathbb{C}^{p\times 1}$ is the input. Compositing the linear mapping ϕ with Θ, we have the output of the transformed network

$$\tilde{h}_s(\boldsymbol{z}) = \sum_{k\neq i,j}^{q}\alpha_{sk}\sigma_{cr}\left(\boldsymbol{\theta}_k^\top\boldsymbol{z}\right) + \tilde{\alpha}_{si}\sigma_{cr}\left((\rho\boldsymbol{\theta}_j + \boldsymbol{\theta}_i)^\top\boldsymbol{z}\right) + \tilde{\alpha}_{sj}\sigma_{cr}\left((1-\rho)\boldsymbol{\theta}_j^\top\boldsymbol{z}\right).$$

Let $\tilde{h}_s(\boldsymbol{z}) = h_s(\boldsymbol{z})$. The equation has at least one solution since the degree of freedom (i.e., 4) of this equation is greater than the number of equations (i.e., 2). It implies that for any ρ, there exists a linear mapping ϕ' with ρ' acts upon the connection weights Λ, such that

$$L(\Theta,\Lambda) = L(\tilde{\Theta},\phi'(\Lambda)) = L(\phi(\Theta),\phi'(\Lambda)),$$

where $L(\Theta,\Lambda)$ is a short notation of loss function described in Eq. (9). From Lemma 8, the stacking of ϕ and ϕ' constitutes the minimum generator of $\mathbb{G}$. Hence, the composition of ϕ and ϕ' is the desired minor equioutput transformation.

Based on Proposition 2 and $\partial L(\Theta,\Lambda)/\partial\boldsymbol{\theta}_k = 0$ ($k \in [q]$), it stands for the original networks

$$\frac{1}{N}\sum_{n=1}^N y_n\exp(-y_n[f(\mathbf{W};\boldsymbol{x}_n)]_R)\Delta(\boldsymbol{\theta_k};\boldsymbol{x}_n) = 0,$$

where

$$\Delta(\boldsymbol{\theta_k};\boldsymbol{x}_n) = \left(\frac{\partial[f(\Theta;\boldsymbol{x}_n)]_R}{\partial\,[\boldsymbol{\theta}_k]_R} - [\boldsymbol{\theta}_k]_R[f(\Theta;\boldsymbol{x}_n)]_R\right) + \left(\frac{\partial[f(\Theta;\boldsymbol{x}_n)]_I}{\partial\,[\boldsymbol{\theta}_k]_I} - [\boldsymbol{\theta}_k]_I[f(\Theta;\boldsymbol{x}_n)]_R\right)\mathrm{i}.$$

For the transformed networks, we consider the following cases.
(a) For $k \neq i,j$, we have

$$\partial L(\tilde{\Theta},\phi'(\Lambda))/\partial\tilde{\boldsymbol{\theta}}_k = \partial L(\Theta,\Lambda)/\partial\boldsymbol{\theta}_k = 0.$$

(b) For i, we have

$$\frac{\partial L(\tilde{\Theta},\phi'(\Lambda))}{\partial\tilde{\boldsymbol{\theta}}_i} \propto \frac{1}{N}\sum_{n=1}^N\frac{1}{r}\sum_{s=1}^r y_nL(\Theta,\Lambda)\,\zeta_i,$$

where $\tilde{\boldsymbol{\alpha}}_i$ denotes the i-th row vector of $\phi'(\Lambda)$, and

$$\zeta_i = \frac{\partial[f(\Theta;\boldsymbol{z}_s,\boldsymbol{x}_n)]_R}{\partial\boldsymbol{h}}\tilde{\boldsymbol{\alpha}}_i\frac{\partial\sigma_{cr}(\Theta\boldsymbol{z})}{\partial\boldsymbol{z}}\boldsymbol{z} + \frac{\partial[f(\Theta;\boldsymbol{z}_s,\boldsymbol{x}_n)]_R}{\partial\bar{\boldsymbol{h}}}\bar{\tilde{\boldsymbol{\alpha}}}_i\frac{\partial\overline{\sigma_{cr}(\Theta\boldsymbol{z})}}{\partial\boldsymbol{z}}\boldsymbol{z}.$$

Thus, we have

$$\frac{\partial L(\tilde{\Theta},\phi')}{\partial\tilde{\boldsymbol{\theta}}_i}\begin{cases} = 0, & \text{if} \quad \rho\in\mathbb{R}; \\ \neq 0, & \text{if} \quad \rho\in\mathbb{C} \quad \text{and} \quad [\rho]_I\neq 0. \end{cases}$$

(c) Similarly, for j, it holds

$$\frac{\partial L(\tilde{\Theta},\phi')}{\partial\tilde{\boldsymbol{\theta}}_j}\begin{cases} = 0, & \text{if} \quad \rho\in\mathbb{R}, \\ \neq 0, & \text{if} \quad \rho\in\mathbb{C} \quad \text{and} \quad [\rho]_I\neq 0. \end{cases}$$

Lemma 11 holds as desired when $\Theta = \mathbf{V}^l$ ($l \neq L$). □

From Lemma 11, it is easy to obtain the following lemma.

Lemma 12. *Let* $\mathbf{Q}$ *be the real-valued parameters of a real-valued neural network. If* $\mathbf{Q}$ *is a critical point of* $L(\mathbf{Q})$*, then for any equioutput transformation* $\mathcal{T}$ *with* $\rho\in\mathbb{R}$*,* $\phi(\mathcal{T}(\mathbf{Q}))$ *are critical points.*

Lemma 13. *For any fully-connected feed-forward real-valued neural network with parameter space* Ω_R*, there exists a complex-reaction network with parameter space* Ω_{CR} *such that* $\Omega_{CR} \subseteq \Omega_R$*.*

These lemmas above show that, for any real-valued neural network, we can construct a complex-reaction network such that (i) both networks have the same depth, and (ii) both networks have the same number of parameters for each layer. The detailed proof is given by Appendix B.4.

Proof of Theorem 4. Let $\mathbb{G}_R$ and $\mathbb{G}_{CR}$ denote the groups of the equioutput transformations with real-valued and complex-valued ρ (Definition 1), respectively. Let $\mathcal{G}$ be the optima set, $\mathcal{S}_R$ and $\mathcal{S}_{CR}$ denote the critical point sets of the real-valued and complex-reaction networks, respectively. It suffices to consider the real-valued and complex-reaction networks with the same parameters, generated from Lemma 13, and this follows $\Omega_{CR} \subseteq \Omega_R$.

From Lemma 6, we divide the 4 × 4 matrices in Eq. (B.1) and (B.2) into the block formations by 2 × 2, and find that two block matrices relative to $\boldsymbol{\theta_j}$ are anti-symmetric for $\rho\in\mathbb{C}$ (see Eq. (B.1)), and thus ϕ is a linear mapping from complex manifold to complex manifold. For $\rho\in\mathbb{R}$, two block matrices relative to $\boldsymbol{\theta_j}$ are diagonal (see Eq. (B.2)), which implies that $\phi(\boldsymbol{\theta_j})$ is not on the complex manifold. In other words, the equioutput transformation with real-valued ρ projects a critical point on complex manifold onto the real manifold, specially *almost-complex manifold* (Newlander & Nirenberg, 1957; Wells, 1980). The transformed points are still critical points from Lemma 11(i), whereas any critical point in complex manifold after any equioutput transformation with complex-valued ρ cannot derive new critical points from Lemma 11(ii). On the other hand, Lemma 12 shows that a critical point on real manifold after any equioutput transformation still dwells on the real manifold and derives new critical points.

In summary, the critical point set of real-valued networks is closed to the equioutput transformation with real-valued ρ, where this property does not hold for that of complex-reaction networks. Therefore, we have

$$\begin{cases} \mathcal{G} \subseteq \mathbb{G}_R \circ \mathcal{S}_{CR} \subseteq \mathbb{G}_R \circ \mathcal{S}_R = \mathcal{S}_R \\ \mathcal{G} \subseteq \mathcal{S}_{CR} \subsetneqq \mathbb{G}_{CR} \circ \mathcal{S}_{CR}. \end{cases} \tag{12}$$

From Lemma 8, both groups $\mathbb{G}_R$ and $\mathbb{G}_{CR}$ are isomorphic to a direct product of Weyl groups. Thus, $\mathbb{G}_R$ is isomorphic to $\mathbb{G}_{CR}$, i.e., $\mathbb{G}_R \cong \mathbb{G}_{CR}$. This completes the proof. □

5. Discussions and future issues

Theorems 2 and 4 show the advantages of complex-valued neural networks in comparisons with real-valued ones from the perspectives of approximation complexity and optimization dynamics, respectively. These conclusions benefit from unitary transformation and anti-symmetric multiplication, perhaps two most important characteristics of the complex-valued neural networks. In Theorem 2, we adopt a considerably direct way, i.e., regard the relevance of the radius and phase to the target function as a prior (e.g., radial function) that imposes more constraints on the complex-valued neural network than a real-valued one would, and thus, the unitary transformation that allows radius scaling and phase rotation (see Lemma 2) yields an advantageous reduction of the approximation complexity. This means that merely doubling the number of real-valued parameters (or neurons) in each layer does not give the equivalent effect observed in a complex-valued neural network, which is consistent with the conclusions in Hirose (2003), Mönning and Manandhar (2018). Alternatively, a complex number $z = z_1 + z_2\mathrm{i}$ can be written into a matrix form

$$\begin{pmatrix} z_1 & -z_2 \\ z_2 & z_1 \end{pmatrix} = z_1 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + z_2 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

which yields an anti-symmetric multiplication (Abraham & Marsden, 2008) between the partial derivatives and independent variables, e.g., Eq. (1). Once half of entries is known, the other half is fixed. Such an adjoint relation not only reduces the complexity for approximating rotation operations (Joshua, Qian, & Li, 2021; Wu, Zhang, Jiang, & Zhou, 2021b), but also contributes to escaping the saddle points using complex-valued gradient descents (see Theorem 4).

In light of the preceding merits, we feel the complex-valued neural network has the potential and power of representing the functions with unitary transformation and searching the optimal solution in some optimizations. There are two main future directions. One important is to investigate the theoretical advantages of other related networks. The proposed complex-reaction network is of course a general formulation paradigm of some practicable complex-valued neural networks, such as the deep complex networks (Trabelsi et al., 2018), the flexible transmitter networks (Zhang & Zhou, 2021), etc. Hence, our work provides solid support for designating the theoretical legitimacy (Trabelsi, 2019) and characterization (Wu et al., 2021b) of such network models. Besides, it would be interesting to theoretically study feature space transformation (Zhou, 2021) and width-depth representation (Wu, Jing, Du, & Chen, 2021a; Zhang & Fan, 2021) which might be a key to understand mysteries behind the success of deep neural networks.

Another important is to develop some practical complex-valued modules correspondingly. For example, Theorem 2 shows the power of complex-valued activations on representing radial or equally rotation-invariant functions. So we conjecture that adding complex-valued activations may reduce parameter consumption when suffering from data augmentation techniques, such as image rotation. Besides, from Theorem 4, it is apparent that using complex-valued gradients during the whole training procedure or in stages is conducive to finding the optimal point (Yeats, Chen, & Li, 2021). In the future, it is intriguing and reasonable to explore some practical complex-valued modules, just akin to the gating operation or batch normalization, to reduce complexity and accelerate calculation in deep neural networks.

6. Conclusions

This work theoretically presents the advantages of complex-valued neural networks beyond real-valued ones. We investigate the complex-reaction network with fully-connected feed-forward architecture from the perspectives of approximation and optimization dynamics, and then provide two main conclusions. First, we show that a class of radial functions can be approximated by a complex-reaction network using the polynomial number of parameters, yet cannot be approximated by real-valued networks with exponential parameters. Second, we prove that for practical optimization problems, the critical point set of complex-reaction networks is a proper subset of that of real-valued networks, which may shed some insights on finding optimal solutions more easily for complex-reaction networks. These conclusions not only provide a complement support for complex-valued neural networks, but also exhibit the possibility of developing deep neural network with complex-valued modules.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Science Foundation of China (61921006), the Collaborative Innovation Center of Novel Software Technology and Industrialization [https://www.nju.edu.cn/en/98/99/c7892a170137/page.psp], and the Program B for Outstanding Ph.D Candidate of Nanjing University (202101B051).

Appendix A. Full proofs for Theorem 2

We provided the detailed proofs for Theorems 2.

A.1. Constructions in Theorem 2

Here, we first review the candidate radial function in Eldan and Shamir (2016). Define

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x}),$$

where $\epsilon_i \in \{-1, +1\}$, N is a polynomial function of d,

$$f_i(\boldsymbol{x}) = \begin{cases} \mathbf{1}, & \text{if} \quad B_i = 1, \\ \mathbf{0}, & \text{if} \quad B_i = 0, \end{cases}$$

for B_i's are binary indicators, and

$$\Omega_i = \left[\left(1 + \frac{i-1}{N}\right) C_2\sqrt{2d}, \left(1 + \frac{i}{N}\right) C_2\sqrt{2d} \right], \quad i = 1, \ldots, N.$$

Next, the concerned density function ϕ is the Fourier transform of the indicator of a unit-volume Euclidean ball, that is,

$$\int_{\mathbb{R}^{2d}} \phi^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\mathcal{B}_d} \mathbf{1} \, \mathrm{d}\omega = 1.$$

Thus, we have

$$\phi(\boldsymbol{x}) = \int_{\mathcal{B}_d} 1 \cdot \exp\left(-2\pi i \boldsymbol{x}^\top \boldsymbol{\omega}\right) \, \mathrm{d}\boldsymbol{\omega} \quad \text{with}$$

$$\mathcal{B}_d = \left\{ \boldsymbol{\omega} \,\middle|\, \|\boldsymbol{\omega}\|_2 \le \left(\int_{\mathcal{B}} \mathbf{1} \, \mathrm{d}\boldsymbol{\omega}\right)^{-1} \right\}.$$

The proof of Theorem 1 consists of the following two parts:

(i) of Theorem 1. Let f be the radial function described by Eq. (4). For $C_2, C_3 > 0$ with $d > C_2$, any $\delta > 0$, and any choice of $\epsilon_i \in \{-1, +1\}$ ($i \in [N]$), there exists a complex-reaction network f_{CR} of one-hidden layer with range in $[-2, +2]$ and width at most $C_3 C_{cr}(2d)^{3.75}$, such that

$$\|f(\boldsymbol{x}) - f_{CR}(\boldsymbol{x})\|_{L_2(\mu)} \le \frac{\sqrt{3}}{C_2(2d)^{1/4}} + \delta.$$

(ii) of Theorem 1. For positive constants $C_1, C_2, C_3, \rho, \alpha$ with $2d > C_2$ and $\alpha > C_2$, we define

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x}) \quad \text{and} \quad f_R(\boldsymbol{x}) = \sum_{i=1}^{n_r} \tilde{f}_i(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle),$$

where $N \ge 4C_2\alpha^{3/2}d^2$, $n_r \le C_1 e^{2C_1 d}$, and $\tilde{f}_i : \mathbb{R} \to \mathbb{R}$ are measurable functions with $|f_i(x)| \le C_3(1 + |x|^\rho)$. For any $\delta > 0$, there exists a group of $\epsilon_i \in \{-1, +1\}$ ($i \in [N]$) such that

$$\|f(\boldsymbol{x}) - f_R(\boldsymbol{x})\| \ge \delta/\alpha.$$

A.2. Proof of Lemma 1

From the universal approximation theorems (Cybenko, 1989; Leshno et al., 1993), the shallow real-valued neural networks with general sigmoidal and ReLU activation functions have the universal approximation properties for any continuous functions. Here, we provide the tight bounds for the degree of approximation.

For general sigmoidal activation, Chen (1993) has proved that a L-Lipschitz function f can be approximated by real-valued neural networks of one-hidden layer with $\delta \ge L\omega(f; n_r^{-1})$.

Next, we discuss the ReLU activation. Given $2d = 1$, we have $R \ge \delta/(2L)$; Otherwise, Lemma 1 is trivially satisfied once we force the real-valued network $f_R \equiv 0$. In the case of $R \ge \delta/(2L)$, let $n_0 = \lceil \delta/(2RL) \rceil$. For any L-Lipschitz function $f : [-R, R] \to \mathbb{R}$, we have

$$|g_i(\beta) - f(\beta)| \le \delta,$$

where for $i \in [n_0]$,

$$g_i(x) = g_i(-R) + \frac{g_i(\beta + \delta/(2L)) - g_i(\beta - \delta/(2L))}{\delta/L} \sigma_r(x - \beta) \quad \text{with} \quad \beta = \delta/L.$$

Thus, then we have

$$\sup_{x \in \mathbb{R}} \left| f(x) - g(x) \right| \le \delta,$$

where

$$g(x) = g(-R) + \sum_{i}^{n_0} \frac{g_i(\beta_i + \delta/(2L)) - g_i(\beta_i - \delta/(2L))}{\delta/L} \sigma_r(x - \beta_i),$$

and

$$|g_i(\beta_i) - f(\beta_i)| \le \delta, \quad \text{for} \quad \beta_i = i\delta/L \quad \text{and} \quad i \in [n_0].$$

Provided a real-valued neural network of one-hidden layer with ReLU activation

$$f_R(x) = \sum_{i=1}^{n_r} \alpha_i \, \sigma_r(\beta_i x - b_i) - a,$$

we employ that

$$a = -g(-R), \quad \alpha_i = \frac{g_i(\beta_i + \delta/(2L)) - g_i(\beta_i - \delta/(2L))}{\delta/L}, \quad \beta_i = i\delta/L,$$

and

$$n_r \le n_0 \le \frac{\delta}{2RL},$$

and thus, Lemma 1 holds as desired. □

A.3. Proof of Lemma 4

Define a branch function

$$g_i(\boldsymbol{x}) = \begin{cases} \max\{\mathbf{1}, ND_i\}, & \text{if} \quad B_i = 1, \\ \quad \mathbf{0} \qquad\quad , & \text{if} \quad B_i = 0, \end{cases}$$

with

$$D_i = \min\left\{ \left| \|\boldsymbol{x}\| - \left(1 + \frac{i-1}{N}\right) C_2\sqrt{2d} \right|, \left| \|\boldsymbol{x}\| - \left(1 + \frac{i}{N}\right) C_2\sqrt{2d} \right| \right\}.$$

Let

$$g(\boldsymbol{x}) = \sum_{i=1}^{N} \epsilon_i g_i(\boldsymbol{x}).$$

Since $B_i = 1$ and Ω_i's are disjoint intervals, $g_i(\boldsymbol{x})$ is an N-Lipschitz function. Thus, g is also an N-Lipschitz function. So we have

$$\int_{\mathbb{R}^{2d}} \left(g(\boldsymbol{x}) - \sum_{i=1}^{N} \epsilon_i f_i(\boldsymbol{x}) \right)^2 \phi^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^{2d}} \sum_{i=1}^{N} \epsilon_i^2 \left(g_i(\boldsymbol{x}) - f_i(\boldsymbol{x})\right)^2 \phi^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$= \sum_{i=1}^{N} \int_{\mathbb{R}^{2d}} \left(g_i(\boldsymbol{x}) - f_i(\boldsymbol{x})\right)^2 \phi^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \le (3/(C_2)^2\sqrt{2d}),$$

where the last inequality holds from Eldan and Shamir (2016, Lemma 22). This completes the proof. □

A.4. Proof of Lemma 5

We first list some useful lemmas or propositions.

Proposition 4. *Let $f_R(\boldsymbol{x}) = \sum_{i=1}^{k} \tilde{f}_i(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$, where $\tilde{f}_i : \mathbb{R} \to \mathbb{R}$ are measurable functions satisfying $|f_i(x)| \le C_3(1 + |x|^\rho)$ and ρ is an integer satisfying $\rho \le C_1 e^{2C_1 d}$. If $f_R\phi \in L_2$, then*

$$Supp(\widehat{f_R\phi}) \subseteq \bigcup_{i=1}^{k} \left(span\{\boldsymbol{w}_i\} + \mathcal{B}\right),$$

where $\mathcal{B}$ is a unit ball. Furthermore, there exists a pair of functions (p, q) that satisfies

- *$p \in Supp(\widehat{f_R\phi})$;*
- *q is radial and $\int_B q(\boldsymbol{x})^2 d\boldsymbol{x} \le 1 - \delta$ for some $\delta \in [0, 1]$;*
- *$\|p\|_{L_2} = \|q\|_{L_2} = 1$. Then*

$$1 - \langle p, q \rangle_{L_2} \ge \delta/2 - ke^{-2Cd}.$$

This proposition is proved by Eldan and Shamir (2016).

Proposition 5. *Provided $\|p\|_{L_2} = \|q\|_{L_2} = 1$, for any real-valued scalars $a, b > 0$, we have*

$$\|ap - bq\|_{L_2} \ge \frac{b}{2} \|p - q\|_{L_2}.$$

Proposition 6. *According to Eldan and Shamir (2016, Lemma 11), we have*

$$\|f_R\phi\|_{L_2} = \|f_R\|_{L_2(\mu)} \geq \theta/\alpha .$$

The proofs of Propositions 5, 4, and 6 are shown in Appendix C. Based on the results above, we define the pair (p, q)

$$p = \frac{\widehat{f\phi}}{\|f\phi\|_{L_2}} \quad \text{and} \quad q = \frac{\widehat{f_R\phi}}{\|f_R\phi\|_{L_2}},$$

which satisfy the conditions of Propositions 4 and 5. Thus, we have

$$\begin{aligned}\|f - f_R\|_{L_2(\mu)} &= \| f\phi - f_R\phi \|_{L_2} = \| \widehat{f\phi} - \widehat{f_R\phi} \|_{L_2} \\ &= \| \left(\| f\phi \|_{L_2}\right) p - \left(\| f_R\phi \|_{L_2}\right) q \|_{L_2} \\ &\geq \frac{1}{2} \| p - q \|_{L_2} \| f_R\phi \|_{L_2} \geq \frac{\theta}{2\alpha} \| p - q \|_{L_2} \\ &\geq \frac{\theta}{2\alpha} \sqrt{2(1 - \langle p, q \rangle_{L_2})} \\ &\geq \frac{\theta}{2\alpha} \sqrt{\max\{\delta/2 - n_r e^{-2Cd}, 0\}}.\end{aligned}$$

Provided $n_r \leq C_1 e^{2C_1 d}$ and $C_1 = \min\{\delta/4, C\}$, we have

$$\|f - f_R\|_{L_2(\mu)} \geq \frac{\theta\sqrt{\delta}}{4\alpha},$$

and

$$\sqrt{\max\{\delta/2 - n_r e^{-2Cd}, 0\}} \geq \sqrt{\delta/4}.$$

This completes the proof. □

Appendix B. Full proofs for Theorem 4

We provided the detailed proofs for Theorem 4.

B.1. Proof of Lemma 6

For i, j and $\rho \in \mathbb{R}$, let $\boldsymbol{v}_i = \rho\boldsymbol{\theta}_j + \boldsymbol{\theta}_i$ and $\boldsymbol{v}_j = (1 - \rho)\boldsymbol{\theta}_j$, then we have

$$\begin{aligned}\begin{pmatrix} [\boldsymbol{v}_i]_R \\ [\boldsymbol{v}_i]_I \\ [\boldsymbol{v}_j]_R \\ [\boldsymbol{v}_j]_I \end{pmatrix} &= \begin{pmatrix} [\boldsymbol{\theta}_i]_R \\ [\boldsymbol{\theta}_i]_I \\ 0 \\ 0 \end{pmatrix} + \rho \begin{pmatrix} [\boldsymbol{\theta}_j]_R \\ [\boldsymbol{\theta}_j]_I \\ 0 \\ 0 \end{pmatrix} + (1-\rho) \begin{pmatrix} 0 \\ 0 \\ [\boldsymbol{\theta}_j]_R \\ [\boldsymbol{\theta}_j]_I \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & \rho & 0 \\ 0 & 1 & 0 & \rho \\ 0 & 0 & (1-\rho) & 0 \\ 0 & 0 & 0 & (1-\rho) \end{pmatrix} \begin{pmatrix} [\boldsymbol{\theta}_i]_R \\ [\boldsymbol{\theta}_i]_I \\ [\boldsymbol{\theta}_j]_R \\ [\boldsymbol{\theta}_j]_I \end{pmatrix}.\end{aligned} \tag{B.1}$$

For the case $\rho = \rho_1 + \rho_2 \mathrm{i} \in \mathbb{C}$, one has

$$\begin{pmatrix} [\boldsymbol{v}_i]_R \\ [\boldsymbol{v}_i]_I \\ [\boldsymbol{v}_j]_R \\ [\boldsymbol{v}_j]_I \end{pmatrix} = \begin{pmatrix} 1 & 0 & \rho_1 & -\rho_2 \\ 0 & 1 & \rho_2 & \rho_1 \\ 0 & 0 & (1-\rho_1) & \rho_2 \\ 0 & 0 & -\rho_2 & (1-\rho_1) \end{pmatrix} \begin{pmatrix} [\boldsymbol{\theta}_i]_R \\ [\boldsymbol{\theta}_i]_I \\ [\boldsymbol{\theta}_j]_R \\ [\boldsymbol{\theta}_j]_I \end{pmatrix}. \tag{B.2}$$

So each linear mapping $\phi(\Theta; \rho, i, j)$ leads to a straight line in an 2-dimensional affine space. This completes the proof. □

B.2. Proof of Lemma 7

The existence of real-valued equioutput transformations is proved by Chen, Lu, and Hecht-Nielsen (1993). The proof sketch of the complex-valued equioutput transformations is similar. There are three facts that (1) zReLU is even on the complex-valued domain, that is, $\mathrm{zReLU}(-\boldsymbol{z}) = \mathrm{zReLU}(\boldsymbol{z})$; (2) for $l \neq L$ and any i, j, ρ_1, ρ_2, there exists a pair of scalar parameters $\tilde{\alpha}_i$ and $\tilde{\alpha}_j$ in the next layer (i.e., $(l+1)$-th layer) such that

$$L\left(\phi_2 \circ \phi_1(\Theta^l), \tilde{\alpha}_i, \tilde{\alpha}_j\right) = L(\Theta^l, \alpha_i, \alpha_j).$$

(3) $(\tilde{\alpha}_i, \tilde{\alpha}_j)$ is led by a composition $\phi'_1 \circ \phi'_2$ of other linear mappings for some ρ'_1, ρ'_2.

Next, we are going to prove facts (2) and (3). Provided the parameters $\{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \alpha_i, \alpha_j\}$ relative to neurons i, j, we have

$$\begin{aligned}h(\boldsymbol{z}) &= \alpha_i \sigma_{cr}(\boldsymbol{\theta}_i^\top \boldsymbol{z}) + \alpha_j \sigma_{cr}(\boldsymbol{\theta}_j^\top \boldsymbol{z}) \\ &= \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \\ \mathrm{i} \end{pmatrix}^\top \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} [\alpha_i^\top z_i]_R \\ [\alpha_i^\top z_i]_I \\ [\alpha_j^\top z_j]_R \\ [\alpha_j^\top z_j]_I \end{pmatrix},\end{aligned}$$

where σ_{cr} denotes the element-wise activation, and

$$\begin{pmatrix} [z_i]_R \\ [z_i]_I \\ [z_j]_R \\ [z_j]_I \end{pmatrix} = \sigma_{cr} \circ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} [\boldsymbol{\theta}_i^\top \boldsymbol{x}]_R \\ [\boldsymbol{\theta}_i^\top \boldsymbol{x}]_I \\ [\boldsymbol{\theta}_j^\top \boldsymbol{x}]_R \\ [\boldsymbol{\theta}_j^\top \boldsymbol{x}]_I \end{pmatrix}.$$

Let $\rho_1 = \rho_{11} + \rho_{12}\mathrm{i}$, $\rho_2 = \rho_{21} + \rho_{22}\mathrm{i}$, $\rho'_1 = \rho'_{11} + \rho'_{12}\mathrm{i}$, and $\rho'_2 = \rho'_{21} + \rho'_{22}\mathrm{i}$. According to Eq. (B.2), one has

$$\begin{aligned}\tilde{h}(\boldsymbol{x}) &= \tilde{\alpha}_i \sigma_{cr}((\rho\boldsymbol{\theta}_j + \boldsymbol{\theta}_i)^\top \boldsymbol{x}) + \tilde{\alpha}_j \sigma_{cr}((1-\rho)\boldsymbol{\theta}_j^\top \boldsymbol{x}) \\ &= \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \\ \mathrm{i} \end{pmatrix}^\top \begin{pmatrix} 1 & 0 & (1-\rho'_{21}) & \rho'_{22} \\ 0 & 1 & -\rho'_{22} & (1-\rho'_{21}) \\ 0 & 0 & \rho'_{21} & -\rho'_{22} \\ 0 & 0 & \rho'_{22} & \rho'_{21} \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & 0 & \rho'_{11} & -\rho'_{12} \\ 0 & 1 & \rho'_{12} & \rho'_{11} \\ 0 & 0 & (1-\rho'_{11}) & \rho'_{12} \\ 0 & 0 & -\rho'_{12} & (1-\rho'_{11}) \end{pmatrix} \begin{pmatrix} [\alpha_i^\top z'_i]_R \\ [\alpha_i^\top z'_i]_I \\ [\alpha_j^\top z'_j]_R \\ [\alpha_j^\top z'_j]_I \end{pmatrix},\end{aligned}$$

where

$$\begin{aligned}\begin{pmatrix} [z'_i]_R \\ [z'_i]_I \\ [z'_j]_R \\ [z'_j]_I \end{pmatrix} &= \sigma_{cr} \circ \begin{pmatrix} 1 & 0 & (1-\rho_{21}) & \rho_{22} \\ 0 & 1 & -\rho_{22} & (1-\rho_{21}) \\ 0 & 0 & \rho_{21} & -\rho_{22} \\ 0 & 0 & \rho_{22} & \rho_{21} \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & 0 & \rho_{11} & -\rho_{12} \\ 0 & 1 & \rho_{12} & \rho_{11} \\ 0 & 0 & (1-\rho_{11}) & \rho_{12} \\ 0 & 0 & -\rho_{12} & (1-\rho_{11}) \end{pmatrix} \begin{pmatrix} [\boldsymbol{\theta}_i^\top \boldsymbol{x}]_R \\ [\boldsymbol{\theta}_i^\top \boldsymbol{x}]_I \\ [\boldsymbol{\theta}_j^\top \boldsymbol{x}]_R \\ [\boldsymbol{\theta}_j^\top \boldsymbol{x}]_I \end{pmatrix}.\end{aligned}$$

Let $\tilde{h}(\boldsymbol{x}) = h(\boldsymbol{x})$, we obtain a semi-linear equation with eight free parameters. This equation has at least one solution, which completes the proof of facts (2) and (3). Lemma 7 holds as desired. □

B.3. Proof of Lemma 8

Write the parameter space as $\Theta^1 \times \cdots \times \Theta^L$, where Θ^l denotes the subspace concerning the l-th layer. Let $\mathbb{G}_l$ denote the set of linear mapping ϕ upon Θ^l. According to Lemmas 6 and 7, $\mathbb{G}_l$ forms a cube symmetry group, that is, isomorphic to the Weyl group (Warner, 1983; Weyl, 1946). So the equioutput transformation, i.e., group action upon each hidden layer (except the case $l = L$) can be regarded as the direct operation of $\mathbb{G}_l$ on the corresponding subspace Θ^1 and as indirect but isomorphic operation led by some sequence $\{\phi\}$. According to Lemma 7 and the fact that the hidden layers only have symmetry groups associated with themselves, each hidden layer contributes exactly one cube symmetry group to the overall group action. In other words, $\mathbb{G}_l$ 1–1 corresponds to Θ^l. Thus, group $\mathbb{G}$ is isomorphic to the direct product of these groups

$$\mathbb{G} \cong \mathbb{G}_1 \times \cdots \times \mathbb{G}_{L-1}, \tag{B.3}$$

since the actions of the individual groups operating on different layers commute. Next, we are going to bound the order of $\mathbb{G}$, denoted as $|\mathbb{G}|_\#$. Based on Eq. (B.3), we have

$$|\mathbb{G}|_\# = \prod_{l=1}^{L-1} |\mathbb{G}_l|_\#.$$

Suppose that the l-th layer has n_l neurons, then there are $n_l!$ different pairs (i,j) in this layer. Further, if the equation $L(\Theta^l, \alpha_i, \alpha_j) = L(\phi(\Theta^l, \rho, i, j), \tilde{\boldsymbol{\alpha}}_i, \tilde{\boldsymbol{\alpha}}_j)$ in the proof of Lemma 7, led by the minimum generator, has finite solution, then from the Weyl group theory, the order of the group $|\mathbb{G}_l|_\#$ is bounded by $n_l! 2^{n_l}$. Otherwise, $\mathbb{G}$ is an infinite group. This completes the proof. □

B.4. Proof of Lemma 13

For real-valued neural networks, let p and q denote the neuron numbers of some adjacent layers. According to Section 2, there are pq parameters. Consider the following cases. (1) Both p, q are even, that is, $p = 2n$ and $q = 2m$. We construct a complex-reaction network with two layers, where the first and second layers have $2n$ and m neurons, respectively. Then the constructed complex-reaction network has $2(2n)m$ real-valued parameters. (2) When $p = 2n+1$ and $q = 2m$, we can construct a complex-reaction network with two layers, where the first and second layers have p and m neurons, respectively. Then the constructed complex-reaction network has $2pm$ real-valued parameters. (3) When $p = 2n+1$ and $q = 2m+1$, we can construct a complex-reaction network as follows: (3a) this network consists of two layers where the first and second layers have $n+1$ and $2m+1$ neurons; (3b) we force imaginary part of the last neuron of the first layer to be zero. Thus, half of the connection weights that link this neuron and $2m+1$ neurons of the second layer are useless. Then the constructed complex-reaction network has $2(n+1)(2m+1) - (2m+1)$ real-valued parameters.

Summing up the above, for any fully-connected feed-forward real-valued neural network, we can construct a complex-reaction network, which has the same parameter structure with the real-valued one. According to the differential manifold theory, the coordinate transformation of the complex manifold needs to satisfy the holomorphic condition. So the complex manifold led by the parameters of the constructed network is a subset of the real manifold led by that of the real-valued one, that is,

$$\Omega_{CR}^l \subseteq \Omega_R^l,$$

where the superscript l denotes the l-th layer. According to Lemma 8, we can write the connection weights of each network as $\{\Theta^1 \dots \Theta^L\}$, where Θ^l denotes the connection weight matrix concerning the l-th layer. Thus, the parameter space of each network is a direct product of the sub-manifolds led by each layer, that is,

$$\Omega_{CR} = \Omega_{CR}^1 \times \cdots \times \Omega_{CR}^L \quad \text{and} \quad \Omega_R = \Omega_R^1 \times \cdots \times \Omega_R^L.$$

Finally, we have

$$\Omega_{CR} \subseteq \Omega_R.$$

This completes the proof. □

B.5. Full proof of Proposition 2

We begin our proof with some useful lemmas.

Lemma 14 (*Normalization Matrix*)**.** *Let $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{1\times n}$ with $\boldsymbol{w} = \gamma \boldsymbol{v}$ and $\|\boldsymbol{v}\| = 1$. Define $\mathbf{S} = \mathbf{I}_{n\times n} - \boldsymbol{v}^\top \boldsymbol{v}$, then we have*

$$(1)\ \mathbf{S} = \mathbf{I}_{n\times n} - \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}; \quad (2)\ \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{w}} = \frac{\mathbf{S}}{\gamma}; \quad (3)\ \mathbf{S}\boldsymbol{w}^\top = \mathbf{S}\boldsymbol{v}^\top = \mathbf{0}; \quad (4)\ \mathbf{S}^2 = \mathbf{S}.$$

The detailed proof of Lemma 9 is given by Appendix C.1.

Lemma 15 (*Weight Norms*)**.** *During the gradient descent procedure, the change rate of $\|\gamma_j^l\|$ (i.e., weight norms) is the same for each layer.*

Lemma 10 shows that $(\gamma_j^l)^2 = \|\mathbf{W}_j^l\|^2$ grows at a rate independent of the row j and layer l. Thereby, using gradient descent to solve the optimization problem, there is no difference in the change rate of connection weights layer by layer. This result also holds for real-valued networks. The detailed proof is presented by Appendix C.2.

Here, we study the following minimization optimization problem of complex-reaction networks.

$$\min_{\mathbf{V}_j^l, \gamma_j^l} L(\mathbf{W}; \mathbf{X}) = \frac{1}{N} \sum_{n=1}^N \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right),$$
$$s.t. \quad \mathbf{W}_j^l = \gamma_j^l\, \mathbf{V}_j^l, \quad \|\mathbf{V}_i^l\| = 1.$$

Solving this problem by the standard gradient descent, the optimization procedure concerning $\mathbf{W}^l$ induces the following dynamical system:

$$\frac{d\mathbf{W}^l}{dt} = -\frac{\partial L}{\partial \mathbf{W}^l} = \frac{1}{N} \sum_{n=1}^N \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \frac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \mathbf{W}^l}.$$

Based on the Euler's theorem for homogeneous functions, we have

$$\begin{aligned} [f(\mathbf{W}; \mathbf{X})]_R &= \mathbf{W}_i^l \frac{\partial [f(\mathbf{W}; \mathbf{X})]_R}{\partial \mathbf{W}_i^l} \\ &= \mathbf{W}_j^l \left[\frac{\partial [f(\mathbf{W}; \mathbf{X})]_R}{\partial \mathbf{W}_j^l} + \frac{\partial [f(\mathbf{W}; \mathbf{X})]_R}{\partial \mathbf{W}_j^l}\mathrm{i} \right] \end{aligned}$$

and

$$\frac{\partial f(\mathbf{W}^l)}{\partial \mathbf{W}^l} \propto \frac{\partial f(\mathbf{V}^l)}{\partial \mathbf{V}^l}.$$

The formula above implies that the standard gradient descents of non-normalized and normalized complex-reaction networks are proportional. Thus, the gradient descent procedure with weight normalization induces the following dynamical systems

$$\begin{cases} \dfrac{d\gamma^l}{dt} = -\dfrac{\partial L}{\partial \gamma^l} = \displaystyle\sum_{n=1}^N \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \frac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \gamma^l}, \\ \dfrac{d\mathbf{V}^l}{dt} = -\dfrac{\partial L}{\partial \mathbf{V}^l} = \displaystyle\sum_{n=1}^N \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) y_n \frac{\partial [f(\mathbf{W}; \boldsymbol{x}_n)]_R}{\partial \mathbf{V}^l}. \end{cases} \tag{B.4}$$

Let $\boldsymbol{w}_j = ([\mathbf{W}_j^l]_R, -[\mathbf{W}_j^l]_I)$, $\boldsymbol{v}_j = ([\mathbf{V}_j^l]_R, -[\mathbf{V}_j^l]_I)$, and $\mathbf{S}_j = \mathbf{I} - \boldsymbol{v}_j^\top \boldsymbol{v}_j$. It is observed that if $[\mathbf{W}_j^l]_R$ has d elements, $\boldsymbol{w}_j$ is a $2d$-dimensional row vector. According to Lemma 9, one has

$$\mathbf{S}_j = \mathbf{I} - \frac{\boldsymbol{w}_j^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2^2} \quad \text{and} \quad \mathbf{S}_j \boldsymbol{w}_j^\top = \mathbf{S}_j \boldsymbol{v}_j^\top = \mathbf{0}.$$

Thus, we have

$$\frac{d\gamma_j^l}{dt} = \frac{d\|\boldsymbol{w}_j\|}{dt} = \boldsymbol{v}_j \left(\frac{d\boldsymbol{w}_j}{dt}\right)^\top \quad \text{and} \quad \frac{d\boldsymbol{v}_j}{dt} = \frac{\mathbf{S}_j}{\gamma_j^l} \left(\frac{d\boldsymbol{w}_j}{dt}\right)^\top.$$

So Eq. (B.4) becomes

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{\gamma_j^l} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \boldsymbol{v}_j \boldsymbol{u}_j^\top, \\ \dfrac{d\boldsymbol{v}_j}{dt} = \dfrac{\eta}{(\gamma^l)^2} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \left(\boldsymbol{u}_j - \boldsymbol{v}_j [f(\mathbf{V}; \boldsymbol{x}_n)]_R\right), \end{cases}$$

where

$$\boldsymbol{u}_j = \left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_R}, -\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_I} \right).$$

Due to

$$\begin{aligned}\boldsymbol{v}_j \boldsymbol{u}_j^\top &= [\mathbf{V}_j^l]_R \left[\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_R} \right]^\top + [\mathbf{V}_j^l]_I \left[\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_I} \right]^\top \\ &= [f(\mathbf{V}; \boldsymbol{x}_n)]_R,\end{aligned}$$

we have

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{\gamma_j^l} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) [f(\mathbf{V}; \boldsymbol{x}_n)]_R, \\ \dfrac{d\boldsymbol{v}_j}{dt} = \dfrac{\eta}{\left(\gamma_j^l\right)^2} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \left(\boldsymbol{u}_j - \boldsymbol{v}_j [f(\mathbf{V}; \boldsymbol{x}_n)]_R\right). \end{cases} \tag{B.5}$$

Let $\mathbf{I}_\mathbb{C} = (\mathbf{I}_{d\times d}, -\mathbf{I}_{d\times d}\mathrm{i})^T$ and multiply it by Eq. (B.5). Thus, we can obtain the gradient descent dynamics concerning the normalized network

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{\gamma_j^l} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) [f(\mathbf{V}; \boldsymbol{x}_n)]_R, \\ \dfrac{d\mathbf{V}_j^l}{dt} = \dfrac{d\boldsymbol{v}_j}{dt} \mathbf{I}_\mathbb{C} = \dfrac{\eta}{\left(\gamma_j^l\right)^2} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n [f(\mathbf{W}; \boldsymbol{x}_n)]_R\right) \varDelta_j, \end{cases}$$

where

$$\begin{aligned}\varDelta_j = &\left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_R}{\partial \left[\mathbf{V}_j^l\right]_R} - [\mathbf{V}_j^l]_R [f(\mathbf{V}; \boldsymbol{x}_n)]_R \right) \\ &+ \left(\frac{\partial [f(\mathbf{V}; \boldsymbol{x}_n)]_I}{\partial \left[\mathbf{V}_j^l\right]_I} - [\mathbf{V}_j^l]_I [f(\mathbf{V}; \boldsymbol{x}_n)]_R \right) \mathrm{i}.\end{aligned}$$

This completes the proof. □

B.6. Proof of Proposition 3

Consider a real-valued neural network $f_R : \mathbb{R}^{2d} \to \{-1, +1\}$ with ReLU activation and weight normalization $\mathbf{P}_j^l = \gamma_j^l \mathbf{Q}_j^l$, where $\gamma_j^l \in \mathbb{R}^+$ and $\|\mathbf{Q}_j^l\| = 1$. Given a training set $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ with $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, we employ standard gradient descents to minimize the empirical exponential loss.

First, we should introduce some necessary facts. The ReLU function has the real-valued homomorphism property (Poggio et al., 2020), that is, for any $x \in \mathbb{R}$ and $\alpha \geq 0$, the equation holds $\sigma_r(\alpha x) = \alpha \sigma_r(x)$. Thus, we have

$$\sigma_r(x) = \frac{\partial \sigma_r(x)}{\partial x} x,$$

which implies

$$f_R(\mathbf{P}; \mathbf{X}) = \mathbf{P}_j^l \left(\frac{\partial f_R(\mathbf{P}; \mathbf{X})}{\partial \mathbf{P}_j^l} \right)^\top \quad \text{and} \quad f_R(\mathbf{Q}; \mathbf{X}) = \mathbf{Q}_j^l \left(\frac{\partial f_R(\mathbf{Q}; \mathbf{X})}{\partial \mathbf{Q}_j^l} \right)^\top.$$

The optimization procedure concerning $\mathbf{P}^l$ is led by the following dynamical systems

$$\frac{d\mathbf{P}^l}{dt} = -\frac{\partial L}{\partial \mathbf{P}^l} = \frac{1}{N} \sum_{n=1}^N \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) y_n \frac{\partial f_R(\mathbf{P}; \boldsymbol{x}_n)}{\partial \mathbf{P}^l}.$$

The gradient descent procedure with weight normalization induces the following dynamical systems

$$\begin{cases} \dfrac{d\gamma^l}{dt} = -\dfrac{\partial L}{\partial \gamma^l} = \displaystyle\sum_{n=1}^N \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) y_n \dfrac{\partial f_R(\mathbf{P}; \boldsymbol{x}_n)}{\partial \gamma^l}, \\ \dfrac{d\mathbf{V}^l}{dt} = -\dfrac{\partial L}{\partial \mathbf{Q}^l} = \displaystyle\sum_{n=1}^N \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) y_n \dfrac{\partial f_R(\mathbf{P}; \boldsymbol{x}_n)}{\partial \mathbf{Q}^l}. \end{cases} \tag{B.6}$$

Similar to the proof of Proposition 2, we use the vectorized representation, i.e., denote $\boldsymbol{w}_j = \mathbf{P}_j^l$ and $\boldsymbol{v}_j = \mathbf{Q}_j^l$. Define a matrix

$$\mathbf{S}_j = \mathbf{I} - \boldsymbol{v}_j^\top \boldsymbol{v}_j = \mathbf{I} - \frac{\boldsymbol{w}_j^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_j^\top \boldsymbol{w}_j\|}.$$

According to Lemma 9, we have

$$\frac{d\gamma_j^l}{dt} = \frac{d\|\boldsymbol{w}_j\|}{dt} = \boldsymbol{v}_j \left(\frac{d\boldsymbol{w}_j}{dt} \right)^\top \quad \text{and} \quad \frac{d\boldsymbol{v}_j}{dt} = \frac{\mathbf{S}_j}{\gamma_j^l} \left(\frac{d\boldsymbol{w}_j}{dt} \right)^\top, \quad \text{respectively.}$$

Thus, Eq. (B.6) becomes

$$\begin{cases} \dfrac{d\gamma_j^l}{dt} = \dfrac{\eta}{\gamma_j^l} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) \mathbf{Q}_j^l \left(\dfrac{\partial f_R(\mathbf{Q}; \mathbf{X})}{\partial \mathbf{Q}_j^l} \right)^\top \\ \qquad = \dfrac{\eta}{\gamma_j^l} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) \mathbf{Q}_j^l \left(\dfrac{\partial f_R(\mathbf{Q}; \mathbf{X})}{\partial \mathbf{Q}_j^l} \right)^\top, \\ \dfrac{d\boldsymbol{v}_j}{dt} = \dfrac{\eta}{\left(\gamma^l\right)^2} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) \mathbf{S}_j \dfrac{\partial f_R(\mathbf{Q}; \mathbf{X})}{\partial \mathbf{Q}_j^l} \\ \qquad = \dfrac{\eta}{\left(\gamma^l\right)^2} \dfrac{1}{N} \displaystyle\sum_{n=1}^N y_n \exp\left(-y_n f_R(\mathbf{P}; \boldsymbol{x}_n)\right) \\ \qquad\quad \times \left(\dfrac{\partial f_R(\mathbf{Q}; \mathbf{X})}{\partial \mathbf{Q}_j^l} - \mathbf{Q}_j^l f_R(\mathbf{Q}; \boldsymbol{x}_n) \right), \end{cases}$$

where η is a strictly positive constant relative to γ_j^l, which satisfies

$$f(\mathbf{P}; \mathbf{X}) = \eta f(\mathbf{Q}; \mathbf{X}) \quad \text{and} \quad \frac{\partial f(\mathbf{P}_j^l)}{\partial \mathbf{P}_j^l} = \frac{\eta}{\gamma_j^l} \frac{\partial f(\mathbf{Q}_j^l)}{\partial \mathbf{Q}_j^l}, \quad \text{respectively.}$$

This completes the proof. □

Appendix C. Complete proofs for useful lemmas

This section completes the proofs of some useful lemmas in Appendix A.

C.1. Proof of Lemma 9

Let $\boldsymbol{w} = (w_1, \ldots, w_n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$. Since $\boldsymbol{w} = \gamma \boldsymbol{v}$ and $\|\boldsymbol{v}\| = 1$, we have

$$\|\boldsymbol{v}\| = \sqrt{(v_1)^2 + \cdots + (v_n)^2} = 1 \quad \text{and} \quad w_i = \gamma v_i \quad \text{for any} \quad i \in [n].$$

Let

$$\mathbf{S} = \mathbf{I}_{n\times n} - \boldsymbol{v}^\top \boldsymbol{v} = \begin{pmatrix} 1-(v_1)^2 & -v_1 v_2 & \cdots & -v_1 v_n \\ -v_2 v_1 & 1-(v_2)^2 & \cdots & -v_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ -v_n v_1 & -v_n v_2 & \cdots & 1-(v_n)^2 \end{pmatrix}.$$

On the other hand, we have

$$\mathbf{I}_{n\times n} - \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} = \mathbf{I}_{n\times n} - \frac{1}{\|\boldsymbol{w}\|_2^2} \begin{pmatrix} (w_1)^2 & w_1 w_2 & \cdots & w_1 w_n \\ w_2 w_1 & (w_2)^2 & \cdots & w_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n w_1 & w_n w_2 & \cdots & (w_n)^2 \end{pmatrix}.$$

For $i, j \in [n]$, one has

$$\begin{aligned}1-\frac{(w_i)^2}{(w_1)^2+\cdots+(w_n)^2} &= \frac{\sum_{k\neq i}(w_k)^2}{\sum_k (w_k)^2} \\ &= \frac{\sum_{k\neq i}(v_k)^2}{\sum_k (v_k)^2} = 1-\frac{(v_i)^2}{(v_1)^2+\cdots+(v_n)^2} \\ &= 1-(v_i)^2,\end{aligned}$$

and

$$-\frac{w_iw_j}{(w_1)^2+\cdots+(w_n)^2} = -\frac{v_iv_j}{(v_1)^2+\cdots+(v_n)^2} = -v_iv_j.$$

Thus, we have

$$\mathbf{S} = \mathbf{I}_{n\times n} - \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}.$$

Consider the partial derivative of $\boldsymbol{v}$ with respect to $\boldsymbol{w}$

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{w}} = \begin{pmatrix} \partial v_1/\partial w_1 & \partial v_1/\partial w_2 & \cdots & \partial v_1/\partial w_n \\ \partial v_2/\partial w_1 & \partial v_2/\partial w_2 & \cdots & \partial v_2/\partial w_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial v_n/\partial w_1 & \partial v_n/\partial w_2 & \cdots & \partial v_n/\partial w_n \end{pmatrix}.$$

Thus, we have

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{w}} = \frac{\mathbf{S}}{\gamma}.$$

It is easily to verify that 0 is an eigenvalue of the matrix $\mathbf{S}$ and $\boldsymbol{v}$ is the corresponding eigenvector. So we have

$$\mathbf{S}\boldsymbol{w}^\top = \mathbf{S}\boldsymbol{v}^\top = \mathbf{0}.$$

Let $\mathbf{S}_i$ denote the i-th row vector of matrix $\mathbf{S}$. Thus, we have

$$\begin{aligned}\mathbf{S}_i\,(\mathbf{S}_i)^\top &= \sum_{k\neq i}(v_iv_k)^2 + \left(1-(v_i)^2\right)^2 = \sum_{k\neq i}(v_iv_k)^2 + \left(\sum_{k\neq i}(v_k)^2\right)^2 \\ &= 1-(v_i)^2,\end{aligned}$$

and

$$\begin{aligned}\mathbf{S}_i\left(\mathbf{S}_j\right)^\top &= \sum_{k\neq i,j}(v_iv_j)(v_k)^2 + \left(1-(v_i)^2\right)v_jv_i + v_iv_j\left(1-(v_j)^2\right) \\ &= \sum_{k\neq i,j}(v_iv_j)(v_k)^2 + v_iv_j\left(2\sum_{k\neq i,j}(v_k)^2 + (v_i)^2 + (v_j)^2\right) \\ &= -v_iv_j.\end{aligned}$$

Thus, we have $\mathbf{S}^2 = \mathbf{S}$. This completes the proof. □

C.2. Proof of Lemma 10

Observing the change rate of γ_j^l in Proposition 2, we have

$$\frac{d\left(\gamma_j^l\right)^2}{dt} = 2\,\gamma_j^l\frac{d\gamma_j^l}{dt} = 2\frac{\eta}{N}\sum_{n=1}^{N}\exp\left(-y_n[f(\mathbf{W};\boldsymbol{x}_n)]_R\right)\ [f(\mathbf{V};\boldsymbol{x}_n)]_R.$$

Thus, we have

$$\left\|\frac{d\mathbf{W}_j^l}{dt}\right\| = \frac{\partial\|\mathbf{W}_j^l\|}{\partial\mathbf{W}_j^l}\frac{d\mathbf{W}_j^l}{dt},$$

and then

$$\left\|\frac{d\mathbf{W}_j^l}{dt}\right\|^2 = \frac{2}{N}\sum_{n=1}^{N}\exp\left(-y_n[f(\mathbf{W};\boldsymbol{x}_n)]_R\right)[f(\mathbf{W};\boldsymbol{x}_n)]_R.$$

So the change rate of $\|\mathbf{W}_j^l\|^2$ is independent of the layer index l and row index j. Finally, it is easily to verify that these results above also hold for Proposition 3. □

References

Abraham, R., & Marsden, J. E. (2008). Foundations of mechanics. (364), American Mathematical Soc..

Adali, T., Schreier, P., & Scharf, L. (2011). Complex-valued signal processing: The proper way to deal with impropriety. *IEEE Transactions on Signal Processing*, *59*(11), 5101–5125.

Allen-Zhu, Z., Li, Y., & Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th international conference on machine learning* (pp. 242–252).

Arena, P., Fortuna, L., Re, R., & Xibilia, M. (1993). On the capability of neural networks with complex neurons in complex valued functions approximation. In *Proceedings of the 1993 international symposium on circuits and systems* (pp. 2168–2171).

Arena, P., Fortuna, L., Re, R., & Xibilia, M. (1995). Multilayer perceptrons to approximate complex valued functions. *International Journal of Neural Systems*, *6*(04), 435–446.

Arora, S., Du, S., Hu, W., Li, Z., & Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th international conference on machine learning* (pp. 322–332).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd international conference on learning representations*.

Barron, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, *14*(1), 115–133.

Burkard, A., Zimmermann, G., & Schwarzer, B. (2021). Monitoring systems for checking websites on accessibility. *Frontiers in Computer Science*, *2*.

Chen, D. (1993). Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and Its Applications*, *9*(3), 17–28.

Chen, A., Lu, H., & Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation*, *5*(6), 910–927.

Cybenko, George (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., & Graves, A. (2016). Associative long short-term memory. In *Proceedings of the 33rd international conference on machine learning* (pp. 1986–1994).

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems, vol. 27* (pp. 2933–2941).

Du, Simon, Lee, Jason, Li, Haochuan, Wang, Liwei, & Zhai, Xiyu (2019). Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th international conference on machine learning* (pp. 1675–1685).

Du, S., Zhai, X., Poczos, B., & Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. In *Proceedings of the 6th international conference on learning representations*.

Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. In *Proceedings of the 29th annual conference on learning theory* (pp. 907–940).

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*(3), 183–192.

Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649).

Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd international conference on machine learning* (pp. 1225–1234).

Hirose, A. (2003). *Complex-valued neural networks: Theories and applications, vol. 5*. World Scientific.

Hirose, A. (2012). *Complex-valued Neural Networks, vol. 400*. Springer.

Hirose, A., & Yoshida, S. (2011). Comparison of complex-and real-valued feed-forward neural networks in their generalization ability. In *Proceedings of the 18th international conference on neural information processing* (pp. 526–531).

Hirose, A., & Yoshida, S. (2012). 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.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems, vol. 31* (pp. 8571–8580).

Joshua, B., Qian, L., & Li, X. (2021). A survey of complex-valued neural networks. arXiv:2101.12249.

Kidger, P., & Lyons, T. (2020). Universal approximation with deep narrow networks. In *Proceedings of the 33rd annual conference on learning theory* (pp. 2306–2327).

Koenderink, J., van Doorn, A., & Gegenfurtner, K. (2021). Rgb colors and ecological optics. *Frontiers in Computer Science*, *3*.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems, vol. 25* (pp. 1097–1105).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Leshno, M., Lin, V., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems, vol. 31* (pp. 6232–6240).

Mönning, N., & Manandhar, S. (2018). Evaluation of complex-valued neural networks on real-valued classification tasks. arXiv:1811.12351.

Newlander, A., & Nirenberg, L. (1957). Complex analytic coordinates in almost complex manifolds. *Annals of Mathematics*, 391–404.

Nitta, T. (2002). On the critical points of the complex-valued neural network. In *Proceedings of the 9th international conference on neural information processing, vol. 3* (pp. 1099–1103).

Nitta, T. (2013). Local minima in hierarchical structures of complex-valued neural networks. *Neural Networks*, *43*(2013), 1–7.

Oyallon, E., & Mallat, S. (2015). Deep roto-translation scattering for object classification. In *Proceedings of the 28th conference on computer vision and pattern recognition* (pp. 2865–2873).

Poggio, T., Banburski, A., & Liao, Q. (2020). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, *117*(48), 30039–30045.

Sun, S., Chen, W., Wang, L., Liu, X., & Liu, T.-Y. (2016). On the depth of deep neural networks: A theoretical view. In *Proceedings of the 30th AAAI conference on artificial intelligence* (pp. 2066–2072).

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems, vol. 27* (pp. 3104–3112).

Trabelsi, C. (2019). *Stabilizing and enhancing learning for deep complex and real neural networks*. (Ph.D. thesis), Ecole Polytechnique, Montreal (Canada).

Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J., et al. (2018). Deep complex networks. In *Proceedings of the 6th international conference on learning representations*.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In *Proceedings of the 33rd international conference on machine learning* (pp. 2071–2080).

Tygert, M., Bruna, J., Chintala, S., LeCun, Y., Piantino, S., & Szlam, A. (2016). A mathematical motivation for complex-valued convolutional networks. *Neural Computation*, *28*(5), 815–825.

Virtue, P., Stella, X., & Lustig, M. (2017). Better than real: Complex-valued neural nets for MRI fingerprinting. In *Proceedings of the 2017 international conference on image processing* (pp. 3953–3957).

Voigtlaender, F. (2020). The universal approximation theorem for complex-valued neural networks. arXiv:2012.03351.

Warner, F. (1983). *Foundations of differentiable manifolds and Lie groups, vol. 94*. Springer.

Wells, R. (1980). *Differential analysis on complex manifolds, vol. 21980*. Springer.

Weyl, H. (1946). *The classical groups: Their invariants and representations*. Princeton University Press.

Wolter, M., & Yao, A. (2018). Complex gated recurrent neural networks. In *Advances in neural information processing systems, vol. 31* (pp. 10536–10546).

Worrall, D., Garbin, S., Turmukhambetov, D., & Brostow, G. (2017). Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the 30th conference on computer vision and pattern recognition* (pp. 5028–5037).

Wu, Wei, Jing, Xiaoyuan, Du, Wencai, & Chen, Guoliang (2021a). Learning dynamics of kernel-based deep neural networks in manifolds. *Science China Information Sciences*, *64*(11), 1–15.

Wu, J.-H., Zhang, S.-Q., Jiang, Y., & Zhou, Z.-H. (2021b). Towards theoretical understanding of flexible transmitter networks via approximation and local minima. arXiv:2111.06027.

Yeats, E. C., Chen, Y., & Li, H. (2021). Improving gradient regularization using complex-valued neural networks. In *Proceedings of the 38th international conference on machine learning* (pp. 11953–11963).

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, & Vinyals, Oriol (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of the 7th International Conference on Learning Representations*.

Zhang, S.-Q., & Fan, F.-L. (2021). Neural network gaussian processes by increasing depth. arXiv:2108.12862.

Zhang, H., Liu, X., Xu, D., & Zhang, Y. (2014). Convergence analysis of fully complex backpropagation algorithm based on wirtinger calculus. *Cognitive Neurodynamics*, *8*(3), 261–266.

Zhang, H., & Mandic, D. P. (2015). Is a complex-valued stepsize advantageous in complex-valued gradient learning algorithms? *IEEE Transactions on Neural Networks and Learning Systems*, *27*(12), 2730–2735.

Zhang, S.-Q., & Zhou, Z.-H. (2021). Flexible transmitter network. *Neural Computation*, *33*(11), 2951–2970.

Zhou, Zhi-Hua (2021). Why over-parameterization of deep neural networks does not overfit?. *Science China Information Sciences*, *64*(1), 1–3.