Journal of Machine Learning Research 27 (2026) 1-51 Submitted 5/25; Revised 1/26; Published 2/26

Reparameterized Complex-valued Neurons Can Efficiently Learn More than Real-valued Neurons via Gradient Descent

Jin-Hui Wu WUJH@LAMDA.NJU.EDU.CN
Shao-Qun Zhang ZHANGSQ@LAMDA.NJU.EDU.CN
Yuan Jiang JIANGY@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China

Editor: Qiang Liu

Abstract

Complex-valued neural networks potentially possess better representations and performance than real-valued counterparts when dealing with some complicated tasks such as acoustic analysis, radar image classification, etc. Despite empirical successes, it remains unknown theoretically when and to what extent complex-valued neural networks outperform real-valued ones. We take one step in this direction by comparing the learnability of real-valued neurons and complex-valued neurons via gradient descent. We theoretically show that a complex-valued neuron can learn functions expressed by any one real-valued neuron and any one complex-valued neuron with convergence rates $O(t^{-3})$ and $O(t^{-1})$ where t is the iteration index of gradient descent, respectively, whereas a two-layer real-valued neural network with finite width cannot learn a single non-degenerate complex-valued neuron. We prove that a complex-valued neuron learns a real-valued neuron with rate $\Omega(t^{-3})$, exponentially slower than the linear convergence rate of learning one real-valued neuron using a real-valued neuron. We then reparameterize the phase parameter of the complex-valued neuron and prove that a reparameterized complex-valued neuron can efficiently learn a real-valued neuron with a linear convergence rate. We further verify and extend these results via simulation experiments in more general settings.

Keywords: Neural Network Theory, Single-Neuron Learning, Gradient Descent, Convergence Rate, Reparameterization

1. Introduction

Complex-valued neural networks (CVNNs) use the neuron models and operations in the complex-valued domain and are good at handling many complicated scenarios. Pioneering works successfully apply CVNNs to various areas, such as synthetic aperture radar image classification (Zhang et al., 2017; Gabot et al., 2024), acoustic analysis (Shafran et al., 2018; Linhardt et al., 2025), and magnetic resonance image reconstruction (Wang et al., 2020). In these applications, input signals naturally contain phase information. CVNNs seem more suitable than real-valued neural networks (RVNNs) in phase-dependent tasks since empirical experiments and intuitive explanations suggest that CVNNs can possess better data representations of phase information and grasp the phase-rotational dynamics more accurately (Bassey et al., 2021; Lee et al., 2022).

Beyond the promising progress of CVNNs in practice, many efforts have been devoted to the theoretical understanding of CVNNs. Most existing works demonstrate some desirable properties of CVNNs, such as universal approximation (Kim and Adali, 2002; Voigtlaender, 2023), the minimum width for universal approximation (Geuchen et al., 2023), boundedness and complete stability (Zhou and Song, 2013), most critical points not being local minimum (Nitta, 2013), and local-minimum-free conditions (Wu et al., 2024). Some recent works demonstrate the approximation advantage of CVNNs in phase-invariant tasks by proving that neuromorphic networks with complex-valued operations can approximate radial functions with exponentially fewer parameters than RVNNs (Zhang et al., 2022; Wu et al., 2024). However, those studies do not explain why CVNNs outperform RVNNs mostly in phase-dependent tasks, particularly when considering the fact that there are functions that can be efficiently approximated but not efficiently learned with gradient methods (Malach and Shalev-Shwartz, 2019; Malach et al., 2021). Indeed, the general functional difference between CVNNs and RVNNs remains unknown.

In this paper, we take one step towards understanding when and to what extent one can benefit from common learning paradigms using CVNNs rather than RVNNs. More specifically, we attempt to identify the superiority and inferiority of CVNNs through the following fundamental questions

When do CVNNs outperform RVNNs via gradient descent?
Can we learn everything with CVNNs without paying additional price?

We theoretically study these questions by focusing on learning a single neuron by optimizing the expected square loss via gradient descent under the setting of low-dimensional inputs and no bias term. This paper significantly extends our previous NeurIPS study (Wu et al., 2023), in which we showed that the phase parameter helps complex-valued neurons learn more but slower than real-valued neurons. In this version, we further propose reparameterizations of phase parameters in complex-valued neurons to accelerate learning while maintaining the learning capability. Specifically, we provide a class of reparameterizations for phase parameters, prove that a reparameterized complex-valued neuron can efficiently learn more than a real-valued neuron, and verify the theoretical findings via simulation experiments in Section 6. Our contributions are summarized in Table 1 and further explained as follows.

Table 1: A summary of our contributions. The first column lists the learnable neurons. The second and third columns represent the convergence rates of learning a real-valued neuron and a complex-valued neuron via gradient descent, respectively.

Learnable Neurons	**Real-valued Neuron**	**Complex-valued Neuron**
Real-valued Neuron	$O(\mathrm{e}^{-c_1 t})$ (Lemma 5)[1]	Cannot Learn (Theorem 4)
Complex-valued Neuron	$\Theta(t^{-3})$ (Theorems 1 and 6)	$O(t^{-1})$ (Theorem 2)
Reparameterized CVN	$O(\mathrm{e}^{-c_2 t})$ (Theorem 7)	$O(t^{-1})$ (Theorem 8)

[1] Theorem 6.4 in (Yehudai and Shamir, 2020).

- **Complex-valued neurons can learn more than real-valued neurons.** We prove that using gradient descent, a single complex-valued neuron can learn functions expressed by any one real-valued neuron and any one complex-valued neuron with convergence rate $O(t^{-3})$ and $O(t^{-1})$ in Theorems 1 and 2, respectively. In contrast, we show the lower bound of expressing a non-degenerate complex-valued neuron with a two-layer RVNN in Theorem 4, which implies that a two-layer RVNN with finite width cannot learn a single non-degenerate complex-valued neuron. These results provide positive responses to the first question from at least two perspectives. Firstly, CVNNs outperform RVNNs when dealing with phase-sensitive tasks. Secondly, CVNN is always a conservative choice when we are unwilling to take the risk of failure.

- **Complex-valued neurons learn slower than real-valued neurons.** We present a lower bound $\Omega(t^{-3})$ for learning functions expressed by any one real-valued neuron using a complex-valued neuron via gradient descent in Theorem 6. This conclusion, together with the well-known linear convergence of learning functions expressed by any one real-valued neuron using a real-valued neuron (Yehudai and Shamir, 2020), implies that CVNNs suffer from slower convergence than RVNNs when handling simple phase-independent tasks. This phenomenon answers the second question and reveals the additional price for learning everything with CVNNs.

- **Reparameterized complex-valued neurons can efficiently learn more than real-valued neurons.** We propose a reparameterization of the phase parameter in complex-valued neurons and show that a reparameterized complex-valued neuron can learn real-valued and complex-valued neurons with convergence rate $O(\mathrm{e}^{-c_2 t})$ and $O(t^{-1})$ in Theorems 7 and 8, respectively. As reparameterization does not affect the hardness of learning a complex-valued neuron using RVNNs, these results indicate that reparameterized complex-valued neurons can learn more than real-valued neurons without the price of slower convergence. This conclusion provides a positive answer for the second question and highlights the importance of reparameterization of the phase parameter.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 details our settings and notations. Section 4 demonstrates that complex-valued neurons can learn more than real-valued neurons. Section 5 proves that complex-valued neurons learn slower than real-valued neurons. Section 6 proposes reparameterization of phase parameters and shows that reparameterized complex-valued neurons can efficiently learn more than real-valued neurons. Section 7 concludes our work with prospects.

2. Related Works

Complex-valued Neural Networks. CVNNs originate in the 1990s when parameters of networks and the commonly used back-propagation algorithm are generalized to the complex-valued domain (Leung and Haykin, 1991; Benvenuto and Piazza, 1992; Georgiou and Koutsougeras, 1992). The motivation of CVNNs is at least threefold. From the representation perspective, CVNNs consider the phase information and model complex-valued problems more efficiently and properly than RVNNs (Arena et al., 1995; Amin et al., 2011; Hirose and Yoshida, 2012). From the computation perspective, a complex-valued neuron is capable of solving the exclusive-or problem and the detection of symmetry, whereas a real-valued neuron cannot (Nitta, 2003). From the biological perspective, the recently proposed flexible transmitter neuron (Zhang and Zhou, 2021), which has a natural complex implementation, formulates the communication between pre-synapse and post-synapse precisely rather than considering only the pre-synapse in traditional MP neuron (McCulloch and Pitts, 1943). CVNNs achieve better performance in versatile applications, especially those with naturally phase-related signals, such as radio frequency signals (Zhang and Wu, 2017), sonar signals (Gao et al., 2018), and audio signals (Trabelsi et al., 2018). We refer to two surveys for more detailed discussions (Bassey et al., 2021; Lee et al., 2022).

From the aspect of theories, several works provide preliminary support for CVNNs by proving fundamental properties of CVNNs, such as shallow CVNNs are universal approximators (Kim and Adali, 2002; Voigtlaender, 2023), most critical points are not spurious local minimum (Nitta, 2013; Wu et al., 2024), and CVNNs are bounded and completely stable (Zhou and Song, 2013). These theoretical insights only consider CVNNs without comparison with RVNNs. Another line of research verifies the superiority of CVNNs by comparing the approximation complexity of RVNNs and CVNNs and finding that CVNNs can express radial functions more efficiently (Wu et al., 2024; Zhang et al., 2022). This line of work only takes approximation into account and does not explicitly consider learning processes, which is of more interest in practice. This work takes the first step toward analyzing and comparing the learning behaviors of CVNNs and RVNNs.

Neuron Learning. Neuron learning is the simplest case of neural network learning, and existing works mainly focus on learning real-valued neurons (Auer et al., 1995; Frei et al., 2020; Diakonikolas et al., 2020, 2022). Some studies demonstrate the possibility of learning one real-valued neuron or a network using meticulously designed algorithms (Kalai and Sastry, 2009; Kakade et al., 2011; Goel et al., 2017). Later, researchers investigate the learnability of neurons using standard gradient methods. An exponential convergence rate is established for learning one real-valued neuron with a real-valued neuron under different assumptions (Yehudai and Shamir, 2020; Soltanolkotabi, 2017; Du et al., 2018; Mei et al., 2018; Kalan et al., 2019). We consider the problem of learning between one real-valued neuron and one complex-valued neuron, as well as learning one complex-valued neuron using a complex-valued neuron. The heterogeneity between real-valued and complex-valued neurons makes the analysis of optimization behaviors more complicated.

3. Preliminaries

Notations. Suppose that the input dimension is an even number. For any vector $\boldsymbol{x} \in \mathbb{R}^{2d}$, we denote x_i as the i-th coordinate of $\boldsymbol{x}$. Let $\boldsymbol{x}_{\mathbb{C}} = (x_1; \dots; x_d) + (x_{d+1}; \dots; x_{2d})\mathrm{i} \in \mathbb{C}^d$ be the folded complex-valued representation of $\boldsymbol{x}$, and $\overline{\boldsymbol{x}}_{\mathbb{C}}$ is the complex conjugate of $\boldsymbol{x}_{\mathbb{C}}$. For any two vectors $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{2d}$, $\theta_{\boldsymbol{w},\boldsymbol{v}} = \arccos(\boldsymbol{w}^\top \boldsymbol{v} \|\boldsymbol{w}\|^{-1} \|\boldsymbol{v}\|^{-1})$ denotes the angle between $\boldsymbol{w}$ and $\boldsymbol{v}$. For any $x \in \mathbb{R}$, $\tau(x) = \max\{0, x\}$ indicates the ReLU activation function. Let $\mathrm{Re}(z)$ be the real part of a complex number z. For any $z \in \mathbb{C}$ and $\psi \in [0, \pi/2]$, $\sigma_\psi(z)$ denotes the real part of the symmetrical version of zReLU activation function (Guberman, 2016), i.e.,

$$\sigma_\psi(z) = \begin{cases} \mathrm{Re}(z) \ , & \theta_z \in [-\psi, \psi] \ , \\ 0 \ , & \text{otherwise} \ , \end{cases}$$

where θ_z represents the argument of z. For any proposition p, we use $\mathbb{I}(p)$ to represent the indicator function of p, i.e., $\mathbb{I}(p) = 1$ if p is true and $\mathbb{I}(p) = 0$ otherwise. A table of frequently used notations is provided in Table 2.

Table 2: Frequently used notations.

Notation	Description
$\mathbb{C}^d, \mathbb{R}^d$	the d-dimensional complex or real space
$\mathbb{E}$	expectation
$\mathbb{I}(\cdot)$	the indicator function
L	the expected square loss of learning a neuron
$\mathcal{N}(\mathbf{0}, \mathbf{I})$	the standard Gaussian distribution
O, Ω, Θ	asymptotic notations
Pr	probability
$P_{\mathcal{Q}}(\boldsymbol{x})$	the projection of $\boldsymbol{x}$ on $\mathcal{Q}$
$\mathrm{Re}(z)$	the real part of a complex number z
t	the iteration index of gradient descent
$\mathcal{U}(a, b)$	the uniform distribution on the interval $[a, b]$
$\boldsymbol{v}, \boldsymbol{w}$	the weight vector of a learning or target neuron
$\boldsymbol{x}, x_i$	an input vector $\boldsymbol{x} \in \mathbb{R}^{2d}$ with the i-th coordinate x_i
$\boldsymbol{x}_{\mathbb{C}}, \overline{\boldsymbol{x}}_{\mathbb{C}}$	$\boldsymbol{x}_{\mathbb{C}} = (x_1; \dots; x_d) + (x_{d+1}; \dots; x_{2d})\mathrm{i} \in \mathbb{C}^d$ with complex conjugate $\overline{\boldsymbol{x}}_{\mathbb{C}}$
$\theta_{\boldsymbol{a},\boldsymbol{b}}$	the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$
θ_z	the argument of a complex number z
$\sigma_\psi(z)$	the real part of the symmetrical version of zReLU activation function
η	the step size of gradient descent
τ	the ReLU activation function $\tau(x) = \max\{0, x\}$
ψ, ϕ	phase parameters of zReLU before or after reparameterization
$\\|\cdot\\|$	the 2-norm of a vector

Learning a Single Neuron. We consider learning a target neuron with a learning neuron. A neuron generally takes the form $\boldsymbol{x} \to \sigma_\psi(\boldsymbol{w}; \boldsymbol{x})$, where the weight $\boldsymbol{w} \in \mathbb{R}^{2d}$ and phase $\psi \in [0, \pi/2]$ indicate learnable parameters, and we omit the bias term for technical reasons. This general formulation includes a real-valued neuron with ReLU activation $\boldsymbol{x} \to \tau(\boldsymbol{w}^\top \boldsymbol{x})$ and a complex-valued neuron with zReLU activation $\boldsymbol{x} \to \sigma_\psi(\boldsymbol{w}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}})$ as special cases. For any target neuron with parameters $(\boldsymbol{v}, \psi_v)$, the learning process consists of finding a neuron with parameters $(\boldsymbol{w}, \psi_w)$ to minimize the expected square loss

$$L(\boldsymbol{w}, \psi_w) = \frac{1}{2} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left(\sigma_{\psi_w}(\boldsymbol{w}; \boldsymbol{x}) - \sigma_{\psi_v}(\boldsymbol{v}; \boldsymbol{x}) \right)^2 \right], \tag{1}$$

where the learnable parameter ψ_w occurs only when the learning neuron is complex-valued.

Learning Algorithm and Initialization. We use the projected gradient descent as the learning algorithm, where the projection guarantees the constraint on phase $\psi \in [0, \pi/2]$. To minimize a function $f(\boldsymbol{x})$ with an initialization $\boldsymbol{x}_0$, projected gradient descent iteratively updates weights via $\boldsymbol{x}_{t+1} = P_{\mathcal{Q}}(\boldsymbol{x}_t - \eta_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t))$, where η_t represents the step size, $\mathcal{Q}$ denotes the constraint set, and $P_{\mathcal{Q}}$ indicates the projection operator $P_{\mathcal{Q}}(\boldsymbol{x}_0) = \arg\min_{\boldsymbol{x} \in \mathcal{Q}} \|\boldsymbol{x} - \boldsymbol{x}_0\|$. We initialize weights of neurons with Gaussian distribution, which includes most standard initialization schemes in practice (Glorot and Bengio, 2010). The learnable parameter of the zReLU activation is initialized with $\mathcal{U}(0, \pi/2)$, i.e., the uniform distribution on $[0, \pi/2]$.

4. Complex-valued Neurons Can Learn More

In this section, we provide theoretical support for the learning advantage of complex-valued neurons by providing two positive learning scenarios for complex-valued neurons and one negative learning result for real-valued neurons. Specifically, Subsections 4.1 and 4.2 confirm the learning power of complex-valued neurons, by verifying that a complex-valued neuron can learn functions expressed by any one real-valued neuron and any one complex-valued neuron, respectively. Subsection 4.3 points out the limited learning capability of real-valued neurons, by proving that a two-layer RVNN with finite width cannot learn a single non-degenerate complex-valued neuron.

4.1 Learning One Real-valued Neuron with a Complex-valued Neuron

We first study the problem of whether one can learn one real-valued neuron with ReLU activation function using a complex-valued neuron with zReLU activation function, where the expected square loss in Eq. (1) becomes

$$L_{\mathrm{cr}}(\boldsymbol{w}, \psi) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\sigma_\psi(\boldsymbol{w}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) - \tau(\boldsymbol{v}^\top \boldsymbol{x})\right)^2\right], \qquad (2)$$

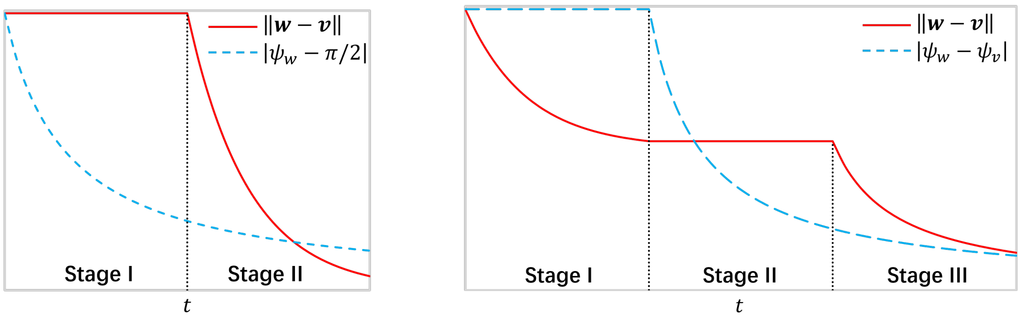
where we abbreviate the phase parameter ψ_w as ψ since the target real-valued neuron does not have a phase parameter, $\boldsymbol{w} \in \mathbb{R}^{2d}$ and $\boldsymbol{v} \in \mathbb{R}^{2d}$ represent the weight vectors of the complex-valued and the real-valued neurons, respectively. We assume $\|\boldsymbol{v}\| = 1$ without loss of generality. Then we present the first theorem for complex-valued neuron learning.

Theorem 1 *Let $d = 1$. Suppose that $\boldsymbol{w}_0 \sim \mathcal{N}(0, \mathbf{I}_2)$ and $\psi_0 \sim \mathcal{U}(0, \pi/2)$. Let $\{(\boldsymbol{w}_t, \psi_t)\}_{t=0}^{\infty}$ denote the parameter sequence of the complex-valued neuron generated by projected gradient descent when optimizing L_{cr}, the expected loss of learning a real-valued neuron using a complex-valued neuron. If the step size $\eta_t = \eta \in (0, 1/(12\pi))$, then we have*

$$\Pr\left[L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \leqslant \frac{8000}{\eta^3 t^3} + \left(1 - \frac{\eta}{48}\right)^{t+1-32/\eta}\right] \geqslant \frac{1}{32}\ .$$

Theorem 1 shows that a complex-valued neuron can learn the functions expressed by any one real-valued neuron with convergence rate $O(t^{-3})$ using projected gradient descent. It should be mentioned that the large constants are results of loose bounds and rescaling in the proof, which can be improved by more careful analysis. We do not attempt to decrease these large constants, as they do not hurt the learnability and convergence rate.

Notice that Theorem 1 adopts the constant probability 1/32, rather than a high probability. First, this constant probability comes from the intrinsical difference between real-valued neurons and complex-valued neurons. A real-valued neuron activates half of the phase domain, whereas a complex-valued neuron may only activate a small part as controlled by the parameter ψ, which makes the expected loss a piecewise function. When the initialization of $\boldsymbol{w}$ falls into the opposite direction of $\boldsymbol{v}$ and ψ is small, the activated regions of the real-valued and complex-valued neurons are not overlapped. Such a bad initialization happens with a constant probability and encourages the complex-valued neuron to decrease phase to minimize the loss. As a result, the phase of the complex-valued neuron will shrink to zero, which leads to a constant loss and the failure of learning. Second, it is possible



(a) Convergence stages of Theorem 1.

(b) Convergence stages of Theorem 2.

Figure 1: Subfigures (a) and (b) demonstrate the convergence stages of Theorems 1 and 2, respectively. The horizontal axis represents the iteration index of gradient descent. The black dotted line denotes the separation of convergence stages.

to amplify the constant probability to a high probability by training multiple independent complex-valued neurons simultaneously and then choosing the one with the smallest loss.

Challenges. Although $(\boldsymbol{w},\psi)=(\boldsymbol{v},\pi/2)$ is an obvious global minimum of the expected loss, the convergence conclusion in Theorem 1 is non-trivial. As one can see in the proof, the landscape of the expected loss possesses a stationary point $(\boldsymbol{w},\psi)=\mathbf{0}$. If we initialize $\boldsymbol{w}=-k\boldsymbol{v}$ with $k>0$, then it is easy to verify that $\boldsymbol{w}$ converges to $\mathbf{0}$ and ψ decreases to 0 when the step size is sufficiently small. This implies that the landscape is not convex, and $\mathbf{0}$ is an attractor. The existence of this spurious stationary point becomes a critical obstacle in the proof and provides another reason for the hardness of a high-probability conclusion.

The proof idea of Theorem 1 mainly consists of estimating the gradients and finding an ideal region with both constant probability and convergence guarantees. We provide a proof sketch as follows. Firstly, we analyze the optimization behaviors of $\boldsymbol{w}$ and ψ separately. Then we identify an ideal region with desirable properties: the gradient $\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w},\psi)$ can be bounded by $O((\psi-\pi/2)^2)$, which implies $\psi-\pi/2=O(t^{-1})$. Meanwhile, gradient descent on $\boldsymbol{w}$ performs like a contraction mapping with fixed point $\boldsymbol{v}$ and Lipschitz constant $1-\Theta(\psi)$, i.e., $\boldsymbol{w}$ converges to $\boldsymbol{v}$ linearly when ψ is large enough. Based on these observations, the convergence consists of two stages, as shown in Fig. 1(a). In Stage I, the phase ψ converges to the global minimum, and the weight $\boldsymbol{w}$ remains in the ideal region. When the phase grows above some threshold, one enters Stage II where the weight converges linearly and the phase maintains its slow convergence rate. Finally, we estimate the order of loss and probability of the ideal region to complete the proof. Full proofs are available in Appendix B.

4.2 Learning One Complex-valued Neuron with a Complex-valued Neuron

We investigate whether one can learn one complex-valued neuron using a complex-valued neuron. In this case, the expected square loss in Eq. (1) can be rewritten as

$$L_{\mathrm{cc}}(\boldsymbol{w},\psi)=\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\sigma_\psi(\boldsymbol{w}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})-\sigma_{\psi_v}(\boldsymbol{v}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})\right)^2\right],$$

where $(\boldsymbol{v},\psi_v)$ denotes the parameter of the target complex-valued neuron, and $(\boldsymbol{w},\psi)$ is the learnable parameter. In general, we still assume $\|\boldsymbol{v}\|=1$. Here, we use gradient descent with vanishing step size $x_{t+1}=x_t-\eta_t\nabla f(x_t)$, where the positive step size η_t satisfies $\eta_t\to 0$ as $t\to\infty$. Then we present the second theorem for complex-valued neuron learning.

Theorem 2 *Let $d = 1$, and $\psi_v \in [7\pi/20, 2\pi/5]$. Suppose that $\boldsymbol{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and $\psi_{w,0} \sim \mathcal{U}(0, \pi/2)$. Let $\{(\boldsymbol{w}_t, \psi_{w,t})\}_{t=0}^{\infty}$ denote the parameter sequence of the complex-valued neuron generated by projected gradient descent when optimizing L_{cc}, the expected loss of learning a complex-valued neuron using a complex-valued neuron. If we use vanishing step size $\eta_t = \min\{c_1, c_2/t\}$ with $c_1 \leqslant 1/3000$ and $c_2 \geqslant 20$, then*

$$\Pr\left[L_{\mathrm{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \leqslant \frac{400c_2^3}{c_1 t}\right] \geqslant 10^{-5} \ .$$

Theorem 2 shows that a complex-valued neuron can learn functions expressed by any one complex-valued neuron with convergence rate $O(t^{-1})$ and constant probability.

Challenges. It is observed that the $O(t^{-1})$ convergence rate in Theorem 2 is slower than the $O(t^{-3})$ convergence rate in Theorem 1. The deceleration of convergence comes from the intrinsic difficulties of learning functions expressed by any one complex-valued neuron. These difficulties become the main challenges in the analysis and can be understood from at least two perspectives. Firstly, there emerge new spurious stationary points. As one can see in the proof, the gradient with respect to ψ_w becomes 0 once ψ_w reaches $\pi/2$ and $\boldsymbol{w}$ is close to $\boldsymbol{v}$, i.e., $(\boldsymbol{w}, \psi_w) = (\boldsymbol{v}, \pi/2)$ is a spurious stationary point. Secondly, the landscape of the loss function is no longer smooth. For both $\boldsymbol{w}$ and ψ_w, the local landscape around the global minimum is roughly an absolute function, which declares the non-smoothness of the loss and the failure of gradient descent with a constant step size.

To overcome these obstacles, we apply mild conditions and slight modifications to guarantee convergence and maintain the generality of our conclusion. We separate the phase of the target complex-valued neuron far from 0 and $\pi/2$ in consideration of spurious local stationary points: As ψ_v tends to 0, the initialization of the learning neuron is less likely to overlap with the target neuron. Then we will take the risk of falling into the spurious local minimum $(\boldsymbol{w}, \psi_w) = (\mathbf{0}, 0)$. As ψ_v approaches $\pi/2$, we are confronted by another spurious stationary point $\psi_w = \pi/2$. We use a vanishing step size to cope with the non-smoothness of the loss function since a constant step size inevitably suffers from oscillation.

We summarize the proof idea of Theorem 2 as follows. The overall procedure is similar to that of Theorem 1 but every step is different and more challenging due to non-smoothness and more spurious stationary points. Firstly, we identify an ideal region with nice gradient properties: the gradient with respect to $\boldsymbol{w}_\perp$, the weight component perpendicular to $\boldsymbol{v}$, points to the global minimum $\mathbf{0}$ and maintains constant order. The gradient ∇_{ψ_w} is bounded and points towards ψ_v when the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ is small enough. Meanwhile, the gradient with respect to $\boldsymbol{w}_{\boldsymbol{v}}$ performs like a contraction mapping with fixed point $[1 - \Theta(\psi_v \psi_w^{-1})]\boldsymbol{v}$ and Lipschitz constant $1 - \Theta(\psi)$, i.e., there exists a deviation of the fixed point from the global minimum. Based on these observations, we then prove the convergence with three stages, as demonstrated in Fig 1(b): In Stage I, $\boldsymbol{w}_\perp$, the weight component perpendicular to $\boldsymbol{v}$, converges to 0 with an inversely proportional rate, and ψ_w and $\boldsymbol{w}_{\boldsymbol{v}}$ remain in the ideal region. Thus, the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ decreases with an inversely proportional rate. When $\theta_{\boldsymbol{w},\boldsymbol{v}}$ declines below some threshold, we come to Stage II where phase ψ_w converges to ψ_v with rate $O(t^{-1})$. As ψ_w approaches ψ_v, the fixed point becomes close to $\boldsymbol{v}$ and we step into Stage III where $\boldsymbol{w}$ converges to $\boldsymbol{v}$ with the same rate as ψ_w. Finally, we estimate the order of loss and provide a lower bound for the probability of falling into the ideal region with Gaussian initialization to complete the proof. We provide detailed proofs in Appendix C.

4.3 RVNNs Cannot Learn a Complex-valued Neuron

This subsection focuses on whether one can learn one complex-valued neuron with zReLU activation using real-valued neurons. However, it is naturally unfair to learn a complex-valued neuron with a single real-valued neuron since a complex-valued neuron has more parameters than a real-valued neuron. To tackle this issue, we consider the problem of learning a complex-valued neuron with a two-layer RVNN. A two-layer RVNN with n hidden neurons can be represented by $\boldsymbol{x} \to \boldsymbol{\alpha}^\top \tau(\mathbf{W}\boldsymbol{x})$, where $\mathbf{W} \in \mathbb{R}^{n\times 2d}$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$ indicate weight parameters of the network, and τ is the ReLU activation function applied componentwisely. We still focus on the expected square loss, which takes the form

$$L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\boldsymbol{\alpha}^\top \tau(\mathbf{W}\boldsymbol{x}) - \sigma_\psi(\boldsymbol{v}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}})\right)^2\right],$$

where we abbreviate the phase parameter ψ_v as ψ since RVNN has no phase parameter. We are mainly interested in learning a non-degenerate complex-valued neuron, which is distinct from a real-valued neuron and defined as follows.

Definition 3 *A complex-valued neuron is a non-degenerate one if $\psi \notin \{0, \pi/2\}$ and $\boldsymbol{v} \neq \mathbf{0}$.*

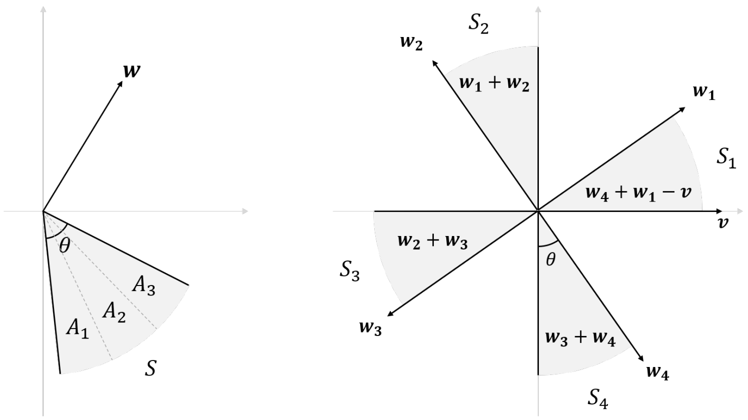
For a complex-valued neuron with phase $\psi = 0$ or $\boldsymbol{v} = \mathbf{0}$, the zReLU activation always outputs 0. Then the complex-valued neuron is equivalent to a real-valued neuron with all zero weights. For a complex-valued with phase $\psi = \pi/2$, the zReLU activation is equivalent to the ReLU activation. Thus, a non-degenerate complex-valued neuron is a non-real-valued neuron. Then we present the third theorem for complex-valued neuron learning.

Theorem 4 *Let $d = 1$. For any non-degenerate complex-valued neuron, denote by L_{rc} the expected square loss of learning this complex-valued neuron using a one-hidden-layer RVNN with n hidden neurons. Then the loss satisfies*

$$L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) \geqslant \frac{\|\boldsymbol{v}\|^2 \min\{2\psi, \pi - 2\psi\}^3}{24\pi(n+2)^2} > 0\ .$$

Theorem 4 provides a positive lower bound for the expected squared loss of approximating a non-degenerate complex-valued neuron using a two-layer RVNN with a fixed number of hidden neurons. This positive lower bound indicates that there always remains a positive gap between the target non-degenerate complex-valued neuron and the two-layer RVNN of fixed width no matter how the parameters of the RVNN are learned. Thus, a finite-width RVNN cannot learn a single non-degenerate complex-valued neuron.

The lower bound decreases at rate $\Theta(\|\boldsymbol{v}\|^2 \min\{2\psi, \pi - 2\psi\}^3 n^{-2})$. The norm term $\|\boldsymbol{v}\|$ depicts the magnitude of the problem, which affects the expected square loss quadratically from the homogeneity of zReLU. In the extreme case of $\boldsymbol{v} = \mathbf{0}$, a trivial real-valued neuron with $\boldsymbol{w} = \mathbf{0}$ reaches the lower bound 0. Meanwhile, the lower bound possesses a positive relation with a phase-dependent term $\{2\psi, \pi - 2\psi\}$. Intuitively, this term indicates the difference between a complex-valued neuron and a real-valued neuron. A real-valued neuron corresponds to $\psi{=}0$ or $\psi{=}\pi/2$ and this term measures the distance between the phase of a complex-valued neuron and a real-valued one. Finally, the lower bound decreases with a rate


$\boldsymbol{w}$
θ
A_3
A_2
A_1
S
S_2
$\boldsymbol{w}_2$
$\boldsymbol{w}_1 + \boldsymbol{w}_2$
$\boldsymbol{w}_1$
S_1
$\boldsymbol{w}_4 + \boldsymbol{w}_1 - \boldsymbol{v}$
$\boldsymbol{v}$
$\boldsymbol{w}_2 + \boldsymbol{w}_3$
S_3
θ
$\boldsymbol{w}_3$
$\boldsymbol{w}_3 + \boldsymbol{w}_4$
$\boldsymbol{w}_4$
S_4

(a) The loss in one sector.

(b) The loss in four sectors.

Figure 2: An illustration of the proof idea of Theorem 4. Shaded areas are sectors with infinite radii. (a) The expectation $\mathbb{E}_{\boldsymbol{x}\in S}[(\boldsymbol{w}^\top \boldsymbol{x})^2]$ in the sector S equals the sum of the expectations on three subareas A_1, A_2, and A_3. The minimum expectation on a subarea can be bounded by $\Omega(\theta^2\|\boldsymbol{w}\|^2)$. (b) The expected loss of learning a complex-valued neuron using four symmetric real-valued neurons in four symmetric sectors can be bounded by $\Omega(\theta^2\|\boldsymbol{v}\|^2)$, where $\boldsymbol{w}_i$ and $\boldsymbol{v}$ are weights of real-valued and complex-valued neurons, respectively.

inversely proportional to the square of hidden size n. We conjecture that this dependence is tight and cannot be improved: a two-layer RVNN with n neurons divides the space into n pieces, in each of which RVNN acts as a linear function. Choosing the n weight vectors of RVNN suitably, the difference between the RVNN and the complex-valued neuron remains small (of order n^{-1}) in each piece, which leads to the expected loss of order $O(n^{-2})$.

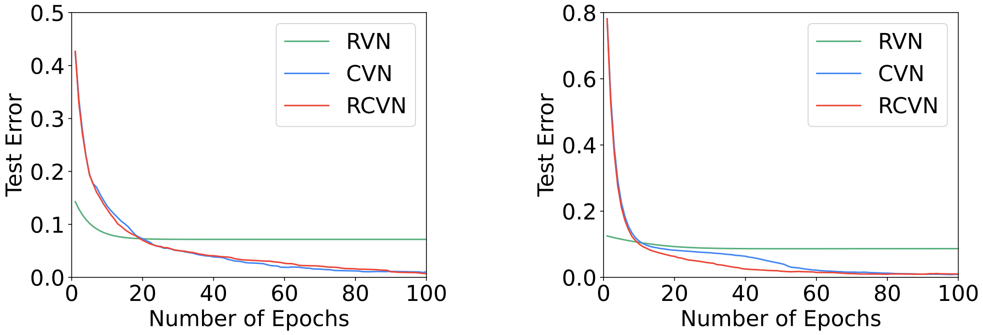
The conditions in Theorem 4 are made for conciseness of proof and we believe the conclusion holds in more general cases. The dimension $d = 1$ corresponds to the intrinsic dimension of expressing a complex-valued neuron because of the rotational invariance of the inner product and the spherical symmetry of Gaussian distributions. Thus, additional dimensions contain no information and cannot improve the efficiency of approximation when $d > 1$. It is necessary to consider non-degenerate complex-valued neurons since degenerate complex-valued neurons are equivalent to real-valued ones.

We provide proof ideas of Theorem 4 as follows. The expected square loss L_{rc} is a piecewise quadratic function, and each piece forms a sector centered at the origin with an infinite radius. In each piece, L_{rc} takes the form $\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x})^2]$. The proof mainly consists of a lower bound of L_{rc} in a sector and summation over all sectors with suitable weights and order. Firstly, we consider the expected loss in a sector with a small central angle θ, as shown in Fig. 2(a). We divide the sector into three identical subareas A_1, A_2, and A_3. Then at least one of A_1 and A_3 remains $\theta/6$ away from the vertical direction of $\boldsymbol{w}$, which leads to a lower bound $\Omega(\theta^2\|\boldsymbol{w}\|^2)$. Secondly, we consider the loss in four rotationally symmetric sectors, as shown in Fig. 2(b), where $\boldsymbol{v}$ and $\boldsymbol{w}_i$ are complex-valued and real-valued neurons, respectively, and the expression in each sector implies the activated neurons. Then at least one sector possesses a weight vector with norm $\Omega(\|\boldsymbol{v}\|)$, no matter how we choose $\boldsymbol{w}_i$'s. Thus, the overall loss is bounded by $\Omega(\theta^2\|\boldsymbol{v}\|^2)$. Finally, we take the weight $\boldsymbol{\alpha}$ into consideration and sum over all sectors. For RVNN with n neurons, the best choice of $\theta = \Theta(n^{-1})$ arrives at the lower bound $\Omega(n^{-2}\|\boldsymbol{v}\|^2)$. Detailed proofs are provided in Appendix D.

Table 3: A complex-valued neuron can learn more than a real-valued neuron.

Target	Real-valued Neuron	Complex-valued Neuron
Real-valued Neuron	$O(\mathrm{e}^{-ct})$ (Lemma 5)[1]	$O(t^{-3})$ (Theorem 1)
Complex-valued Neuron	Cannot Learn (Theorem 4)	$O(t^{-1})$ (Theorem 2)

[1] Theorem 6.4 in (Yehudai and Shamir, 2020).



(a) $d = 1$ and no bias term.

(b) $d = 5$ and with bias terms.

Figure 3: The test error of learning a complex-valued neuron. In both the theoretical setting (Fig. 3a) and more general settings (Fig. 3b), both complex-valued neurons and reparameterized complex-valued neurons have vanishing errors, while real-valued neurons converge to positive errors.

The rotational invariance is the most essential factor that causes the difference in expressive power between real-valued and complex-valued neurons. For a real-valued neuron, the sum of weight vectors in vertically opposite sectors (e.g., $S_1 \cup S_3$ in Fig. 2(b)) is a constant. While for a complex-valued neuron, the sum is no longer invariant because of the phase-related activated region. Then it is possible to extend Theorem 4 to all complex-valued activation functions without rotational invariance by using infinitesimal sector angles.

Summary and simulation experiments. We summarize the main conclusions of this section in Table 3. Both real-valued and complex-valued neurons succeed in learning functions expressed by any one real-valued neuron. But difference occurs when learning those expressed by non-degenerate complex-valued neurons: A complex-valued neuron can learn functions expressed by any one complex-valued neuron, but a two-layer RVNN with finite width cannot learn a single non-degenerate complex-valued neuron. Such a disagreement demonstrates that a complex-valued neuron possesses more powerful learning capability, which profits from the consideration of phase information in complex-valued operations. Our theoretical conclusions are based on the setting of low-dimensional inputs and no bias term, and the simulation results in Fig. 3 verify and extend these discoveries in more general settings. Details about the simulation experiments are available in Appendix H.

5. Complex-valued Neurons Learn Slower

In this section, we show that complex-valued neurons learn slower than real-valued neurons. To arrive at this conclusion, we first rephrase the linear convergence of learning functions expressed by any one real-valued neuron using real-valued neurons. Then we prove that a complex-valued neuron learns the same class of functions at an exponentially slower rate.

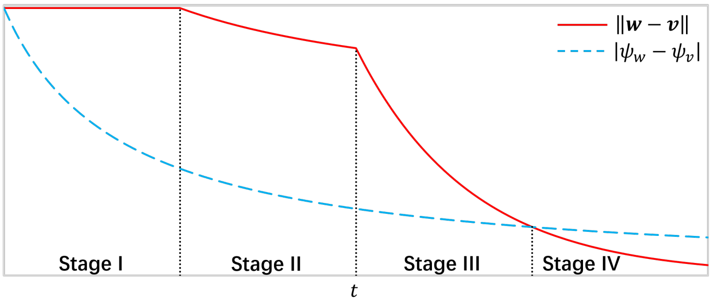


Figure 4: A demonstration of convergence stages of Theorem 6. The horizontal axis is the iteration index of gradient descent. The black dotted line is the separation of stages.

We concentrate on learning one real-valued neuron $\boldsymbol{x} \to \tau(\boldsymbol{v}^\top \boldsymbol{x})$ with $\|\boldsymbol{v}\| = 1$. When learning one real-valued neuron using a real-valued neuron, the expected square loss in Eq. (1) possesses the following simple closed form (Safran and Shamir, 2018)

$$L_{\mathrm{rr}}(\boldsymbol{w}) = \frac{1}{4}\|\boldsymbol{w}\|^2 - \frac{1}{2\pi}\|\boldsymbol{w}\|[\sin\theta_{\boldsymbol{w},\boldsymbol{v}} + (\pi - \theta_{\boldsymbol{w},\boldsymbol{v}})\cos\theta_{\boldsymbol{w},\boldsymbol{v}}] + \frac{1}{4} \ .$$

It is widely known that a real-valued neuron learns a real-valued neuron with high probability and linear convergence rate from (Yehudai and Shamir, 2020, Theorem 6.4).

Lemma 5 *(Yehudai and Shamir, 2020, Theorem 6.4) Suppose that the weight vector $\boldsymbol{w} \in \mathbb{R}^{2d}$ is initialized by a Gaussian distribution $\mathcal{N}(0, \mathbf{I}/(2d))$. Let L_{rr} denote the expected square loss of learning a real-valued neuron using a real-valued neuron. Then there exist constants c_1, c_2 such that gradient descent with suitable step size satisfies*

$$\Pr[L_{\mathrm{rr}}(\boldsymbol{w}_t) \leqslant \mathrm{e}^{-c_1 t}] \geqslant 1 - \mathrm{e}^{-c_2 d} \ .$$

Then we present the fourth theorem for complex-valued neuron learning.

Theorem 6 *Let $d = 1$. Suppose that $\|\boldsymbol{w}_0 - \boldsymbol{v}\| < 1$. Let $\{(\boldsymbol{w}_t, \psi_t)\}_{t=0}^{\infty}$ denote the parameter sequence of the complex-valued neuron generated by projected gradient descent when optimizing L_{cr} in Eq. (2). If the step size $\eta_t = \eta \in (0, 1/(12\pi))$, then we have*

$$L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \geqslant \frac{(1 - 12\eta)^{3T_3/2}(\psi^* - \psi_0)^3}{8\pi(t - T_3 + 1)^3} - \frac{1}{2\pi}\left(1 - \frac{\eta}{48}\right)^{t - T_3} \ ,$$

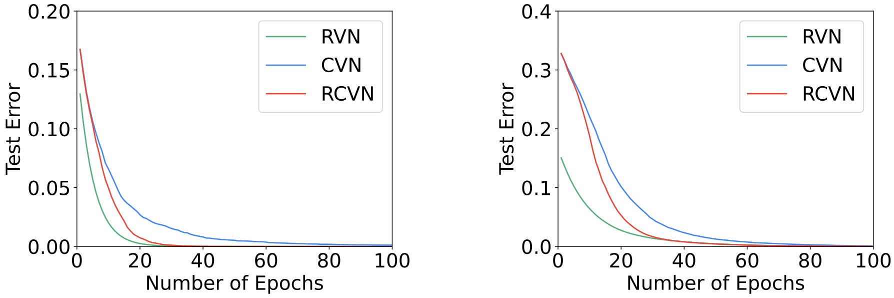
where $\psi^* = \pi/2$, and T_3 is a constant dependent on $\|\boldsymbol{w}_0 - \boldsymbol{v}\|$, η, and $\psi^* - \psi_0$.

Theorem 6 presents a lower bound for the expected loss of learning one real-valued neuron with a complex-valued neuron. It is observed that the negative term in the lower bound becomes 0 exponentially fast as t increases, and the positive term decreases with order $\Omega(t^{-3})$. Thus, the expected loss possesses a lower bound $\Omega(t^{-3})$ since the positive term dominates the loss when t grows sufficiently large. This lower bound matches the upper bound in Theorem 1. Thus, $O(t^{-3})$ becomes the utmost limit of learning with a complex-valued neuron via gradient descent, i.e., we cannot expect a complex-valued neuron to learn faster than this utmost limit.

Table 4: A real-valued neuron learns faster than a complex-valued neuron.

Target	Real-valued Neuron	Complex-valued Neuron
Real-valued Neuron	$O(\mathrm{e}^{-ct})$ (Lemma 5)[1]	$\Omega(t^{-3})$ (Theorem 6)

[1] Theorem 6.4 in (Yehudai and Shamir, 2020).



(a) $d = 1$ and no bias term.

(b) $d = 5$ and with bias terms.

Figure 5: The test error of learning a real-valued neuron. In both the theoretical setting (Fig. 5a) and more general settings (Fig. 5b), a reparameterized complex-valued neuron learns almost as efficiently as a real-valued neuron and faster than a complex-valued neuron.

The conditions in Theorem 6 are technical and reasonable. The condition on $\boldsymbol{w}_0$ is made for the conciseness of proof and can be removed. It is observed that $(\boldsymbol{w}, \psi) = (\boldsymbol{v}, \psi^*)$ is the unique global minimum with $L_{\mathrm{cr}} = 0$. Meanwhile, it is easy to verify that the loss goes to infinity when $\|\boldsymbol{w}\| \to \infty$. Thus, if we aim to obtain a small loss, the parameter sequence must fall into a small neighborhood of the global minimum, which is depicted by condition $\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant R < 1$. The condition $\psi_0 \neq \psi^*$ holds with probability 1 when we initialize ψ_0 with a continuous distribution. This condition is necessary to obtain a meaningful lower bound since the numerator of the positive term equals 0 when $\psi_0 = \psi^*$. We emphasize that this condition is essential and cannot be removed because a complex-valued neuron with $\psi_0 = \psi^*$ is equivalent to a real-valued neuron, which enjoys linear convergence as stated in Lemma 5. The condition $d = 1$ corresponds to the simplest optimization problem of learning one real-valued neuron with a complex-valued neuron since high-dimensional optimization brings more difficulties. Thus, we cannot expect a complex-valued neuron to learn a real-valued neuron with a convergence rate faster than $O(t^{-3})$ in a higher dimension.

We summarize the proof idea of Theorem 6 as follows. The gradient with respect to ψ possesses the order $(\psi^* - \psi)^2 + (\psi^* - \psi)\theta_{\boldsymbol{w},\boldsymbol{v}}$. The key intuition is that ψ converges fast to the global minimum when $\theta_{\boldsymbol{w},\boldsymbol{v}}$ remains large, but $\theta_{\boldsymbol{w},\boldsymbol{v}}$ diminishes as $\boldsymbol{w}$ converges to the global minimum $\boldsymbol{v}$. The detailed proofs are complicated and consist of several stages, depicting the entangled convergence between $\boldsymbol{w}$ and ψ as shown in Figure 4. In Stage I, ψ increases above a positive constant, which is a necessary condition for fast convergence of $\boldsymbol{w}$ in Stage II. When the distance between $\boldsymbol{w}$ and $\boldsymbol{v}$ declines below a threshold, the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ becomes small. Then we enter Stage III, where $\boldsymbol{w}$ converges faster than ψ. Stage IV begins when $\psi^* - \psi$ dominates $\theta_{\boldsymbol{w},\boldsymbol{v}}$. Then the gradient degenerates to order $(\psi^* - \psi)^2$, which implies a lower bound of convergence $\psi^* - \psi = \Omega(t^{-1})$. Finally, estimating the loss around the global minimum leads to the conclusion. Detailed proofs are provided in Appendix E.

Summary and simulation experiments. Table 4 summarizes the conclusions in this section, which shows that a complex-valued neuron learns slower than a real-valued one. A complex-valued neuron is more flexible since it can learn the phase. But this flexibility becomes redundant and slows down the convergence when learning a phase-independent function. Our theories are based on the setting of low-dimensional inputs and no bias term, and the simulation results in Fig. 3 verify and extend these discoveries in more general settings. Details about the experiments are available in Appendix H.

6. Reparameterized Complex-valued Neurons Can Efficiently Learn More

In this section, we reparameterize the phase parameter of complex-valued neurons and present learning guarantees for reparameterized complex-valued neurons. Specifically, Subsection 6.1 proposes the reparameterization of the phase parameter. Subsections 6.2 and 6.3 provide convergence rates for reparameterized complex-valued neurons when learning a real-valued neuron and a complex-valued neuron, respectively.

6.1 Reparameterization of the Phase Parameter

In Sections 4 and 5, a complex-valued neuron uses the symmetrical version of the zReLU activation function parameterized by a phase parameter ψ

$$\sigma_\psi(z) = \begin{cases} \mathrm{Re}(z) \ , & \theta_z \in [-\psi, \psi] \ , \\ 0 \ , & \text{otherwise} \ . \end{cases}$$

The phase parameter ψ is directly optimized by projected gradient descent when the learnable neuron is a complex-valued neuron. As mentioned in the proof idea of Theorem 6, the gradient $\nabla_\psi L_{\mathrm{cr}}$ degenerates to order $(\psi - \psi^*)^2$ in Stage IV of learning a real-valued neuron with the global minimum $\psi^* = \pi/2$. Then we have $L_{\mathrm{cr}} \approx (\psi^* - \psi)^3$, i.e., the loss is approximately a cubic function of the deviation $\psi^* - \psi$, whereas a linear convergence rate requires that the loss should be approximately a quadratic function of the deviation.

Motivated by the above observations, we propose the following reparameterization of the phase parameter ψ

$$\psi(\phi) = \psi^* \left[1 - g(\phi)^{2/3}\right] \quad \text{with} \quad \phi \in \mathbb{R} \ , \tag{3}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a mapping satisfying the following two requirements.

1. The range of mapping g satisfies $\mathrm{ran}(g) = [-1, 1]$.
2. The expansion of mapping g around 0 satisfies $g(\phi) = a_1\phi + O(\phi^2)$ with $a_1 \neq 0$. Furthermore, choose $r \in (0, 1/(10a_1))$ such that

$$\frac{a_1}{2} \leqslant g'(\phi) \leqslant 2a_1 \quad \text{for all} \quad \phi \in [-r, r] \ . \tag{4}$$

The reparameterization in Eq. (3) does not restrict the expression of g as long as g satisfies the two requirements. Many elementary functions are potential choices for the mapping g, such as the sine function $g(\phi) = \sin\phi$, hyperbolic tangent function $g(\phi) = \tanh\phi$, and arctangent function $g(\phi) = \arctan\phi$.

The reparameterization in Eq. (3) introduces a new phase parameter ϕ and enjoys at least two desirable properties. First, the new phase parameter ϕ is unbounded. Based on $\mathrm{ran}(g) = [-1, 1]$, the phase parameter ψ is a function of $\phi \in \mathbb{R}$ with range $[0, \psi^*]$. Thus, the reparameterization converts the constrained optimization of ψ to the unconstrained optimization of ϕ, avoiding the necessity of projection after gradient descent. Second, the loss L_{cr} is approximately a quadratic function of the deviation of ϕ in a small angle region. According to the second requirement, one knows that $g(0) = 0$ and $\phi^* = 0$ is the new phase parameter for a real-valued neuron. Then the loss L_{cr} in Stage IV of Theorem 6 satisfies

$$L_{\mathrm{cr}} \approx (\psi^* - \psi)^3 = (\psi^*)^3 g(\phi)^2 \approx a_1^2 (\psi^*)^3 (\phi - \phi^*)^2 \ , \tag{5}$$

i.e., the loss of learning a real-valued neuron using a complex-valued neuron L_{cr} is approximately a quadratic function of $\phi - \phi^*$, making linear convergence rate possible.

6.2 Learning a Real-valued Neuron with Reparameterization

We first consider learning functions expressed by any one real-valued neuron using a reparameterized complex-valued neuron. The expected square loss in Eq. (1) becomes

$$L_{\mathrm{cr}}(\boldsymbol{w}, \psi(\phi)) = \frac{1}{2} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left(\sigma_{\psi(\phi)}(\boldsymbol{w}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) - \tau(\boldsymbol{v}^\top \boldsymbol{x}) \right)^2 \right] , \tag{6}$$

where we abbreviate the phase parameter ϕ_w as ϕ since the target real-valued neuron does not have a phase parameter, and $\psi(\phi)$ is the reparameterization of the phase parameter in Eq. (3). We assume $\|\boldsymbol{v}\| = 1$ without loss of generality. Then we present the first theorem for reparameterized complex-valued neuron learning.

Theorem 7 *Let $d = 1$. Suppose that $\boldsymbol{w}_0 \sim \mathcal{N}(0, \mathbf{I}_2)$ and $\phi_0 \sim \mathcal{N}(0, 1)$. Let $\{(\boldsymbol{w}_t, \phi_t)\}_{t=0}^{\infty}$ denote the parameter sequence of the reparameterized complex-valued neuron generated by gradient descent when optimizing $L_{\mathrm{cr}}(\boldsymbol{w}, \psi(\phi))$ in Eq. (6). If the step size $\eta_t = \eta \in (0, 4)$ and $a_1 \leqslant 1/12$, then we have*

$$\Pr \left[L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi(\phi_t)) \leqslant \frac{1}{3} \left(1 - \frac{a_1^4 \pi^2}{2} \eta \right)^{2t} \right] \geqslant \frac{a_1^{4/3} r^{10/3} \mathrm{e}^{-r^2/2}}{100} \ .$$

Theorem 7 shows that a reparameterized complex-valued neuron can learn functions expressed by one real-valued neuron with a linear convergence rate and constant probability. The condition $a_1 \leqslant 1/12$ can be satisfied by rescaling g, e.g., $g(\phi) = \arctan(\phi/12)$.

Challenges. The main challenge is the maintenance of the small angle condition that guarantees the quadratic approximation in Eq. (5). The loss L_{cr} is a piecewise function around the global minimum. In the small angle region, i.e., the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ converges faster than the phase parameter, the loss satisfies $L_{\mathrm{cr}} \approx (\phi - \phi^*)^2$ as mentioned in Eq. (5). In the large angle region, i.e., the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ converges more slowly than the phase parameter, the loss L_{cr} is approximately a quadratic function of $\psi - \psi^*$ and thus $L_{\mathrm{cr}} \approx (\phi - \phi^*)^{4/3}$. Before reparameterization, the small angle condition is eventually guaranteed as the phase parameter ψ converges more slowly than the angle. However, after reparameterization, the small angle condition may be violated since both the angle and phase parameters have linear convergence rates in the small angle region.

To address this challenge, we raise a mild condition $a_1 \leqslant 1/12$ on the reparameterization to guarantee the small angle condition. On the one hand, the constant a_1 can control the convergence speed of the phase parameter ϕ. Intuitively, the constant a_1 represents the slope of g and thus the flatness of the reparameterization around the global minimum $\phi^* = 0$. A smaller a_1 indicates a flatter landscape and a slower convergence speed. On the other hand, the constant a_1 does not affect the convergence speed of the weight parameter $\boldsymbol{w}$ since reparameterization does not change the optimization of the weight parameter. Thus, we can select a sufficiently small a_1 to make the convergence speed of the phase parameter slower than that of the weight parameter, which guarantees the small angle condition since the angle $\theta_{\boldsymbol{w},\boldsymbol{v}}$ is directly determined by the convergence speed of the weight parameter.

We summarize the proof idea of Theorem 7 as follows. The most important ingredient in the proof is providing a lower bound on the convergence speed of the phase parameter ϕ. As derived in Eq. (5), the loss is approximately a quadratic function of the deviation of ϕ in the small angle region, and the coefficient of the quadratic function is positively related to a_1. Then the convergence of the phase parameter has a lower bound $|\phi_t| \geqslant (1-\Theta(a_1^2))|\phi_{t-1}|$. The main idea of the remaining proof is similar to that of Theorem 1. Firstly, we identify an ideal region where the small angle condition is satisfied. Meanwhile, the lower bound on the convergence speed of ϕ implies that this region is closed under gradient descent during the whole optimization. Secondly, we analyze the convergence rate of the weight parameter $\boldsymbol{w}$ and phase parameter ϕ. Finally, we estimate the convergence of the loss and calculate the probability of random initialization falling into the ideal region to complete the proof. Detailed proofs are provided in Appendix F.

6.3 Learning a Complex-valued Neuron with Reparameterization

We proceed to investigate learning functions expressed by any one complex-valued neuron using a reparameterized complex-valued neuron. Then the expected square loss in Eq. (1) can be reformulated as

$$L_{\mathrm{cc}}(\boldsymbol{w}, \psi(\phi_w)) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\sigma_{\psi(\phi_w)}(\boldsymbol{w}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) - \sigma_{\psi(\phi_v)}(\boldsymbol{v}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}})\right)^2\right],$$

where $(\boldsymbol{v}, \psi(\phi_v))$ is the parameter of the target complex-valued neuron, and $(\boldsymbol{w}, \psi(\phi_w))$ denotes the learnable parameter. Without loss of generality, we still assume $\|\boldsymbol{v}\| = 1$. Here, we use gradient descent with vanishing step size $x_{t+1} = x_t - \eta_t \nabla f(x_t)$, where the positive step size η_t satisfies $\eta_t \to 0$ as $t \to \infty$. Then we present the second theorem for reparameterized complex-valued neuron learning.

Theorem 8 *Let $d = 1$, and $\psi(\phi_v) \in [7\pi/20, 2\pi/5]$. Suppose that $\boldsymbol{w}_0 \sim \mathcal{N}(0, \mathbf{I}_2)$ and $\phi_0 \sim \mathcal{N}(0,1)$. Let $\{(\boldsymbol{w}_t, \phi_{w,t})\}_{t=0}^{\infty}$ denote the parameter sequence of the reparameterized complex-valued neuron generated by gradient descent when optimizing $L_{\mathrm{cc}}(\boldsymbol{w}, \psi(\phi_w))$, the expected loss of learning a complex-valued neuron using a reparameterized complex-valued neuron. Suppose that g is strictly increasing, K-Lipschitz continuous, and satisfies $g'(\phi) > k > 0$ for all $\phi \in \{\phi \mid \psi(\phi) \in [7\pi/20 - 2/25, \pi/2]\}$. If we use vanishing step size $\eta_t = \min\{c_1, c_2/t\}$ with $c_1 \leqslant 1/(3000\pi^2 K^2)$ and $c_2 \geqslant \max\{20, 10/k^2\}$, then we have*

$$\Pr\left[L_{\mathrm{cc}}(\boldsymbol{w}, \psi(\phi_w)) \leqslant \frac{100c_2^3(K+1)^2}{c_1 t}\right] \geqslant \frac{\mathrm{e}^{-1/k^2}}{10^6 K}\ .$$

Table 5: A reparameterized complex-valued neuron can efficiently learn more.

Target	**Real-valued Neuron**	**Reparameterized CVN**
Real-valued Neuron	$O(\mathrm{e}^{-c_1 t})$ (Lemma 5)[1]	$O(\mathrm{e}^{-c_2 t})$ (Theorem 7)
Complex-valued Neuron	Cannot Learn (Theorem 4)	$O(t^{-1})$ (Theorem 8)

[1] Theorem 6.4 in (Yehudai and Shamir, 2020).

Theorem 8 demonstrates that a reparameterized complex-valued neuron can learn functions expressed by one complex-valued neuron with convergence rate $O(t^{-1})$ and constant probability. The strictly increasing, Lipschitz, and positive gradient conditions can be satisfied by many continuously differentiable sigmoidal functions.

Challenges. The proof inherits all challenges and assumptions of learning a complex-valued neuron before reparameterization, as discussed after Theorem 2. Besides, the reparameterization brings new challenges. First, an improper reparameterization may introduce new spurious local minima. The chain rule says $\nabla_\phi L_{\mathrm{cc}} = \nabla_\psi L_{\mathrm{cc}} \psi'(\phi)$, which implies that a stationary point of the loss after reparameterization is a stationary point of the loss before reparameterization or a stationary point of the reparameterization function. Second, reparameterization changes the landscape of the loss, and an improper reparameterization may cause undesired optimization behaviors, such as oscillation or slow convergence speed.

To tackle these issues, we focus on proper reparameterizations whose g enjoys strict monotonicity and bounded gradients. The strict monotonicity of g, together with $g(0) = 0$, ensures that $\phi = 0$ is the only stationary point of $\psi(\phi)$. As discussed after Theorem 2, $\psi = \pi/2$, which corresponds to $\phi = 0$, is a stationary point of L_{cc}. Thus, strict monotonicity prevents the introduction of new stationary points. The bounded gradients of g imply lower and upper bounds of $\psi'(\phi)$ and restrict the distortion of the loss landscape. Then the gradient $\nabla_\phi L_{\mathrm{cc}}$ is close to the gradient $\nabla_\psi L_{\mathrm{cc}}$. Thus, the bounded gradients avoid the risks of oscillation and slow convergence speed.

The proof idea of Theorem 8 is similar to that of Theorem 2. We first define an ideal region that is closed during optimization. Then we analyze the three stages of convergence, where the angle $\theta_{w,v}$, the phase ϕ, and the weight $\boldsymbol{w}$ begin to converge to the global minimum in the corresponding stages. Finally, we estimate the convergence of loss and the probability of the ideal region. The monotonicity implies local invertibility of ψ, which makes the ideal region well-defined and the global minimum unique. The bounded gradients indicate a smooth distortion of the loss landscape, which guarantees the boundedness of the ideal region and maintains the convergence rate of the phase parameter.

Summary and simulation experiments. We summarize the main conclusions of this section in Table 5. A reparameterized complex-valued neuron can learn any real-valued neuron faster than an unreparameterized one and as efficiently as a real-valued neuron. Meanwhile, a reparameterized complex-valued neuron can learn any complex-valued neuron, which cannot be achieved by a two-layer RVNN with finite width. These theoretical results confirm that the phase parameter provides a stronger learning capability for complex-valued neurons, and the reparameterization helps complex-valued neurons learn efficiently. Our theoretical conclusions are based on the setting of low-dimensional inputs and no bias term, and the simulation results in Figures 3 and 5 verify and extend these findings in more general settings. Experimental details are provided in Appendix H.

7. Conclusions and Prospects

In this paper, we investigate the problem of learning a single neuron using another neuron by optimizing the expected square loss via gradient descent. Firstly, we prove that complex-valued neurons can learn more than real-valued neurons since CVNNs benefit from the phase parameter, which helps CVNNs learn phase information more efficiently. Secondly, we show that complex-valued neurons learn slower than real-valued neurons in phase-independent tasks. This phenomenon captures the additional price for learning simpler tasks with more complicated models, where the redundant phase consideration exponentially slows down the convergence. Thirdly, we prove that a reparameterized complex-valued neuron can efficiently learn more than a real-valued neuron, which highlights the importance of reparameterization for the phase parameter. These theoretical conclusions imply that complex-valued operations can be more suitable in phase-related tasks. Meanwhile, new training techniques are important to accelerate the convergence rate of learning phase information, and reparameterization may be a feasible approach.

Our study serves as a preliminary attempt to compare the learning process of artificial neural networks with different functional operations. The current theories assumes 2-dimensional inputs, i.e., $d = 1$, to leverage the closed-form expression of the expected loss. It is much more complicated to precisely calculate the expected loss and analyze the gradient in high-dimensional scenarios since a 4-dimensional input space is required to fully characterize the learning process of complex-valued neurons, which may require new theoretical techniques to simplify the analysis. In the future, it is important to extend our theoretical results to more general settings, such as cases of high-dimensional inputs, equipped with bias terms, and over-parameterized architectures (Zhou, 2021). Meanwhile, it is prospective to investigate complex-valued neuron learning from finite samples and derive a high-probability convergence condition. Since the empirical loss is a piecewise constant function with respect to the learnable phase parameter, it might be necessary to explore new learning algorithms, which is also encouraged by the neural tangent kernel aspect (Tan et al., 2022). Besides, it is promising to consider the more practical and challenging procedure of learning general functions with deep architectures.

Acknowledgments

This research was supported by NSFC (62576165, 62406138) and JSF Leading-edge Technology Program (BK20232003). We would like to thank the anonymous reviewers and the action editor for their invaluable comments.

Appendix A. Useful Lemmas

The next two lemmas are helpful for calculating the closed form of the expected loss.

Lemma 9 *Let $d = 1$. For any $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{2d}$, and $a \leqslant b \leqslant a + 2\pi$, we have*

$$\begin{aligned} A(\boldsymbol{w}, \boldsymbol{v}, a, b) &= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\boldsymbol{w}^\top \boldsymbol{x} \cdot \boldsymbol{v}^\top \boldsymbol{x} \cdot \mathbb{I}(\theta_x \in [a, b]) \right] \\ &= \frac{\|\boldsymbol{w}\| \|\boldsymbol{v}\|}{4\pi} \left[2(b - a) \cos \theta_{\boldsymbol{w}, \boldsymbol{v}} + \sin(\theta_{\boldsymbol{w}} + \theta_{\boldsymbol{v}} - 2a) - \sin(\theta_{\boldsymbol{w}} + \theta_{\boldsymbol{v}} - 2b) \right] . \end{aligned}$$

Proof. According to the probability density function of Gaussian distribution, we can calculate A in the polar coordinate system as

$$\begin{aligned} A(\boldsymbol{w},\boldsymbol{v},a,b) &= \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{2\pi}\int_0^\infty \int_a^b r^3 \mathrm{e}^{-\frac{1}{2}r^2} \cos(\theta_{\boldsymbol{w}}-\phi)\cos(\theta_{\boldsymbol{v}}-\phi)\ \mathrm{d}\phi\ \mathrm{d}r \\ &= \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{\pi}\int_a^b \cos(\theta_{\boldsymbol{w}}-\phi)\cos(\theta_{\boldsymbol{v}}-\phi)\ \mathrm{d}\phi \\ &= \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{4\pi}\left[2(b-a)\cos\theta_{\boldsymbol{w},\boldsymbol{v}} + \sin(\theta_{\boldsymbol{w}}+\theta_{\boldsymbol{v}}-2a) - \sin(\theta_{\boldsymbol{w}}+\theta_{\boldsymbol{v}}-2b)\right]\ , \end{aligned}$$

where the second and third equalities hold from integrating over r and ϕ, respectively. Thus, we have completed the proof. □

Lemma 10 *Let $d=1$. For any $\boldsymbol{w},\boldsymbol{v}\in\mathbb{R}^{2d}$, denote by $\theta=\theta_{\boldsymbol{w},\boldsymbol{v}}$ the angle between $\boldsymbol{w}$ and $\boldsymbol{v}$. Then for any $\psi_w,\psi_v\in[0,\pi/2]$, define $\psi_m=\min\{\psi_w,\psi_v\}$. Then we have*

$$\begin{aligned} &B(\boldsymbol{w},\boldsymbol{v},\psi_w,\psi_v) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sigma_{\psi_w}(\boldsymbol{w}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})\sigma_{\psi_v}(\boldsymbol{v}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})\right] \\ &= \begin{cases} \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{2\pi}\cos\theta_{\boldsymbol{w},\boldsymbol{v}}[2\psi_m+\sin(2\psi_m)]\ , & \theta_{\boldsymbol{w},\boldsymbol{v}}\in[0,|\psi_v-\psi_w|]\ , \\ \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{4\pi}[2(\psi_w+\psi_v-\theta_{\boldsymbol{w},\boldsymbol{v}})\cos\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_v) & \\ \quad -\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_w)]\ , & \theta_{\boldsymbol{w},\boldsymbol{v}}\in[|\psi_v-\psi_w|,\psi_v+\psi_w]\ , \\ 0\ , & \theta_{\boldsymbol{w},\boldsymbol{v}}\in[\psi_v+\psi_w,\pi]\ . \end{cases} \end{aligned}$$

Proof. We only consider the case of $\psi_w\leqslant\psi_v$. The other case $\psi_w\geqslant\psi_v$ can be proven similarly. We prove the conclusion by discussion.

1. Suppose $\theta_{\boldsymbol{w},\boldsymbol{v}}\in[0,\psi_v-\psi_w]$. Then Lemma 9 leads to

$$B(\boldsymbol{w},\boldsymbol{v},\psi_w,\psi_v) = A(\boldsymbol{w},\boldsymbol{v},\theta_{\boldsymbol{w}}-\psi_w,\theta_{\boldsymbol{w}}+\psi_w) = \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{2\pi}\cos\theta_{\boldsymbol{w},\boldsymbol{v}}[2\psi_w+\sin(2\psi_w)]\ .$$

2. Suppose $\theta_{\boldsymbol{w},\boldsymbol{v}}\in[\psi_v-\psi_w,\psi_v+\psi_w]$ and $\theta_{\boldsymbol{w}}\leqslant\theta_{\boldsymbol{v}}$. Then one knows from Lemma 9 that

$$\begin{aligned} B(\boldsymbol{w},\boldsymbol{v},\psi_w,\psi_v) &= A(\boldsymbol{w},\boldsymbol{v},\theta_{\boldsymbol{v}}-\psi_v,\theta_{\boldsymbol{w}}+\psi_w) \\ &= \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{4\pi}[2(\psi_w+\psi_v-\theta_{\boldsymbol{w},\boldsymbol{v}})\cos\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_v)-\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_w)]\ . \end{aligned}$$

3. Suppose $\theta_{\boldsymbol{w},\boldsymbol{v}}\in[\psi_v-\psi_w,\psi_v+\psi_w]$ and $\theta_{\boldsymbol{w}}\geqslant\theta_{\boldsymbol{v}}$. Based on Lemma 9, we have

$$\begin{aligned} B(\boldsymbol{w},\boldsymbol{v},\psi_w,\psi_v) &= A(\boldsymbol{w},\boldsymbol{v},\theta_{\boldsymbol{w}}-\psi_w,\theta_{\boldsymbol{v}}+\psi_v) \\ &= \frac{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}{4\pi}[2(\psi_w+\psi_v-\theta_{\boldsymbol{w},\boldsymbol{v}})\cos\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_v)-\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}-2\psi_w)]\ . \end{aligned}$$

4. Suppose $\theta_{\boldsymbol{w},\boldsymbol{v}}\in[\psi_v+\psi_w,\pi]$. Then the support of $\sigma_{\psi_w}(\boldsymbol{w}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})$ does not overlap with that of $\sigma_{\psi_v}(\boldsymbol{v}_{\mathbb{C}}^\top\overline{\boldsymbol{x}}_{\mathbb{C}})$, which leads to $B(\boldsymbol{w},\boldsymbol{v},\psi_w,\psi_v)=0$.

Combining the cases above completes the proof. □

Appendix B. Proof of Theorem 1

In the main part of this section, we provide the closed form of the loss, definition of the ideal region, and the detailed proof of Theorem 1. Subsection B.1 presents the optimization behaviors in the ideal region. Subsection B.2 proves several convergence rate lemmas. Subsection B.3 gives some technical lemmas to bound small terms in the proof.

Proof of Theorem 1. Let $\boldsymbol{w} = (w_1, w_2)$. By spherical symmetry, we assume $\boldsymbol{v} = (1, 0)$ without loss of generality. According to Lemma 10, the expected loss can be calculated by

$$
\begin{aligned}
L_{\mathrm{cr}}(\boldsymbol{w}, \psi) &= \frac{1}{2} B(\boldsymbol{w}, \boldsymbol{w}, \psi, \psi) - B(\boldsymbol{w}, \boldsymbol{v}, \psi, \pi/2) + \frac{1}{2} B(\boldsymbol{v}, \boldsymbol{v}, \pi/2, \pi/2) \\
&= \left\{ \begin{array}{ll}
\frac{1}{4} - \frac{1}{4\pi}[\sin(2\psi) + 2\psi][1 - (w_1 - 1)^2 - w_2^2] \ , & \theta \in [0, \pi/2 - \psi] \ , \\
\frac{1}{4} - \frac{1}{2\pi}[\frac{1}{2}\sin(2\psi) w_1 - \frac{1}{2}\cos(2\psi)|w_2| + \frac{1}{2}|w_2| + (\frac{\pi}{2} + \psi - \theta) w_1] & \\
\quad + \frac{1}{4\pi}[\sin(2\psi) + 2\psi](w_1^2 + w_2^2) \ , & \theta \in (\pi/2 - \psi, \pi/2 + \psi) \ , \\
\frac{1}{4} + \frac{1}{4\pi}[2\psi + \sin(2\psi)](w_1^2 + w_2^2) \ , & \theta \in [\pi/2 + \psi, \pi] \ ,
\end{array} \right.
\end{aligned}
\tag{7}
$$

where $\theta = \theta_{\boldsymbol{w},\boldsymbol{v}} = \arccos(w_1/\sqrt{w_1^2 + w_2^2})$. For any $R \in (0, 1)$, define

$$
\begin{aligned}
D_1 &= \{(\boldsymbol{w}, \psi) \mid \|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R, \psi \in [0, \pi/2], \theta \in [0, \pi/2 - \psi]\} \ , \\
D_2 &= \{(\boldsymbol{w}, \psi) \mid \|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R, \psi \in [0, \pi/2], \theta \in (\pi/2 - \psi, \pi/2 + \psi)\} \ .
\end{aligned}
\tag{8}
$$

Let $D = D_1 \cup D_2 = \{(\boldsymbol{w}, \psi) \mid \|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R, \psi \in [0, \pi/2], \theta \in [0, \pi/2 + \psi]\}$ denote the ideal region. The rest of the proof procedure consists of four steps.

Step 1: D is closed under gradient descent. We first prove the maintenance of inclusion by mathematical induction, i.e., $(\boldsymbol{w}_0, \psi_0) \in D$ indicates $(\boldsymbol{w}_t, \psi_t) \in D$.

1. Base case. The conclusion holds for $t = 0$ from the condition.

2. Induction. Suppose that the conclusion holds for $t = k$ with $k \in \mathbb{N}$. Then based on Lemmas 13 and 14, one knows

$$
-6(\psi^* - \psi_k) \leqslant \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi_k) \leqslant -\frac{1 - R^2}{4\pi} (\psi^* - \psi_k)^2 \leqslant 0 \ , \tag{9}
$$

where $\psi^* = \pi/2$, the first inequality holds based on the induction hypothesis and $|w_{2,k}| \leqslant 1$, and the third inequality holds from $R < 1$. Thus, the updating rule $\psi_{k+1} = \psi_k - \eta \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi_k)$ with $\eta \in (0, 1/(12\pi))$ leads to

$$
\frac{\pi}{2} \geqslant \psi^* - \psi_k \geqslant \psi^* - \psi_{k+1} \geqslant (1 - 6\eta)(\psi^* - \psi_k) \geqslant 0 \ , \tag{10}
$$

where the first and fourth inequalities hold from the induction hypothesis. Meanwhile, Lemmas 11 and 12 imply

$$
\|\boldsymbol{w}_{k+1} - \boldsymbol{v}\| \leqslant \Big(1 - \frac{\eta}{24\pi}[\sin(2\psi_k) + 2\psi_k]\Big) \|\boldsymbol{w}_k - \boldsymbol{v}\| \leqslant R \ . \tag{11}
$$

Combining Eqs. (10) and (11), the conclusion holds for $t = k + 1$.

Therefore, mathematical induction implies $(\boldsymbol{w}_t, \psi_t) \in D$ when $(\boldsymbol{w}_0, \psi_0) \in D$.

Step 2: parameters converge to the global minimum in D. The convergence process consists of two stages. In stage I, we deal with the convergence of ψ when $(\boldsymbol{w}_0, \psi_0) \in D$. Based on Eq. (9) and the updating rule $\psi_{k+1} = \psi_k - \eta \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi_k)$, one knows

$$\psi^* - \psi_{t+1} \leqslant (\psi^* - \psi_t) \left[1 - \frac{\eta(1-R^2)}{4\pi}(\psi^* - \psi_t) \right] .$$

Define $a_t = \eta(1-R^2)(\psi^* - \psi_t)/(4\pi)$. Then we obtain $a_{t+1} \leqslant a_t(1-a_t)$. From $\psi^* - \psi_t \in [0, \pi/2]$ and $\eta < 1/(12\pi) \leqslant 4$, one knows $a_t \in [0, 1/2]$. Applying Lemma 16 to a_t leads to

$$\psi^* - \psi_t = \frac{4\pi a_t}{\eta(1-R^2)} \leqslant \frac{4\pi}{\eta(1-R^2)(t+1)} \ . \tag{12}$$

In stage II, we consider the convergence of $\boldsymbol{w}$ when $(\boldsymbol{w}_0, \psi_0) \in D$. Based on Eq. (12), choosing $T_1 \geqslant 16 \lceil \eta(1-R^2) \rceil^{-1}$ leads to $\psi^* - \psi_t \leqslant \pi/4$ for any $t \geqslant T_1$, i.e., $\psi_t \geqslant \pi/4$ for any $t \geqslant T_1$. Thus, for any $t \geqslant T_1$, Eq. (11) indicates

$$\|\boldsymbol{w}_t - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{48}\right) \|\boldsymbol{w}_{t-1} - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{48}\right)^{t-T_1} \ , \tag{13}$$

where the first inequality holds from the monotonic increasing of $\sin(x) + x$ and $\psi_t \geqslant \pi/4$, and the second inequality holds because of $\|\boldsymbol{w}_{T_1} - \boldsymbol{v}\| \leqslant R < 1$.

Step 3: the loss converges to 0 in D. We estimate the convergence of the expected loss when $(\boldsymbol{w}_0, \psi_0) \in D$. For any $(\boldsymbol{w}, \psi) \in D$, define non-negative quantities $\Delta_{\boldsymbol{w}} = \|\boldsymbol{w} - \boldsymbol{v}\|$ and $\Delta_\psi = \psi^* - \psi$. We provide an upper bound for L_{cr} by discussion.

1. Suppose $(\boldsymbol{w}, \psi) \in D_1$. Then we have

$$L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \leqslant \frac{1}{4} - \frac{1}{2\pi}(\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) \leqslant \frac{1}{2\pi}\Delta_\psi^3 + \frac{1}{4}\Delta_{\boldsymbol{w}}^2 \ , \tag{14}$$

where the first inequality holds based on $\sin(2\psi) + 2\psi = \sin(2\Delta_\psi) + 2\psi^* - 2\Delta_\psi \geqslant 2\psi^* - 2\Delta_\psi^3$, and the second inequality holds from non-negative Δ_ψ.

2. Suppose $(\boldsymbol{w}, \psi) \in D_2$. The expected loss can be rewritten as

$$\begin{aligned}
L_{\mathrm{cr}}(\boldsymbol{w}, \psi) &= \frac{1}{4} - \frac{1}{4\pi}[\sin(2\psi) + 2\psi](1 - \Delta_{\boldsymbol{w}}^2) \\
&\quad + \frac{1}{4\pi}[(\cos(2\psi) - 1)|w_2| + (\sin(2\psi) + 2\psi + 2\theta - 2\psi^*)w_1] \\
&\leqslant \frac{1}{4} - \frac{1}{2\pi}(\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) + \frac{1}{4\pi}[(\pi + 2\theta - 2\psi^*)w_1] \\
&\leqslant \frac{1}{4} - \frac{1}{2\pi}(\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) + \frac{1}{2\pi}\Delta_{\boldsymbol{w}}(1 + \Delta_{\boldsymbol{w}}) \\
&\leqslant \frac{1}{2\pi}\Delta_\psi^3 + \frac{1}{2\pi}\Delta_{\boldsymbol{w}} + \frac{1}{2}\Delta_{\boldsymbol{w}}^2 \ ,
\end{aligned} \tag{15}$$

where the first inequality holds from $\pi \geqslant \sin(2\psi) + 2\psi \geqslant 2\psi^* - 2\Delta_\psi^3$ and $\cos(2\psi) - 1 \leqslant 0$, the second inequality holds based on $\theta \leqslant \tan\theta \leqslant \Delta_{\boldsymbol{w}}$ and $w_1 \leqslant 1 + \Delta_{\boldsymbol{w}}$, and the third inequality holds from $\Delta_\psi \geqslant 0$.

Combining Eqs. (14) and (15), for any $(\boldsymbol{w}_0, \psi_0) \in D$ and $t \geqslant T_1$, $\Delta_{\boldsymbol{w}}^2 \leqslant \Delta_{\boldsymbol{w}}$ implies

$$L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \leqslant \frac{1}{2\pi}\Delta_{\psi,t}^3 + \Delta_{\boldsymbol{w},t} \leqslant \frac{32\pi^3}{\eta^3(1-R^2)^3t^3} + \left(1 - \frac{\eta}{48}\right)^{t-T_1} \ , \tag{16}$$

where the second inequality holds by Eqs. (12) and (13).

Step 4: initialization falls into D with constant probability. For simplicity, define $p_0 = \Pr[(\boldsymbol{w}_0, \psi_0) \in D]$.. From $\psi_0 \sim \mathcal{U}(0, \pi/2)$, the requirement $\psi \in [0, \pi/2]$ is satisfied. Denote by $p(\boldsymbol{w})$ the probability density function of $\mathcal{N}(0, \mathbf{I}_2)$. Then one has

$$p_0 = \Pr[\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant R] = \int_{\boldsymbol{w} \in B(\boldsymbol{v},R)} p(\boldsymbol{w}) \ \mathrm{d}\boldsymbol{w} \geqslant \mu(B(\boldsymbol{v}, R)) \min_{\boldsymbol{w} \in B(\boldsymbol{v},R)} p(\boldsymbol{w}) \geqslant \frac{R^2}{16} \ . \tag{17}$$

Let $R^2 = 1/2$. We obtain from Eqs. (16) and (17) that

$$\Pr\left[L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \leqslant \frac{8000}{\eta^3 t^3} + \left(1 - \frac{\eta}{48}\right)^{t+1-32/\eta}\right] \geqslant \frac{1}{32} \ ,$$

which completes the proof. □

B.1 Optimization Behaviors

The following two lemmas indicate the linear convergence of $\boldsymbol{w}$ in D_1 and D_2, respectively.

Lemma 11 *Let $\boldsymbol{w}' = \boldsymbol{w} - \eta\nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}, \psi)$. If $(\boldsymbol{w}, \psi) \in D_1$ and $\eta \in (0, 4)$, then we have*

$$\|\boldsymbol{w}' - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{4\pi}[\sin(2\psi) + 2\psi]\right) \|\boldsymbol{w} - \boldsymbol{v}\| \ .$$

Proof. For any $(\boldsymbol{w}, \psi) \in D_1$, one has

$$\langle \nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}, \psi), \boldsymbol{w} - \boldsymbol{v} \rangle = \left\langle \frac{1}{4\pi}[\sin(2\psi) + 2\psi](\boldsymbol{w} - \boldsymbol{v}), \boldsymbol{w} - \boldsymbol{v} \right\rangle = \frac{1}{4\pi}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ .$$

Meanwhile,

$$\|\nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}, \psi)\|^2 = \frac{1}{(4\pi)^2}[\sin(2\psi) + 2\psi]^2 \|(\boldsymbol{w} - \boldsymbol{v})\|^2 \ .$$

Then according to Lemma 15 and $\psi \in [0, \pi/2]$, for any $\eta \in (0, 4)$, one has

$$\|\boldsymbol{w}' - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{4\pi}[\sin(2\psi) + 2\psi]\right) \|\boldsymbol{w} - \boldsymbol{v}\| \ ,$$

which completes the proof. □

Lemma 12 *Let $\boldsymbol{w}' = \boldsymbol{w} - \eta\nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}, \psi)$. If $(\boldsymbol{w}, \psi) \in D_2$ and $\eta \in (0, 1/(12\pi))$, then*

$$\|\boldsymbol{w}' - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{24\pi}[\sin(2\psi) + 2\psi]\right) \|\boldsymbol{w} - \boldsymbol{v}\| \ .$$

Proof. Firstly, we prove the strong convexity in D_2. For any $(\boldsymbol{w}, \psi) \in D_2$, one has

$$\begin{aligned}
&2\pi\langle \nabla_{\boldsymbol{w}} L_{\rm cr}(\boldsymbol{w}, \psi), \boldsymbol{w} - \boldsymbol{v}\rangle \\
= &-\left[\frac{1}{2}\sin(2\psi) + \left(\frac{\pi}{2} + \psi - \theta\right) + \frac{w_1|w_2|}{w_1^2 + w_2^2}\right](w_1 - 1) + [\sin(2\psi) + 2\psi]w_1(w_1 - 1) \\
&-\left[-\frac{1}{2}\cos(2\psi) + \frac{1}{2} - \frac{w_1^2}{w_1^2 + w_2^2}\right]|w_2| + [\sin(2\psi) + 2\psi]w_2^2 \\
= &[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 - R_1 - R_2 \ ,
\end{aligned} \tag{18}$$

where

$$R_1 = \left[\left(\frac{\pi}{2} - \psi - \theta\right) - \frac{1}{2}\sin(2\psi)\right](w_1 - 1) \quad \text{and} \quad R_2 = \left[\frac{1}{2} - \frac{1}{2}\cos(2\psi) - \frac{w_1}{w_1^2 + w_2^2}\right]|w_2| \ .$$

According to Lemmas 17 and 18, Eq. (18) can be bounded by

$$\langle \nabla_{\boldsymbol{w}} L_{\rm cr}(\boldsymbol{w}, \psi), \boldsymbol{w} - \boldsymbol{v}\rangle \geqslant \frac{1}{2\pi}\left(\frac{1}{2} - \frac{1}{\pi}\right)[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \geqslant \frac{1}{12\pi}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ . \tag{19}$$

Secondly, we provide an upper bound of gradient in D_2. For any $(\boldsymbol{w}, \psi) \in D_2$, one has

$$4\pi^2\|\nabla_{\boldsymbol{w}} L_{\rm cr}(\boldsymbol{w}, \psi)\|^2 = T_1 + T_2 \ ,$$

where

$$\begin{aligned}
T_1 &= \left([\sin(2\psi) + 2\psi]w_1 - \frac{1}{2}\sin(2\psi) - \left(\frac{\pi}{2} + \psi - \theta\right) - \frac{w_1|w_2|}{w_1^2 + w_2^2}\right)^2 \ , \\
T_2 &= \left(\left[\frac{1}{2}\cos(2\psi) - \frac{1}{2} + \frac{w_1^2}{w_1^2 + w_2^2}\right]\operatorname{sgn}(w_2) + [\sin(2\psi) + 2\psi]w_2\right)^2 \ .
\end{aligned} \tag{20}$$

From Lemmas 19 and 20, one knows

$$\|\nabla_{\boldsymbol{w}} L_{\rm cr}(\boldsymbol{w}, \psi)\|^2 \leqslant [\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ . \tag{21}$$

Finally, based on Eqs. (19) and (21) and Lemma 15, we conclude

$$\|\boldsymbol{w}' - \boldsymbol{v}\| \leqslant \sqrt{1 - \left(\frac{1}{6\pi} - \eta\right)\eta[\sin(2\psi) + 2\psi]}\|\boldsymbol{w} - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{24\pi}[\sin(2\psi) + 2\psi]\right)\|\boldsymbol{w} - \boldsymbol{v}\| \ ,$$

where the first inequality holds based on $\sqrt{1-x} \leqslant 1 - x/2$ for any $x \in [0,1]$ and $\eta \in (0, 1/(12\pi))$. Thus, we have completed the proof. □

The next two lemmas depict the gradient with respect to ψ in D_1 and D_2, respectively.

Lemma 13 *Let $\psi' = \psi - \eta\nabla_\psi L_{\rm cr}(\boldsymbol{w}, \psi)$. If $(\boldsymbol{w}, \psi) \in D_1$, then*

$$-\frac{1}{\pi}\left(\frac{\pi}{2} - \psi\right)^2 \leqslant \nabla_\psi L_{\rm cr}(\boldsymbol{w}, \psi) \leqslant -\frac{1 - R^2}{4\pi}\left(\frac{\pi}{2} - \psi\right)^2 \ .$$

Proof. For any $(\boldsymbol{w}, \psi) \in D_1$, one has

$$\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) = -\frac{1}{2\pi}[\cos(2\psi) + 1](1 - \|\boldsymbol{w} - \boldsymbol{v}\|^2) \ .$$

For any $\psi \in [0, \pi/2]$, we have $\frac{1}{2}(\pi/2 - \psi)^2 \leqslant \cos(2\psi) + 1 \leqslant 2(\pi/2 - \psi)^2$. Meanwhile, one has $0 \leqslant \|\boldsymbol{w}_t - \boldsymbol{v}\| \leqslant R$. Thus, the gradient with respect to ψ can be bounded by

$$-\frac{1}{\pi}\left(\frac{\pi}{2} - \psi\right)^2 \leqslant \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \leqslant -\frac{1 - R^2}{4\pi}\left(\frac{\pi}{2} - \psi\right)^2 \ ,$$

which completes the proof of the lower bound. □

Lemma 14 *If $(\boldsymbol{w}, \psi) \in D_2$, then*

$$-2\left(\frac{\pi}{2} - \psi\right)^2 - 2\left(\frac{\pi}{2} - \psi\right)|w_2| \leqslant \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \leqslant -\frac{1 - R^2}{2}\left(\frac{\pi}{2} - \psi\right)^2 \ .$$

Proof. The gradient of L_{cr} with respect to ψ in D_2 can be calculated by

$$\begin{aligned} 2\pi\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) &= [1 + \cos(2\psi)]w_1^2 - [1 + \cos(2\psi)]w_1 + [1 + \cos(2\psi)]w_2^2 - \sin(2\psi)|w_2| \\ &= [1 + \cos(2\psi)][\|\boldsymbol{w} - \boldsymbol{v}\|^2 - 1] + [1 + \cos(2\psi)]w_1 - \sin(2\psi)|w_2| \ . \end{aligned} \tag{22}$$

Firstly, we prove the upper bound for $\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi)$. It is observed that

$$[1 + \cos(2\psi)]w_1 - \sin(2\psi)|w_2| \leqslant 2\cos\psi(w_1 \sin\theta - |w_2|\cos\theta) = 0 \ ,$$

where the first inequality holds based on $\pi/2 \geqslant \psi \geqslant \pi/2 - \theta \geqslant 0$, and the first equality holds from $w_1 = r\cos\theta$ and $|w_2| = r\sin\theta$. Substituting Eq. (30) into Eq. (22), we obtain

$$2\pi\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \leqslant [1 + \cos(2\psi)][\|\boldsymbol{w} - \boldsymbol{v}\|^2 - 1] \leqslant -\frac{1 - R^2}{2}\left(\frac{\pi}{2} - \psi\right)^2 \ ,$$

where the second inequality holds by $1 + \cos(2\psi) \geqslant \frac{1}{2}(\pi/2 - \psi)^2$ and $\|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R$.

Secondly, we verify the lower bound for $\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi)$. It is observed that

$$2\pi\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \geqslant -[1 + \cos(2\psi)] - \sin(2\psi)|w_2| \geqslant -2\left(\frac{\pi}{2} - \psi\right)^2 - 2\left(\frac{\pi}{2} - \psi\right)|w_2| \ ,$$

where the first inequality holds because of $[1 + \cos(2\psi)]w_1 \geqslant 0$ and $\|\boldsymbol{w} - \boldsymbol{v}\| \geqslant 0$, and the second inequality holds according to $1 + \cos(2\psi) \leqslant 2(\pi/2 - \psi)^2$ and $\sin(2\psi) \leqslant \pi - 2\psi$ for $\psi \in [0, \pi/2]$. Thus, we have completed the proof. □

B.2 Convergence Rate Lemmas

Lemma 15 provides a sufficient condition for linear convergence of gradient descent.

Lemma 15 *If there exist two constants c_1 and c_2 such that*

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geqslant c_1\|\boldsymbol{w} - \boldsymbol{v}\|^2 \quad \textit{and} \quad \|\nabla f(\boldsymbol{w})\|^2 \leqslant c_2\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ ,$$

then $\boldsymbol{w}' = \boldsymbol{w} - \eta\nabla f(\boldsymbol{w})$ with $\eta \in (0, 2c_1/c_2)$ and $c = \sqrt{1 - 2c_1\eta + c_2\eta^2} \in (0, 1)$ satisfies

$$\|\boldsymbol{w}' - \boldsymbol{v}\| \leqslant c\|\boldsymbol{w} - \boldsymbol{v}\| \ .$$

Proof. One knows from $\boldsymbol{w}' = \boldsymbol{w} - \eta\nabla f(\boldsymbol{w})$ that

$$\begin{aligned}\|\boldsymbol{w}' - \boldsymbol{v}\|^2 &= \|\boldsymbol{w} - \boldsymbol{v}\|^2 - 2\eta\langle\nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v}\rangle + \eta^2\|\nabla f(\boldsymbol{w})\|^2 \\ &\leqslant (1 - 2c_1\eta + c_2\eta^2)\|\boldsymbol{w} - \boldsymbol{v}\|^2 \\ &\leqslant \|\boldsymbol{w} - \boldsymbol{v}\|^2 \ ,\end{aligned}$$

where the last inequality holds since $\eta \in (0, 2c_1/c_2)$. Thus, we have completed the proof. □

Lemma 16 gives sufficient conditions for convergence with an inversely proportional rate.

Lemma 16 *Let $\{a_t\}_{t=0}^{\infty} \subset [0, 1/2]$ represent a real-valued sequence. If $a_{t+1} \leqslant a_t(1 - a_t)$, then $a_t \leqslant (t+1)^{-1}$. If $a_{t+1} \geqslant a_t(1 - a_t)$, then $a_t \geqslant a_0(t+1)^{-1}$.*

Proof. We prove the first conclusion by mathematical induction.

1. Base case. For $t = 0$, the conclusion holds from $a_0 \leqslant 1/2 \leqslant 1$.
2. Induction. Suppose that the conclusion holds for $t = k$ with $k \in \mathbb{N}$. Then one has

$$a_{t+1} \leqslant \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right) = \frac{k}{(k+1)^2} \leqslant \frac{1}{k+2} \ ,$$

where the first inequality holds from the induction hypothesis and the monotonicity of $x(1-x)$ for $x \in [0, 1/2]$. Thus, the conclusion holds for $t = k + 1$.

Therefore, mathematical induction completes the proof of the first conclusion.

We proceed to verify the second conclusion by mathematical induction.

1. Base case. For $t = 0$, the conclusion holds from $a_0 \geqslant a_0$.
2. Induction. Suppose that the conclusion holds for $t = k$ with $k \in \mathbb{N}$. Then one has

$$a_{t+1} \geqslant \frac{a_0}{k+1}\left(1 - \frac{a_0}{k+1}\right) = \frac{a_0(k+1-a_0)}{(k+1)^2} \geqslant \frac{a_0}{k+2} \ ,$$

where the first inequality holds from the induction hypothesis and the monotonicity of $x(1-x)$ for $x \in [0, 1/2]$, and the second inequality holds based on $a_0 \leqslant 1/2$. Thus, the conclusion holds for $t = k + 1$.

Therefore, mathematical induction completes the proof. □

B.3 Technical Lemmas

We present upper bounds for some small terms used in the proof.

Lemma 17 *Let $R_1 = \left[\left(\frac{\pi}{2} - \psi - \theta\right) - \frac{1}{2}\sin(2\psi)\right](w_1 - 1)$. If $(\boldsymbol{w}, \psi) \in D_2$, then*

$$R_1 \leqslant \frac{1}{2}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ .$$

Proof. Let $r = \sqrt{w_1^2 + w_2^2}$ be the norm of $\boldsymbol{w}$. By the definition of θ, one has $w_1 = r\cos\theta$ and $|w_2| = r\sin\theta$. Thus, we can rewrite R_1 as $R_1 = \left[(\pi/2 - \psi - \theta) - 2^{-1}\sin(2\psi)\right](r\cos\theta - 1)$. We provide the upper bound for R_1 by discussion.

1. Suppose $r\cos\theta - 1 \geqslant 0$. Based on the definition of D_2, we have $\frac{\pi}{2} - \psi - \theta \leqslant 0$. Meanwhile, $\psi \in [0, \pi/2]$ indicates $\sin(2\psi) \geqslant 0$. Thus, one knows $R_1 \leqslant 0$.

2. Suppose $r\cos\theta - 1 < 0$. R_1 can be rewritten as

$$R_1 = \frac{1}{2}[\sin(2\psi) + 2\psi](1 - 2r\cos\theta + r^2) + \widetilde{R} \ , \tag{23}$$

where $\widetilde{R} = 2^{-1}[\sin(2\psi) + 2\psi]r(\cos\theta - r) + (\pi/2 - \theta)\,(r\cos\theta - 1)$. If $\cos\theta - r \leqslant 0$, it is observed that $\widetilde{R} \leqslant 0$ because of $\psi, \theta \in [0, \pi/2]$ and $r\cos\theta - 1 < 0$. If $\cos\theta - r > 0$, then

$$\widetilde{R} \leqslant \frac{\pi}{2}r(\cos\theta - r) + \left(\frac{\pi}{2} - \theta\right)(r\cos\theta - 1) = -\frac{\pi}{2}r^2 + (\pi - \theta)\cos\theta r - \left(\frac{\pi}{2} - \theta\right) =: f(r) \ ,$$

where the inequality holds since $\sin(2\psi) + 2\psi$ is increasing. The discriminant of f is

$$\Delta(\theta) = (\pi - \theta)^2\cos^2\theta - \pi(\pi - 2\theta) \leqslant \frac{1}{\pi^2}\theta^2(\pi - 2\theta)(2\theta - 3\pi) \ ,$$

where the first inequality holds since $\cos^2\theta \leqslant 1 - 4\theta^2/\pi^2$ on $[0, \pi/2]$. According to $\theta \in [0, \pi/2]$, one knows $\Delta(\theta) \leqslant 0$, which indicates $f(r) \leqslant 0$, and thus, $\widetilde{R} \leqslant 0$ when $\cos\theta - r \leqslant 0$. Combining the cases above, we obtain $\widetilde{R} \leqslant 0$, which, together with Eq. (23), implies $R_1 \leqslant 2^{-1}[\sin(2\psi) + 2\psi](1 - 2r\cos\theta + r^2) = 2^{-1}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2$.

Combining the cases above completes the proof. □

Lemma 18 *Let* $R_2 = \left[\frac{1}{2} - \frac{1}{2}\cos(2\psi) - \frac{w_1}{w_1^2 + w_2^2}\right]|w_2|$. *If* $(\boldsymbol{w}, \psi) \in D_2$, *then*

$$R_2 \leqslant \frac{1}{\pi}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ .$$

Proof. Let $r = \sqrt{w_1^2 + w_2^2}$ be the norm of $\boldsymbol{w}$. By the definition of θ, one has $w_1 = r\cos\theta$ and $|w_2| = r\sin\theta$. Thus, we can rewrite R_2 as $R_2 = \left[2^{-1}r(1 - \cos(2\psi)) - \cos\theta\right]\sin\theta$. We provide the upper bound for R_2 by discussion.

1. Suppose $\frac{r}{2}[1 - \cos(2\psi)] - \cos\theta \leqslant 0$. From $\theta \in [0, \pi/2]$, we have $R_2 \leqslant 0$.

2. Suppose $\frac{r}{2}[1 - \cos(2\psi)] - \cos\theta > 0$. It is observed that $r < 2\cos\theta$ since $\|\boldsymbol{w} - \boldsymbol{v}\|^2 \leqslant r_0^2 < 1$ holds from the definition of D_2. Thus, the supposition indicates $\cos\theta < \frac{r}{2}[1 - \cos(2\psi)] < [1 - \cos(2\psi)]\cos\theta$, which, together with $\theta \in [0, \pi/2]$, implies $\psi \geqslant \pi/4$. It is observed that

$$f(r) = \frac{1}{2}(1 - 2r\cos\theta + r^2) - (r - \cos\theta)\sin\theta = \frac{1}{2}(r - \cos\theta - \sin\theta)^2 \geqslant 0 \ ,$$

which indicates

$$\frac{1}{\pi}[\sin(2\psi) + 2\psi](1 - 2r\cos\theta + r^2) \geqslant \frac{1}{2}(1 - 2r\cos\theta + r^2) \geqslant (r - \cos\theta)\sin\theta \geqslant R_2 \ ,$$

where the first inequality holds according to $\psi \geqslant \pi/4$, and the third inequality holds because of $\cos(2\psi) \geqslant -1$.

Combining the cases above, we obtain

$$R_2 \leqslant \frac{1}{\pi}[\sin(2\psi) + 2\psi](1 - 2r\cos\theta + r^2) = \frac{1}{\pi}[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ ,$$

which completes the proof. □

Lemma 19 *Define T_1 as in Eq. (20). If $(\boldsymbol{w}, \psi) \in D_2$. then $T_1 \leqslant 7\pi[\sin(2\psi)+2\psi]\|\boldsymbol{w}-\boldsymbol{v}\|^2$.*

Proof. It is observed that $T_1 = \big[[\sin(2\psi)+2\psi](w_1-1)+T_{11}+T_{12}\big]^2$ with

$$T_{11} = \frac{1}{2}\sin(2\psi) + \left(\psi+\theta-\frac{\pi}{2}\right) \quad \text{and} \quad T_{12} = -\frac{w_1|w_2|}{w_1^2+w_2^2} \ . \tag{24}$$

Firstly, we define $r_0 \in (0,1)$ and calculate an upper bound for T_{11} by discussion.
1. Suppose $|w_1-1|+|w_2| \geqslant r_0$. Then one knows from $\theta \leqslant \frac{\pi}{2}$ that

$$|T_{11}| \leqslant \frac{1}{2}\sin(2\psi)+\psi \leqslant \frac{1}{2r_0}[\sin(2\psi)+2\psi][|w_1-1|+|w_2|] \ .$$

2. Suppose $|w_1-1|+|w_2| \leqslant r_0$. Then it is observed that $w_1 \geqslant 1-r_0+|w_2| \geqslant 0$. Thus,

$$r = \sqrt{w_1^2+w_2^2} \geqslant \sqrt{(1-r_0)^2+2|w_2|(|w_2|+1-r_0)} \geqslant 1-r_0 \ ,$$

where the second inequality holds from $r_0 \leqslant 1$. Then we can bound $|w_2|$ from below as

$$|w_2| = r\sin\theta \geqslant (1-r_0)\sin\theta \geqslant \frac{1-r_0}{2}\theta \ , \tag{25}$$

where the second inequality holds since $\theta \leqslant 2\sin\theta$ for all $\theta \in [0,\pi/2]$. Meanwhile, we bound θ from above as

$$\theta \leqslant \tan\theta = \frac{|w_2|}{w_1} \leqslant \left(\frac{1-r_0}{|w_2|}+1\right)^{-1} \leqslant \left(\frac{1-r_0}{r_0}+1\right)^{-1} = r_0 \ , \tag{26}$$

where the second inequality holds from $w_1 \geqslant 1-r_0+|w_2|$, and the third inequality holds based on $|w_2| \leqslant r_0$. Then we obtain an upper bound of T_{11} as follows

$$|T_{11}| \leqslant \theta \leqslant \frac{2|w_2|}{1-r_0} \leqslant \frac{4\psi|w_2|}{(1-r_0)(\pi-2r_0)} \leqslant \frac{2}{(1-r_0)(\pi-2r_0)}[\sin(2\psi)+2\psi][|w_1-1|+|w_2|] \ ,$$

where the first inequality holds from $\psi \leqslant \frac{\pi}{2}$, the second inequality holds from Eq. (25), and the third inequality holds based on $\psi \geqslant \frac{\pi}{2}-\theta$ and Eq. (26).

Combining the cases above, we have proven

$$|T_{11}| \leqslant \max\left\{\frac{1}{2r_0}, \frac{2}{(1-r_0)(\pi-2r_0)}\right\}[\sin(2\psi)+2\psi][|w_1-1|+|w_2|] \ .$$

Choosing $r_0 = \frac{1}{4}\left[\pi+6-\sqrt{\pi^2+4\pi+36}\right]$, we obtain an upper bound of T_{11} as follows

$$|T_{11}| \leqslant \frac{3}{2}[\sin(2\psi)+2\psi][|w_1-1|+|w_2|] \ . \tag{27}$$

Secondly, we provide an upper bound for T_{12}. We claim and prove by discussion that

$$|w_2| \leqslant 2\sqrt{w_1^2+w_2^2}(|w_1-1|+|w_2|) \ . \tag{28}$$

1. Suppose $w_1 \leqslant 1/2$. Then it is observed that $|w_1 - 1| \geqslant 1/2$, which implies
$$|w_2| \leqslant \sqrt{w_1^2 + w_2^2} \leqslant \sqrt{w_1^2 + w_2^2} \cdot 2|w_1 - 1| \leqslant 2\sqrt{w_1^2 + w_2^2}(|w_1 - 1| + |w_2|) \ .$$

2. Suppose $w_1 \geqslant 1/2$. Then one has $\sqrt{w_1^2 + w_2^2} \geqslant 1/2$, which indicates
$$|w_2| \leqslant |w_1 - 1| + |w_2| \leqslant 2\sqrt{w_1^2 + w_2^2}(|w_1 - 1| + |w_2|) \ .$$

From the definition of D_2, one has $\frac{\pi}{2} \geqslant \psi \geqslant \frac{\pi}{2} - \theta \geqslant 0$, which indicates
$$\frac{1}{2}[\sin(2\psi) + 2\psi] \geqslant \psi \geqslant \sin\psi \geqslant \sin\left(\frac{\pi}{2} - \theta\right) = \cos\theta = \frac{w_1}{\sqrt{w_1^2 + w_2^2}} \ . \tag{29}$$

Then we obtain an upper bound of $|T_{12}|$ as
$$|T_{12}| \leqslant \frac{2w_1}{\sqrt{w_1^2 + w_2^2}}(|w_1 - 1| + |w_2|) \leqslant [\sin(2\psi) + 2\psi](|w_1 - 1| + |w_2|) \ , \tag{30}$$

where the first inequality holds according to Eq. (28), and the second inequality holds based on Eq. (29). Finally, combining Eqs. (27) and (30), we conclude
$$T_1 \leqslant \Big[\big|[\sin(2\psi) + 2\psi](w_1 - 1)\big| + \max\{|T_{11}|, |T_{12}|\}\Big]^2 \leqslant 7\pi[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ ,$$

where the first inequality holds based on $T_{11} \geqslant 0$ and $T_{12} \leqslant 0$, and the second inequality holds from $\sin(2\psi) + 2\psi \leqslant \pi$ for any $\psi \in [0, \pi/2]$. Thus, we have completed the proof. □

Lemma 20 *Define T_2 as in Eq. (20). If $(\boldsymbol{w}, \psi) \in D_2$, then $T_2 \leqslant 7\pi[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2$.*

Proof. From $\cos\theta = w_1/\sqrt{w_1^2 + w_2^2}$, one has $\cos(\pi - 2\theta) = 1 - 2\cos^2\theta = 1 - 2w_1^2/(w_1^2 + w_2^2)$. Thus, we have
$$\left|\left[\frac{1}{2}\cos(2\psi) - \frac{1}{2} + \frac{w_1^2}{w_1^2 + w_2^2}\right]\operatorname{sgn}(w_2)\right| = \frac{1}{2}|\cos(2\psi) - \cos(\pi - 2\theta)| \leqslant \psi + \theta - \frac{\pi}{2} \leqslant T_{11} \ ,$$

where the first inequality holds because of $|\cos a - \cos b| \leqslant |a - b|$, and the second inequality holds based on the definition of T_{11} in Eq. (24) and $\sin(2\psi) \geqslant 0$. Recalling the upper bound of T_{11} in Eq. (27), we obtain
$$\begin{aligned} T_2 &\leqslant \left(\left|\left[\frac{1}{2}\cos(2\psi) - \frac{1}{2} + \frac{w_1^2}{w_1^2 + w_2^2}\right]\operatorname{sgn}(w_2)\right| + |[\sin(2\psi) + 2\psi]w_2|\right)^2 \\ &\leqslant 7\pi[\sin(2\psi) + 2\psi]\|\boldsymbol{w} - \boldsymbol{v}\|^2 \ , \end{aligned}$$

which completes the proof. □

Appendix C. Proof of Theorem 2

In the main part of this section, we present the closed form of the loss, definition and properties of the ideal region, and the detailed proof of Theorem 2. Subsection C.1 provides the optimization behaviors. Subsection C.2 gives some convergence rate lemmas.

According to Lemma 10, the expected square loss L_{cc} can be calculated by
$$L_{\text{cc}}(\boldsymbol{w}, \psi_w) = \frac{1}{2}B(\boldsymbol{w}, \boldsymbol{w}, \psi_w, \psi_w) - B(\boldsymbol{w}, \boldsymbol{v}, \psi_w, \psi_v) + \frac{1}{2}B(\boldsymbol{v}, \boldsymbol{v}, \psi_v, \psi_v) \ . \tag{31}$$

For $R \in (0,1)$ and $0 \leqslant \psi_l \leqslant \psi_u \leqslant \pi/2$, define

$$D_1 = \{(\boldsymbol{w}, \psi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \psi_w \in [\psi_l, \psi_u], \theta_{\boldsymbol{w},\boldsymbol{v}} \in [0, |\psi_w - \psi_v|]\} \ ,$$
$$D_2 = \{(\boldsymbol{w}, \psi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \psi_w \in [\psi_l, \psi_u], \theta_{\boldsymbol{w},\boldsymbol{v}} \in (|\psi_w - \psi_v|, \psi_w + \psi_v)\} \ .$$

Let $D = D_1 \cup D_2 = \{(\boldsymbol{w}, \psi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \psi_w \in [\psi_l, \psi_u], \theta_{\boldsymbol{w},\boldsymbol{v}} \in [0, \psi_w + \psi_v]\}$ be the ideal region. By spherical symmetry, we assume $\boldsymbol{v} = (1,0)$ without loss of generality in the rest proof. For conciseness, define $s_w = \sin(2\psi_w) + 2\psi_w$ and $s_v = \sin(2\psi_v) + 2\psi_v$. The following lemma discusses the properties of the ideal region, concerning the closeness of the region under gradient descent and the probability that an initialization falls into this region.

Lemma 21 *Let $\psi_v \in [7\pi/20, 2\pi/5]$. If we choose the parameters as*

$$R = \frac{1}{25} \ , \quad \psi_l = \psi_v - \frac{109}{100}R \ , \quad \psi_u = \psi_v + \frac{109}{100}R \ , \quad \text{and} \quad 0 < \eta \leqslant \frac{1}{120}R \ ,$$

then conditions in Lemmas 22-27 are satisfied. If $\boldsymbol{w}_0 \sim \mathcal{N}(0, \mathbf{I}_2)$ and $\psi_{w,0} \sim \mathcal{U}(0, \pi/2)$, then

$$\Pr\left[(\boldsymbol{w}_0, \psi_{w,0}) \in D\right] \geqslant 10^{-5} \ .$$

Proof. We first prove that all conditions in the lemmas are satisfied.

- Lemma 22. The first condition holds from $\eta \leqslant R/120 < 2$. According to $\psi_u > \psi_v > \pi/4$, we have $\psi_v \sin(2\psi_u) \leqslant \psi_u \sin(2\psi_v)$, which implies the second condition

$$s_v \geqslant \frac{\psi_v s_u}{\psi_u} = \frac{\psi_v s_u}{\psi_v + 109R/100} \geqslant \frac{7\pi s_u/20}{7\pi/20 + 109R/100} \geqslant (1-R)s_u \geqslant (1-R)s_w \ ,$$

  where the fourth inequality holds since s_w is monotonic.

- Lemma 23. The first condition $\eta < 2$ has been verified. The second condition holds according to $\psi_l/20 \geqslant (7\pi/20 - 109R/100)/20 \geqslant R$. The third condition holds because of $\max\{\psi_u - \psi_v, \psi_v - \psi_l\} = 109R/100 \leqslant 5R\psi_l/3$.
- Lemma 24. The only condition $\eta < 2$ has been verified.
- Lemma 25. The first condition holds by $R = 1/25 \leqslant 1/2$. The second one holds based on $\cos^2 \psi_v \geqslant \cos^2(2\pi/5) \geqslant 1/25$. The third one holds from $\eta \leqslant R/120 \leqslant 3R/2$.
- Lemma 26. The first condition $R \leqslant 1/2$ has been verified. The second and third ones hold from $3^{-1}\pi \min\{\psi_u - \psi_v, \psi_v - \psi_l\} \geqslant \frac{R}{120} \geqslant \eta$.
- Lemma 27. The first condition $R \leqslant 1/2$ has been verified. The second one holds from

$$\arcsin R + 9\eta \leqslant \frac{101R}{100} + \frac{3R}{40} \leqslant \frac{109R}{100} = \psi_u - \psi_v \ .$$

We then prove the second conclusion. Let $p_0 = \Pr[(\boldsymbol{w}_0, \psi_{w,0}) \in D]$. Then we have

$$\begin{aligned} p_0 &= \Pr[\psi_l \leqslant \psi_{w,0} \leqslant \psi_u] \cdot \Pr[1 - R \leqslant w_1 \leqslant 1 + R] \cdot \Pr[-R \leqslant w_2 \leqslant R] \\ &= \frac{109R}{50} \cdot \frac{1}{2}[\operatorname{erf}(1+R) - \operatorname{erf}(1-R)] \cdot \operatorname{erf}(R) \\ &\geqslant 10^{-5} \ , \end{aligned}$$

where $\operatorname{erf}(x)$ denotes the error function. Thus, we have completed the proof. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. Let R, ψ_l, and ψ_u be the same as those in Lemma 21. Suppose that $(\boldsymbol{w}_0, \psi_{w,0}) \in D$. Then Lemma 21 implies $(\boldsymbol{w}_t, \psi_{w,t}) \in D$ for any $t \in \mathbb{N}$. The proof of convergence is divided into several stages.

Step 1: w_2 converges to 0. In stage I, we consider the convergence of $w_{t,2}$ when $(\boldsymbol{w}_0, \psi_{w,0}) \in D$. From Lemmas 24 and 25, the optimization behaviors of w_2 is the combination of minimizing a contraction mapping or an almost absolute function. Thus, Lemma 28 with $r_1 = r_2 = R$, $c_3 = s_w/(2\pi)$, $g_l = (\cos^2\psi_v - \sqrt{2}R)/(2\pi)$, and $g_u = 2/3$ implies

$$|w_{t,2}| \leqslant \frac{c_2^2(\cos^2\psi_v - \sqrt{2}R)}{4\pi c_1 t} \leqslant \frac{c_2^2}{4\pi c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{32}$$

Step 2: ψ_w converges to ψ_v. In stage II, we prove the convergence of $\psi_{w,t}$ when $(\boldsymbol{w}_0, \psi_{w,0}) \in D$. From Lemmas 26 and 27, the convergence of ψ_w is limited by that of w_2, i.e., ψ_w tends to the global minimum with constant-order gradient when the error of ψ_w is larger than that of w_2, while becomes far away from the global minimum otherwise. Then Lemma 29 with $r_1 = r_2 = 109R/100$, $a = c_2^2(\cos^2\psi_v - \sqrt{2}R)/[4\pi c_1(1-R)]$, $g_l = \cos^2\psi_u/(4\pi)$, and $g_u = 9$ indicates

$$|\psi_{w,t} - \psi_v| \leqslant \left[\frac{c_2^2(\cos^2\psi_v - \sqrt{2}R)}{2\pi c_1(1-R)} + 9c_2\right]\frac{1}{t} \leqslant \frac{10c_2^2}{c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{33}$$

Step 3: w_1 converges to 1. In stage III, we investigate the convergence of $w_{t,1}$ when $(\boldsymbol{w}_0, \psi_{w,0}) \in D$. From Lemmas 22 and 23, the gradient points to the global minimum with a remainder controlled by the error of w_1 and ψ_w. Then Lemma 30 with $d_l = 1/4$, $d_u = 1/2$, and $e = 20c_2^2/(\pi c_1)$ leads to

$$|w_{t,1} - 1| \leqslant \frac{20c_2^3}{\pi c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{34}$$

Step 3: the expected loss converges to 0. We now estimate the convergence of the loss when $(\boldsymbol{w}_0, \psi_{w,0}) \in D$. For any $(\boldsymbol{w}, \psi_w) \in D$, define non-negative quantities $\Delta_{\boldsymbol{w}} = \|\boldsymbol{w} - \boldsymbol{v}\|$ and $\Delta_\psi = |\psi_w - \psi_v|$. We provide an upper bound for L_{cc} by discussion.

1. Suppose $(\boldsymbol{w}, \psi_w) \in D_1$. Then we have

$$\begin{aligned} 4\pi L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) &= \|\boldsymbol{w}\|^2 s_w - 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|\cos\theta_{\boldsymbol{w},\boldsymbol{v}} s_m + \|\boldsymbol{v}\|^2 s_v \\ &\leqslant \|\boldsymbol{w}\|^2(s_v + s_\Delta) - 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|(1 - \Delta_{\boldsymbol{w}}^2)(s_v - s_\Delta) + \|\boldsymbol{v}\|^2 s_v \\ &\leqslant 4(\|\boldsymbol{w}\|^2 + 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|)\Delta_\psi + (s_v + 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|)\Delta_{\boldsymbol{w}}^2 \\ &\leqslant 32\Delta_\psi + 8\Delta_{\boldsymbol{w}}^2 \ , \end{aligned}$$

where the first inequality holds from $s_w \leqslant s_v + s_\Delta$, $\cos\theta_{\boldsymbol{w},\boldsymbol{v}} \geqslant \sqrt{1 - \Delta_{\boldsymbol{w}}^2} \geqslant 1 - \Delta_{\boldsymbol{w}}^2$, and $s_m \geqslant s_v - s_\Delta$ with $s_\Delta = 2\Delta_\psi + \sin(2\Delta_\psi)$, the second inequality holds since $|\|\boldsymbol{w}\| - \|\boldsymbol{v}\|| \leqslant \Delta_{\boldsymbol{w}}^2$ and $s_\Delta \leqslant 4\Delta_\psi$, and the third inequality holds based on $\|\boldsymbol{w}\| \leqslant 2$ and $s_v \leqslant \pi$.

2. Suppose $(\boldsymbol{w}, \psi_w) \in D_2$. Let $\theta = \theta_{\boldsymbol{w},\boldsymbol{v}}$. Then one knows

$$\begin{aligned} 4\pi L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) = {} & s_v(\|\boldsymbol{w}\| - \|\boldsymbol{v}\|)^2 + (\|\boldsymbol{w}\|^2 - \|\boldsymbol{w}\|\|\boldsymbol{v}\|\cos\theta)(s_w - s_v) \\ & + \|\boldsymbol{w}\|\|\boldsymbol{v}\|\theta\cos\theta + 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|s_v(1 - \cos\theta) \ . \end{aligned}$$

Then according to $|\|\boldsymbol{w}\|-\|\boldsymbol{v}\|| \leqslant \Delta_{\boldsymbol{w}}$, $s_w - s_v \leqslant 4\Delta_\psi$, $\theta \leqslant \arcsin \Delta_{\boldsymbol{w}} \leqslant 2\Delta_{\boldsymbol{w}}$, and $\cos\theta \geqslant 1-\Delta_{\boldsymbol{w}}^2$, we have

$$\begin{aligned}4\pi L_{\rm cc} &\leqslant 4\big|\|\boldsymbol{w}\|^2 - \|\boldsymbol{w}\|\|\boldsymbol{v}\|\cos\theta\big|\Delta_\psi + 2\|\boldsymbol{w}\|\|\boldsymbol{v}\|\cos\theta\Delta_{\boldsymbol{w}} + (1+2\|\boldsymbol{w}\|\|\boldsymbol{v}\|)s_v\Delta_{\boldsymbol{w}}^2\\ &\leqslant 16\Delta_\psi + 5\Delta_{\boldsymbol{w}}\ ,\end{aligned}$$

where the second inequality hodls based on $\|\boldsymbol{w}\| \leqslant 2$, $s_v \leqslant \pi$, and $\Delta_{\boldsymbol{w}} \leqslant \sqrt{2}R = \sqrt{2}/25$. Combining the cases above, one knows from $\Delta_{\boldsymbol{w}} \leqslant 5/8$ that

$$L_{\rm cc}(\boldsymbol{w},\psi_w) \leqslant 32\Delta_\psi + 5\Delta_{\boldsymbol{w}} \quad \text{for} \quad (\boldsymbol{w},\psi_w)\in D\ . \tag{35}$$

Then based on $(\boldsymbol{w}_t,\psi_{w,t})\in D$ and Eqs. (32)-(34), we obtain from $c_2 \geqslant 1$ that

$$L_{\rm cc}(\boldsymbol{w}_t,\psi_{w,t}) \leqslant \frac{320c_2^2}{c_1 t} + \frac{5c_2^2}{4\pi c_1 t} + \frac{100c_2^3}{\pi c_1 t} \leqslant \frac{400c_2^3}{c_1 t}\ ,$$

which holds with probability at least 10^{-5} from Lemma 21 and completes the proof. □

C.1 Optimization behaviors

The next two lemmas consider the gradient with respect to w_1 in D_1 and D_2, respectively.

Lemma 22 *Let $w_1 = w_1 - \eta\nabla_{w_1}L_{\rm cc}(\boldsymbol{w},\psi_w)$ with $(\boldsymbol{w},\psi_w)\in D_1$. If $\eta\in(0,2)$ and $(1-R)s_w \leqslant s_v$, then we have*

$$\nabla_{w_1}L_{\rm cc}(\boldsymbol{w},\psi_w) = \frac{s_w}{2\pi}(w_1-1) + \frac{1}{2\pi}[s_w - \min\{s_w,s_v\}] \quad \textit{and} \quad |w_1'-1| \leqslant R\ .$$

Proof. For any $(\boldsymbol{w},\psi_w)\in D_1$, one has

$$\nabla_{w_1}L_{\rm cc}(\boldsymbol{w},\psi_w) = \frac{s_w}{2\pi}\left[w_1 - \min\{s_w,s_v\}\right] = \frac{s_w}{2\pi}(w_1-1) + r\ , \tag{36}$$

where r denotes a remainder defined by $r = \frac{1}{2\pi}[s_w - \min\{s_w,s_v\}]$. Then Eq. (36) implies

$$|w_1'-1| \leqslant \left|1-\frac{\eta s_w}{2\pi}\right||w_1-1| + |\eta r| \leqslant \left(1-\frac{\eta s_w}{2\pi}\right)R + \frac{\eta}{2\pi}[s_w - \min\{s_w,s_v\}]\ , \tag{37}$$

where the first inequality holds from the triangle inequality, and the second inequality holds based on $1-\eta s_w/(2\pi) \geqslant 0$ and $|w_1 - 1| \leqslant R$. We proceed to complete the proof by discussion.

- Suppose that $\min\{s_w,s_v\} = s_w$. Then Eq. (37) implies $|w_1'-1| \leqslant [1-\eta s_w/(2\pi)]R \leqslant R$, where the second inequality holds from $\eta > 0$ and $s_w \geqslant 0$.
- Suppose that $\min\{s_w,s_v\} = s_v$. Then one knows from Eq. (37) that

$$|w_1'-1| \leqslant \left(1-\frac{\eta s_w}{2\pi}\right)R + \frac{\eta(s_w-s_v)}{2\pi} \leqslant R\ ,$$

  where the second inequality holds because of $(1-R)s_w \leqslant s_v$.

Combining the cases above completes the proof. □

Lemma 23 *Let $w_1 = w_1 - \eta\nabla_{w_1}L_{\rm cc}(\boldsymbol{w},\psi_w)$ with $(\boldsymbol{w},\psi_w)\in D_2$. If $\eta\in(0,2)$, $R \leqslant \psi_l/20$ and $\max\{\psi_u-\psi_v,\psi_v-\psi_l\} \leqslant 5R\psi_l/3$, then we have*

$$\nabla_{w_1}L_{\rm cc}(\boldsymbol{w},\psi_w) = \frac{s_w - \theta_{\boldsymbol{w},\boldsymbol{v}}}{2\pi}(w_1-1) + \frac{1}{4\pi}[(s_w-s_v) + 2(\theta_{\boldsymbol{w},\boldsymbol{v}} - \sin\theta_{\boldsymbol{w},\boldsymbol{v}})] \quad \textit{and} \quad |w_1'-1| \leqslant R\ .$$

Proof. For any $(\boldsymbol{w},\psi_w)\in D_2$, the gradient of L_{cc} with respect to w_1 can be calculated by

$$\nabla_{w_1}L_{\text{cc}}=\frac{s_w-\theta_{\boldsymbol{w},\boldsymbol{v}}}{2\pi}(w_1-1)+\frac{1}{4\pi}[(s_w-s_v)+2(\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin\theta_{\boldsymbol{w},\boldsymbol{v}})]=\frac{s_w-\theta_{\boldsymbol{w},\boldsymbol{v}}}{2\pi}(w_1-1)+r\ ,$$

where r is a remainder defined by $r=[(s_w-s_v)+2(\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin\theta_{\boldsymbol{w},\boldsymbol{v}})]/(4\pi)$. Then we have

$$|w_1'-1|\leqslant\left|1-\frac{\eta(s_w-\theta_{\boldsymbol{w},\boldsymbol{v}})}{2\pi}\right||w_1-1|+|\eta r|\leqslant R+\eta\left[|r|-\frac{R(s_w-\theta_{\boldsymbol{w},\boldsymbol{v}})}{2\pi}\right]\ , \tag{38}$$

where the first inequality holds from the triangle inequality, and the second inequality holds based on $\eta(s_w-\theta_{\boldsymbol{w},\boldsymbol{v}})\leqslant\eta s_w\leqslant 2\pi$ and $|w_1-1|\leqslant R$. It is observed that

$$s_w-\theta_{\boldsymbol{w},\boldsymbol{v}}\geqslant\frac{7}{2}\psi_l-\theta_{\boldsymbol{w},\boldsymbol{v}}\geqslant\frac{7}{2}\psi_l-2R\ , \tag{39}$$

where the first inequality holds based on $s_w\geqslant 2\psi_l+\sin(2\psi_l)$ and $\sin\psi_l\geqslant 3\psi_l/4$ for $\psi_l\leqslant\pi/4$, and the second inequality holds from $\theta_{\boldsymbol{w},\boldsymbol{v}}\leqslant\arcsin R\leqslant 2R$. Meanwhile, one has

$$|r|\leqslant\frac{1}{4\pi}|s_w-s_v|+\frac{1}{2\pi}|\theta_{\boldsymbol{w},\boldsymbol{v}}-\sin\theta_{\boldsymbol{w},\boldsymbol{v}}|\leqslant\frac{\max\{\psi_u-\psi_v,\psi_v-\psi_l\}}{\pi}+\frac{2R^3}{3\pi}\ , \tag{40}$$

where the first inequality holds from the triangle inequality, and the second inequality holds according to the 4-Lipschitzness of $2\theta+\sin(2\theta)$, $\theta-\sin\theta\leqslant\theta^3/6$ for any $\theta\geqslant 0$, and $\theta_{\boldsymbol{w},\boldsymbol{v}}\leqslant 2R$. Substituting Eqs. (39) and (40) into Eq. (38), we obtain

$$|w_1'-1|\leqslant R+\frac{\eta}{12\pi}\left[12\max\{\psi_u-\psi_v,\psi_v-\psi_l\}+8R^3+12R^2-21R\psi_l\right]\leqslant R\ ,$$

where the second inequality holds from $\max\{\psi_u-\psi_v,\psi_v-\psi_l\}\leqslant 5R\psi_l/3$ and $R\leqslant\psi_l/20\leqslant 1$. Thus, we have completed the proof. □

The next two lemmas focus on the gradient with respect to w_2 in D_1 and D_2, respectively.

Lemma 24 *Let $w_2'=w_2-\eta\nabla_{w_2}L_{\text{cc}}(\boldsymbol{w},\psi_w)$ with $(\boldsymbol{w},\psi_w)\in D_1$. If $\eta\in(0,2)$, then we have*

$$|w_2'|\leqslant\left(1-\frac{\eta s_w}{2\pi}\right)|w_2|\quad\textit{and}\quad|w_2'|\leqslant R\ .$$

Proof. For any $(\boldsymbol{w},\psi_w)\in D_1$, one has $\nabla_{w_2}L_{\text{cc}}(\boldsymbol{w},\psi_w)=\frac{s_w w_2}{2\pi}$. Thus, we have

$$w_2'=\left(1-\frac{\eta s_w}{2\pi}\right)w_2\ . \tag{41}$$

According to $s_w\in[0,\pi]$ and $\eta\in(0,2)$, the coefficient $1-\eta s_w/(2\pi)$ is positive and smaller than 1. Based on $(\boldsymbol{w},\psi_w)\in D_1$, one knows $|w_2|\leqslant R$. Then Eq. (41) implies

$$|w_2'|=\left(1-\frac{\eta s_w}{2\pi}\right)|w_2|\leqslant R\ ,$$

which completes the proof. □

Lemma 25 *Let $w_2'=w_2-\eta\nabla_{w_2}L_{\text{cc}}(\boldsymbol{w},\psi_w)$ with $(\boldsymbol{w},\psi_w)\in D_2$. If $R\leqslant 1/2$, $\sqrt{2}R\leqslant\cos^2\psi_v$, and $\eta\leqslant 3R/2$, then we have*

$$\frac{\cos^2\psi_v-\sqrt{2}R}{2\pi}\leqslant\nabla_{w_2}L_{\text{cc}}(\boldsymbol{w},\psi_w)\text{sgn}(w_2)\leqslant\frac{2}{3}\quad\textit{and}\quad|w_2'|\leqslant R\ .$$

Proof. For any $(\boldsymbol{w}, \psi_w) \in D_2$, the gradient of L_{cc} with respect to w_2 satisfies

$$\nabla_{w_2} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) = \frac{1}{2\pi} s_w w_2 + \frac{1}{4\pi}\left[\cos(2\psi_w) + \cos(2\psi_v) + 2w_1^2 \|\boldsymbol{w}\|^{-1}\right] \mathrm{sgn}(w_2)\ . \tag{42}$$

Since $(\boldsymbol{w}, \psi_w) \in D_2$, one knows that $|w_1 - 1| \leqslant R$ and $|w_2| \leqslant R$. Thus, we have

$$2(1-\sqrt{2}R) \leqslant 2(1-R)^2[(1-R)^2 + R^2]^{-1/2} \leqslant 2w_1^2\|\boldsymbol{w}\|^{-1} \leqslant 2(1+R)\ ,$$

where the first inequality holds because of $R \in [0, 1/2]$. Then we have

$$\cos(2\psi_w) + \cos(2\psi_v) + 2w_1^2\|\boldsymbol{w}\|^{-1} \leqslant 1 + \cos(2\psi_v) + 2(1+R) \leqslant 5\ , \tag{43}$$

where the second inequality holds based on $R \leqslant 1/2$. Meanwhile, one has

$$\cos(2\psi_w) + \cos(2\psi_v) + 2w_1^2\|\boldsymbol{w}\|^{-1} \geqslant -1 + \cos(2\psi_v) + 2(1-\sqrt{2}R) = 2(\cos^2\psi_v - \sqrt{2}R)\ . \tag{44}$$

It is observed that $0 \leqslant s_w|w_2| \leqslant \frac{\pi}{2}$ since $s_w \in [0, \pi]$ and $|w_2| \leqslant R \leqslant \frac{1}{2}$. Then substituting Eqs. (43) and (44) into Eq. (42), we obtain

$$\frac{\cos^2\psi_v - \sqrt{2}R}{2\pi} \leqslant \nabla_{w_2} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w)\mathrm{sgn}(w_2) \leqslant \frac{1}{4} + \frac{5}{4\pi} \leqslant \frac{2}{3}\ .$$

Thus, one knows from Eq. (42) that

$$|w_2'| = \big||w_2| - \eta\nabla_{w_2} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w)\mathrm{sgn}(w_2)\big| \leqslant \max\{|w_2|, \eta\nabla_{w_2} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w)\mathrm{sgn}(w_2)\} \leqslant R\ ,$$

where the first inequality holds from $|a-b| \leqslant \max\{a, b\}$ for $a, b \geqslant 0$, and the second inequality holds based on $|w_2| \leqslant R$ and $\eta \leqslant 3R/2$. Thus, we have completed the proof. □

The next two lemmas investigate gradients with respect to ψ_w in D_1 and D_2, respectively.

Lemma 26 *Let $\psi_w' = \psi_w - \eta\nabla_{\psi_w} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w)$ with $(\boldsymbol{w}, \psi_w) \in D_1$. If $R \leqslant 1/2$, $\eta \leqslant \pi(\psi_u - \psi_v)/3$, and $\eta \leqslant \pi(\psi_v - \psi_l)/3$, then we have*

$$\frac{\cos^2\psi_u}{4\pi} \leqslant \mathrm{sgn}(\psi_w - \psi_v)\nabla_{\psi_w} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) \leqslant \frac{3}{\pi} \quad \textit{and} \quad \psi_w' \in [\psi_l, \psi_u]\ .$$

Proof. For any $(\boldsymbol{w}, \psi_w) \in D_1$, the gradient of L_{cc} with respect to ψ_w satisfies

$$\nabla_{\psi_w} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) = \begin{cases} -\frac{1}{2\pi}[1 + \cos(2\psi_w)][1 - \|\boldsymbol{w} - \boldsymbol{v}\|^2]\ , & \psi_w < \psi_v\ , \\ \frac{1}{2\pi}[1 + \cos(2\psi_w)]\|\boldsymbol{w}\|^2\ , & \psi_w > \psi_v\ , \end{cases}$$

where the gradient at $\psi_w = \psi_v$ can be any subgradient. For any $(\boldsymbol{w}, \psi_w) \in D_2$, we have $\psi_w \in [\psi_l, \psi_u]$, which indicates $2\cos^2\psi_u \leqslant 1 + \cos(2\psi_w) \leqslant 2$. Meanwhile, all points in D_2 satisfies $1 - 2R^2 \leqslant 1 - \|\boldsymbol{w} - \boldsymbol{v}\|^2 \leqslant 1$ and $(1-R)^2 \leqslant \|\boldsymbol{w}\|^2 \leqslant (1+R)^2 + R^2$. Thus, the gradient of L_{cc} with respect to ψ_w can be bounded by

$$(4\pi)^{-1}\cos^2\psi_u \leqslant \mathrm{sgn}(\psi_w - \psi_v)\nabla_{\psi_w} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) \leqslant 3\pi^{-1}\ ,$$

where the first and second inequalities holds based on $R \leqslant 1/2$. Then ψ_w' satisfies

$$\psi_w' = \psi_w - \eta\nabla_{\psi_w} L_{\mathrm{cc}}(\boldsymbol{w}, \psi_w) \leqslant \max\left\{\psi_w, \psi_v + \frac{3\eta}{\pi}\right\} \leqslant \psi_u\ ,$$

where the first inequality holds from discussing the relation between ψ_w and ψ_v, and the second inequality holds based on $\psi_w \leqslant \psi_u$ and $\eta \leqslant \pi(\psi_u - \psi_v)/3$. Meanwhile, one has

$$\psi_w' = \psi_w - \eta \nabla_{\psi_w} L_{\rm cc}(\boldsymbol{w}, \psi_w) \geqslant \min\left\{\psi_w, \psi_v - \frac{3\eta}{\pi}\right\} \geqslant \psi_l \ ,$$

where the first inequality holds from discussing the relation between ψ_w and ψ_v, and the second inequality holds by $\psi_w \geqslant \psi_l$ and $\eta \leqslant \pi(\psi_v - \psi_l)/3$. Thus, the proof is completed. □

Lemma 27 *Let $\psi_w' = \psi_w - \eta \nabla_{\psi_w} L_{\rm cc}(\boldsymbol{w}, \psi_w)$ with $(\boldsymbol{w}, \psi_w) \in D_2$. If $R \leqslant 1/2$ and $\arcsin R + 9\eta \leqslant \psi_u - \psi_v$, then we have*

$$-9 \leqslant -2\left(\frac{\pi}{2} - \psi_w\right)^2 - 2\left(\frac{\pi}{2} - \psi_w\right)|w_2| \leqslant \nabla_{\psi_w} L_{\rm cc} \leqslant -\frac{1}{4}\left(\frac{\pi}{2} - \psi_w\right)^2 \quad \text{and} \quad \psi_w' \in [\psi_l, \psi_u] \ .$$

Proof. For any $(\boldsymbol{w}, \psi_w) \in D_1$, the gradient of $L_{\rm cc}$ with respect to ψ_w satisfies

$$\nabla_{\psi_w} L_{\rm cc}(\boldsymbol{w}, \psi_w) = \frac{\|\boldsymbol{w}\|^2}{2\pi}[1 + \cos(2\psi_w)] - \frac{\|\boldsymbol{w}\|}{2\pi}[\cos \theta_{\boldsymbol{w},\boldsymbol{v}} + \cos(\theta_{\boldsymbol{w},\boldsymbol{v}} - 2\psi_w)] \ .$$

It is observed that the above expression is the same as the gradient of $L_{\rm cr}$ with respect to ψ in Eq. (22). The only difference comes from the domain of $\boldsymbol{w}$, which is $\|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R$ in Lemma 14 and $\|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R$ here. Then according to $\|\boldsymbol{x}\| \leqslant \sqrt{2}\|\boldsymbol{x}\|_\infty$ in $\mathbb{R}^2$, one knows from Lemma 14 that

$$-9 \leqslant -2\left(\frac{\pi}{2} - \psi_w\right)^2 - 2\left(\frac{\pi}{2} - \psi_w\right)|w_2| \leqslant \nabla_{\psi_w} L_{\rm cc}(\boldsymbol{w}, \psi_w) \leqslant -\frac{1}{4}\left(\frac{\pi}{2} - \psi_w\right)^2 \ ,$$

where the first inequality holds according to $|\pi/2 - \psi_w| \leqslant \pi/2$ and $|w_2| \leqslant 1$, and the third inequality holds based on $R \leqslant 1/2$. Then ψ_w' satisfies

$$\psi_w' \leqslant \psi_w + 9\eta \leqslant \psi_v + \theta_{\boldsymbol{w},\boldsymbol{v}} + 9\eta \leqslant \psi_u \ ,$$

where the second inequality holds from the condition $\theta_{\boldsymbol{w},\boldsymbol{v}} \geqslant |\psi_w - \psi_v|$ in the definition of D_2, and the third inequality holds according to $\theta_{\boldsymbol{w},\boldsymbol{v}} \leqslant \arcsin R \leqslant \psi_u - \psi_v - 9\eta$. Meanwhile, it is observed that the gradient is always negative, which implies $\psi_w' \geqslant \psi_w \geqslant \psi_l$. Thus, we have completed the proof. □

C.2 Convergence Rate Lemmas

This section presents sufficient conditions for inversely proportional convergence rates.

Lemma 28 *Let $f : K \to \mathbb{R}$ be a function with a global minimum x^*, where $K \subset \mathbb{R}$ indicates the convex domain satisfying $B(x^*, r_1) \subset K \subset B(x^*, r_2)$. Suppose that there exist constants c_1, c_3, g_l, g_u such that $c_1 \leqslant r_1/g_u$ and for any $x \in K$, then at least one of the following holds.*
1. $|x' - x^*| \leqslant (1 - c_3\eta)|x - x^*|$ and $(x' - x^*)(x - x^*) \geqslant 0$ with $x' = x - \eta\nabla f(x)$ and $\eta \in (0, c_1]$.

2. $g_l \leqslant \operatorname{sgn}(x - x^*)\nabla f(x) \leqslant g_u$ for any $x \neq x^*$ and $|\nabla f(x^*)| \leqslant g_u$.
Then for any $c_2 \geqslant \max\{1/c_3, 2r_2/g_l, 2c_1g_u/g_l\}$, the sequence $\{x_t\}_{t=1}^\infty$ generated by gradient descent $x_{t+1} = x_t - \eta_t\nabla f(x_t)$ with $x_0 \in K$ and $\eta_t = \min\{c_1, c_2/t\}$ satisfies

$$x_t \in K \quad \text{and} \quad |x_t - x^*| \leqslant \frac{a}{t} \quad \text{with} \quad a = \frac{c_2^2 g_l}{2c_1} \ .$$

Proof. Firstly, we prove $x_t \in K$. Suppose $x_k \in K$. We prove $x_{k+1} \in K$ by discussion.

1. If the first condition holds, x_{k+1} is a convex combination of x_k and x^*. Thus, $x_{k+1} \in K$.
2. If the second condition holds and $\operatorname{sgn}(x_{k+1} - x^*) = \operatorname{sgn}(x_k - x^*)$, then x_{k+1} is a convex combination of x_k and x^*. Thus, $x_{k+1} \in K$.
3. If the third condition holds and $\operatorname{sgn}(x_{k+1} - x^*) \neq \operatorname{sgn}(x_k - x^*)$, then one knows from $\eta_t \leqslant c_1$ and $|\nabla f(x)| \leqslant g_u$ that $|x_{k+1} - x^*| \leqslant c_1 g_u \leqslant r_1$, where the second inequality holds based on $c_1 \leqslant r_1/g_u$. Thus, $B(x^*, r_1) \subset K$ leads to $x_{k+1} \in K$.

Combining the cases above, mathematical induction completes the proof of $x_t \in K$.

Secondly, we prove $|x_t - x^*| \leqslant a/t$. Let $t_0 = c_2/c_1$. According to $c_2 \geqslant 2c_1 g_u/g_l \geqslant 2c_1$, one knows $t_0 \geqslant 2$. For $t < t_0$, it is observed that $|x_t - x^*| \leqslant r_2 \leqslant a t_0^{-1} \leqslant a t^{-1}$, where the first inequality holds based on $K \subset B(x^*, r_2)$, the second inequality holds because of $a = c_2^2 g_l/(2c_1) \geqslant r_2 t_0$. Thus, the conclusion holds for any $t < t_0$. Suppose that $|x_k - x^*| \leqslant a/k$ holds for $k \geqslant t_0 - 1$. We then prove $|x_{k+1} - x^*| \leqslant a/(k+1)$ by discussion.

1. If the first condition holds, then we have
$$|x_{k+1} - x^*| \leqslant \left[1 - c_2 c_3 (k+1)^{-1}\right] a k^{-1} \leqslant a(k+1)^{-1} \ ,$$
where the first inequality holds based on the first condition and the induction hypothesis, and the second inequality holds from $c_2 \geqslant 1/c_3$. Thus, the conclusion holds for $t = k+1$.
2. If the second condition holds and $\operatorname{sgn}(x_{k+1} - x^*) = \operatorname{sgn}(x_k - x^*)$, then one knows
$$|x_{k+1} - x^*| \leqslant a k^{-1} - c_2 g_l (k+1)^{-1} \leqslant a(k+1)^{-1} \ ,$$
where the first inequality holds from the induction hypothesis and the second condition, and the second inequality holds because of
$$\frac{a}{k} - \frac{c_2 g_l}{k+1} - \frac{a}{k+1} = \frac{c_2 g_l (t_0/2 - k)}{k(k+1)} \leqslant 0 \ ,$$
where the first equality holds from the choice of a and t_0, and the first inequality holds from $t_0 \geqslant 2$ and $k \geqslant t_0 - 1 \geqslant t_0/2$. Thus, the conclusion holds for $t = k+1$.
3. If the second condition holds and $\operatorname{sgn}(x_{k+1} - x^*) \neq \operatorname{sgn}(x_k - x^*)$, then it is observed that
$$|x_{k+1} - x^*| \leqslant c_2 g_u (k+1)^{-1} \leqslant a(k+1)^{-1} \ ,$$
where the first inequality holds from the second condition, and the second inequality holds based on $a = c_2^2 g_l/(2c_1) \geqslant c_2 g_u$. Thus, the conclusion holds for $t = k+1$.

Combining the cases above, we have completed the proof. □

Lemma 29 *Let $f : K \to \mathbb{R}$ represent a function with a global minimum x^*, where $K \subset \mathbb{R}$ indicates the convex domain satisfying $B(x^*, r_1) \subset K \subset B(x^*, r_2)$. Let $\{\theta_t\}_{t=0}^{\infty}$ be a positive sequence bounded by $\theta_t \leqslant a/t$. Suppose that there exist constants g_l, g_u such that $g_l \leqslant \operatorname{sgn}(x_t - x^*)\nabla f(x_t) \leqslant g_u$ if $|x_t - x^*| \geqslant \theta_t$ and $x \in K$, and $|\nabla f(x_t)| \leqslant g_u$ if $|x_t - x^*| \leqslant \theta_t$ and $x \in K$. Let $c_1 > 0$, and $c_2 \geqslant \max\{2r_2/g_l, 2c_1\}$. Suppose that the sequence $\{x_t\}_{t=1}^{\infty}$ generated by gradient descent $x_{t+1} = x_t - \eta_t \nabla f(x_t)$ with $x_0 \in K$ and $\eta_t = \min\{c_1, c_2/t\}$ satisfies $x_t \in K$ for any $t \in \mathbb{N}^+$. Then the following holds for any $t \in \mathbb{N}^+$*
$$|x_t - x^*| \leqslant \frac{b}{t} \quad \text{with} \quad b = \max\left\{2a + c_2 g_u, \frac{c_2^2 g_l}{2c_1}\right\} \ .$$

Proof. Let $t_0 = 2b/(c_2 g_l) \geqslant c_2/c_1 \geqslant 2$. For any $0 < t < t_0$, it is observed that

$$|x_t - x^*| \leqslant r_2 \leqslant \frac{c_2 g_l}{2} = \frac{b}{t_0} \leqslant \frac{b}{t} \ .$$

Thus, the conclusion holds for $0 < t < t_0$. Suppose that $|x_k - x^*| \leqslant b/k$ holds for $k \geqslant t_0 - 1$. We then prove $|x_{k+1} - x^*| \leqslant b/(k+1)$ by discussion.

1. If the first condition holds and $\operatorname{sgn}(x_{k+1} - x^*) = \operatorname{sgn}(x_k - x^*)$, then we have

$$|x_{k+1} - x^*| \leqslant |x_k - x^*| - \eta_{k+1} g_l \leqslant \frac{b}{k} - \frac{c_2 g_l}{k+1} \leqslant \frac{b}{k+1} \ ,$$

   where the second inequality holds from the induction hypothesis, and the third inequality holds based on $b = c_2 g_l t_0/2$ and $t_0/2 \leqslant t_0 - 1 \leqslant k$. Thus, the conclusion holds for $t = k+1$.

2. If the first condition holds and $\operatorname{sgn}(x_{k+1} - x^*) \neq \operatorname{sgn}(x_k - x^*)$, then we have

$$|x_{k+1} - x^*| \leqslant \eta_{k+1} g_u \leqslant \frac{c_2 g_u}{k+1} \leqslant \frac{b}{k+1} \ ,$$

   which implies that the conclusion holds for $t = k+1$.

3. If the second condition holds, then one knows

$$|x_{k+1} - x^*| \leqslant |x_k - x^*| + \eta_{k+1} g_u \leqslant \frac{a}{k} + \frac{c_2 g_u}{k+1} \leqslant \frac{b}{k+1} \ ,$$

   where the second inequality holds based on $|x_{k+1} - x^*| \leqslant \theta_{k+1} \leqslant a/(k+1)$, and the third inequality holds because of $b \geqslant 2a + c_2 g_u$. Thus, the conclusion holds for $t = k+1$.

Combining the cases above, we have completed the proof. □

Lemma 30 *Let $f : K \to \mathbb{R}$ represent a function with a global minimum x^*, where $K \subset \mathbb{R}$ indicates the convex domain satisfying $K \subset B(x^*, R)$. Let $\{x_t\}_{t=1}^{\infty}$ denote the sequence generated by gradient descent $x_{t+1} = x_t - \eta_t \nabla f(x_t)$ with $x_0 \in K$ and $\eta_t = \min\{c_1, c_2/t\}$, satisfying $x_t \in K$ for $t \in \mathbb{N}^+$. Suppose that the gradient satisfies $\nabla f(x_t) = d(x_t - x^*) + r_t$, where $d_l \leqslant d \leqslant d_u$ and $|r_t| \leqslant e/t$. If $c_1 \leqslant 1/d_u$ and $c_2 \geqslant 2/d_l$, then we have*

$$|x_t - x^*| \leqslant \frac{c}{t} \quad \textit{with} \quad c = \max\left\{\frac{c_2 R}{c_1}, c_2 e\right\} \ .$$

Proof. Let $t_0 = c_2/c_1$. We prove the conclusion by mathematical induction.

1. Base case. For $0 < t \leqslant t_0$, it is observed that $|x_t - x^*| \leqslant R \leqslant c t_0^{-1} \leqslant c t^{-1}$. Thus, the conclusion holds for $0 < t \leqslant t_0$.

2. Induction. Suppose that $|x_k - x^*| \leqslant c/k$ holds for $k \geqslant t_0 - 1$. Then we have

$$|x_{k+1} - x^*| = |(1 - d\eta_k)(x_k - x^*) - \eta_k r_k| \leqslant (1 - d\eta_k)|x_k - x^*| + \eta_k |r_k| \ ,$$

   where the first inequality holds by $d\eta_k \leqslant c_1 d_u \leqslant 1$. Then induction hypothesis implies

$$|x_{k+1} - x^*| \leqslant \left(1 - \frac{2}{k}\right)\frac{c}{k} + \frac{c_2 e}{k^2} \leqslant \frac{c}{k+1} \ ,$$

   where the first inequality holds according to $c_2 d_l \geqslant 2$, and the second inequality holds based on $c \geqslant c_2 e$. Thus, the conclusion holds for $t = k+1$.

Therefore, mathematical induction completes the proof. □

Appendix D. Proof of Theorem 4

We begin the proof with two lemmas. For any vector $\boldsymbol{a} \neq \boldsymbol{0} \in \mathbb{R}^2$ and $\theta \in [0, \pi]$, define $S(\boldsymbol{a}, \theta) = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \theta_{\boldsymbol{x}} \in [\theta_{\boldsymbol{a}} - \theta, \theta_{\boldsymbol{a}} + \theta]\}$ as the sector region with central angle 2θ that is symmetric with respect to $\boldsymbol{a}$. Let $\mathcal{N}_{\boldsymbol{a},\theta}$ be the truncated standard Gaussian distribution on $S(\boldsymbol{a}, \theta)$, of which the probability density function is $p(\boldsymbol{x}) = (2\theta)^{-1}\mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}\|^2}\mathbb{I}(\boldsymbol{x} \in S(\boldsymbol{a}, \theta))$. The next lemma provides a lower bound for the expected squared inner product on $S(\boldsymbol{a}, \theta)$.

Lemma 31 *Let $d = 1$. For any $\boldsymbol{w} \in \mathbb{R}^{2d}$, non-zero $\boldsymbol{a} \in \mathbb{R}^{2d}$, and $\theta \in [0, \pi/2]$, we have*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{a},\theta}}\left[\left(\boldsymbol{w}^\top\boldsymbol{x}\right)^2\right] \geqslant \frac{\theta^2}{3}\|\boldsymbol{w}\|^2 \ .$$

Proof. Let $\theta_{\boldsymbol{w}}$ indicate the phase of $\boldsymbol{w}$, i.e., $\boldsymbol{w} = \|\boldsymbol{w}\|(\sin\theta_{\boldsymbol{w}} + \cos\theta_{\boldsymbol{w}}\mathrm{i})$. Then calculating the expectation in the polar coordinate system leads to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{a},\theta}}\left[\left(\boldsymbol{w}^\top\boldsymbol{x}\right)^2\right] &= \frac{\|\boldsymbol{w}\|^2}{2\theta}\int_0^{+\infty}\int_{\theta_{\boldsymbol{a}}-\theta}^{\theta_{\boldsymbol{a}}+\theta} r^3(\cos\theta_{\boldsymbol{w}}\cos\phi + \sin\theta_{\boldsymbol{w}}\sin\phi)^2\mathrm{e}^{-\frac{1}{2}r^2}\ \mathrm{d}\phi\ \mathrm{d}r\\ &= \theta^{-1}\|\boldsymbol{w}\|^2\left[\theta + 2^{-1}\sin(2\theta)\cos(2\theta_{\boldsymbol{a},\boldsymbol{w}})\right] \ ,\end{aligned}$$

where the second equality holds based on $\cos(\theta_{\boldsymbol{a}} - \theta_{\boldsymbol{w}}) = \cos\theta_{\boldsymbol{a},\boldsymbol{w}}$. Then we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_{\boldsymbol{a},\theta}}\left[\left(\boldsymbol{w}^\top\boldsymbol{x}\right)^2\right] \geqslant \left[1 - (2\theta)^{-1}\sin(2\theta)\right]\|\boldsymbol{w}\|^2 \geqslant 3^{-1}\theta^2\|\boldsymbol{w}\|^2 \ ,$$

where the first inequality holds according to $\theta \in [0, \pi/2]$, and the second inequality holds because of $\sin(x) \leqslant x - x^3/12$ for all $\theta \in [0, \pi/2]$. Thus, we have completed the proof. □

The following lemma provides a lower bound for expressing a complex-valued vector with four real-valued vectors under a symmetric constant.

Lemma 32 *Let $\boldsymbol{v}_k \in \mathbb{R}^d$ with $k \in [4]$ and $\boldsymbol{v} \in \mathbb{R}^d$. If $\boldsymbol{v}_1 + \boldsymbol{v}_3 = \boldsymbol{v}_2 + \boldsymbol{v}_4$, then we have*

$$\sum_{k=1}^{4}\|\boldsymbol{v}_i - \boldsymbol{v}\cdot\mathbb{I}(k=1)\|^2 \geqslant \frac{1}{4}\|\boldsymbol{v}\|^2 \ .$$

Proof. According to the generalized mean inequality, one knows

$$\sum_{k=1}^{4}\|\boldsymbol{v}_i - \boldsymbol{v}\cdot\mathbb{I}(k=1)\|^2 \geqslant \frac{1}{4}\left(\sum_{k=1}^{4}\|\boldsymbol{v}_i - \boldsymbol{v}\cdot\mathbb{I}(k=1)\|\right)^2 \geqslant \frac{1}{4}\|(\boldsymbol{v}_1 - \boldsymbol{v}) - \boldsymbol{v}_2 + \boldsymbol{v}_3 - \boldsymbol{v}_4\|^2 = \frac{1}{4}\|\boldsymbol{v}\|^2 \ ,$$

where the second inequality holds because of the triangle inequality, and the first equality holds based on the condition $\boldsymbol{v}_1 + \boldsymbol{v}_3 = \boldsymbol{v}_2 + \boldsymbol{v}_4$. Thus, we have completed the proof. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. We define $\mathcal{N}_{\boldsymbol{\alpha},\mathbf{W}} = \sum_{i=1}^{n}\alpha_i\tau(\boldsymbol{w}_i^\top\boldsymbol{x})$ for simplicity. From $d = 1$, the weight vector $\boldsymbol{w}_i$ is a 2-dimensional real-valued vector. Let $\theta_{\boldsymbol{w}_i} = \arctan(w_{i,1}^{-1}w_{i,2}) \in (-\psi, 2\pi - \psi]$ denote the phase of $\boldsymbol{w}_i$. We assume $\theta_{\boldsymbol{v}} = 0$ without loss of generality. Denote by $\Theta_{\mathbf{W}}$ the $\pi/2$-symmetrical phase set induced from $\mathbf{W}$ and ψ, i.e.,

$$\Theta_{\mathbf{W}} = \left\{\theta_{\boldsymbol{w}_i} + \frac{(j-1)\pi}{2}\ \middle|\ i \in [n], j \in [4]\right\} \cup \left\{i\psi + \frac{(j-1)\pi}{2}\ \middle|\ i \in \{-1, +1\}, j \in [4]\right\} \ .$$

Then there is an integer $m \leqslant n+2$ such that $|\Theta_{\mathbf{W}}| = 4m$. We sort all phases in $\Theta_{\mathbf{W}}$ as

$$\Theta_{\mathbf{W}} = \{\theta_i\}_{i=1}^{4m} \quad \text{with} \quad -\psi < \theta_1 < \cdots < \theta_{4m} = 2\pi - \psi \ .$$

Let $\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}$ represent an arbitrary two-layer RVNN with weight phases from the set $\Theta_{\mathbf{W}}$, i.e., $\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}(\boldsymbol{x}) = \sum_{i=1}^{4m} \beta_i \tau(\boldsymbol{u}_i^\top \boldsymbol{x})$ with $\theta_{\boldsymbol{u}_i} = \theta_i$. It is observed that $\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}$ degenerates to $\mathcal{N}_{\boldsymbol{\alpha},\mathbf{W}}$ with suitable parameters. Thus, the expected square loss L_{rc} can be bounded as

$$\begin{aligned} L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) &\geqslant \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\left(\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}(\boldsymbol{x}) - \sigma_\psi(\boldsymbol{v}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) \right)^2 \right] \\ &= \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum_{i=1}^{4m} \frac{\Delta\theta_i}{\pi} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{a}_i,\Delta\theta_i)} \left[\left(\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}(\boldsymbol{x}) - \sigma_\psi(\boldsymbol{v}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) \right)^2 \right] , \end{aligned} \tag{45}$$

where $\Delta\theta_i = (\theta_i - \theta_{i-1})/2$ and $\boldsymbol{a}_i = \mathrm{e}^{(\theta_i - \Delta\theta_i)\mathrm{i}}$ with $\theta_0 = \theta_{4(n+1)}$. The indices can be divided into m groups as $\mathcal{I}_i = \{i + (k-1)m \mid k \in [4]\}$ with $i \in [m]$. Denote by i_ψ the index of ψ, i.e., $\theta_{i_\psi} = \psi$. Then Eq. (45) becomes

$$\begin{aligned} L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) &\geqslant \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum_{i=1}^{m} \frac{\Delta\theta_i}{\pi} \sum_{j\in\mathcal{I}_i} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{a}_j,\Delta\theta_j)} \left[\left(\mathcal{N}_{\boldsymbol{\beta},\mathbf{U}}(\boldsymbol{x}) - \sigma_\psi(\boldsymbol{v}_{\mathbb{C}}^\top \overline{\boldsymbol{x}}_{\mathbb{C}}) \right)^2 \right] \\ &= \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum_{i=1}^{m} \frac{\Delta\theta_i}{\pi} \sum_{j\in\mathcal{I}_i} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{a}_j,\Delta\theta_j)} \left[\left((\boldsymbol{v}_j - \boldsymbol{v} \cdot \mathbb{I}(j \leqslant i_\psi))^\top \boldsymbol{x} \right)^2 \right] , \end{aligned} \tag{46}$$

where the first inequality holds since $\Delta\theta_j$ remains the same in $\mathcal{I}_i$, the first equality holds based on the activation regions of ReLU and zReLU, and the definition of $\boldsymbol{v}_j$ as follows

$$\boldsymbol{v}_j = \sum_{l=j-m}^{j+m-1} \beta_{\phi(l)} \boldsymbol{u}_{\phi(l)} \quad \text{with} \quad \phi(l) = \begin{cases} l + 4m \ , & l \leqslant 0 \ , \\ l \ , & 0 < l \leqslant 4m \ , \\ l - 4m \ , & l > 4m \ . \end{cases} \tag{47}$$

Applying Lemma 31 to Eq. (46), we obtain

$$\begin{aligned} L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) &\geqslant \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum_{i=1}^{m} \frac{\Delta\theta_i}{\pi} \sum_{j\in\mathcal{I}_i} \frac{(\Delta\theta_j)^2}{3} \|\boldsymbol{v}_j - \boldsymbol{v} \cdot \mathbb{I}(j \leqslant i_\psi)\|^2 \\ &\geqslant \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum\nolimits_{i=\max\{1,i_\psi-m+1\}}^{\min\{i_\psi,m\}} \frac{(\Delta\theta_i)^3}{3\pi} \sum\nolimits_{k=1}^{4} \|\boldsymbol{v}_{i,k} - \boldsymbol{v} \cdot \mathbb{I}(k=1)\|^2 \ , \end{aligned}$$

where the second inequality holds by the definition of $\boldsymbol{v}_{i,k} = \boldsymbol{v}_{i+(k-1)(n+1)}$ and $\Delta\theta_j = \Delta\theta_i$ for any $j \in \mathcal{I}_i$. Based on Eq. (47), one has $\boldsymbol{v}_{i,1} + \boldsymbol{v}_{i,3} = \boldsymbol{v}_{i,2} + \boldsymbol{v}_{i,4}$. Then Lemma 32 implies

$$\begin{aligned} L_{\mathrm{rc}}(\boldsymbol{\alpha}, \mathbf{W}) &\geqslant \frac{1}{2} \inf_{\boldsymbol{\beta},\mathbf{U}} \sum\nolimits_{i=\max\{1,i_\psi-m+1\}}^{\min\{i_\psi,m\}} \frac{(\Delta\theta_i)^3}{3\pi} \cdot \frac{1}{4} \|\boldsymbol{v}\|^2 \\ &\geqslant \frac{\|\boldsymbol{v}\|^2}{24\pi(\max\{1, i_\psi - m + 1\} - \min\{i_\psi, m\})^2} \left(\sum\nolimits_{i=\max\{1,i_\psi-m+1\}}^{\min\{i_\psi,m\}} \Delta\theta_i \right)^3 \\ &\geqslant \frac{\|\boldsymbol{v}\|^2 \min\{2\psi, \pi - 2\psi\}^3}{24\pi(n+2)^2} \ , \end{aligned}$$

where the second inequality holds from the generalized mean inequality, and the third one holds by $\max\{1, i_\psi - m + 1\} - \min\{i_\psi, m\} \leqslant n + 2$. Thus, we have completed the proof. □

Appendix E. Proof of Theorem 6

We begin with a lemma providing a lower bound for convergence.

Lemma 33 *If there exists a constant c such that $\langle \nabla f(\boldsymbol{w}), \boldsymbol{w}-\boldsymbol{v}\rangle \leqslant c\|\boldsymbol{w}-\boldsymbol{v}\|^2$, then $\boldsymbol{w}' = \boldsymbol{w} - \eta\nabla f(\boldsymbol{w})$ with $\eta \in (0, 1/(2c))$ satisfies $\|\boldsymbol{w}'-\boldsymbol{v}\| \geqslant \sqrt{1-2c\eta}\|\boldsymbol{w}-\boldsymbol{v}\|$.*

Proof. One has $\|\boldsymbol{w}'-\boldsymbol{v}\|^2 \geqslant \|\boldsymbol{w}-\boldsymbol{v}\|^2 - 2\eta\langle \boldsymbol{w}-\boldsymbol{v}, \nabla f(\boldsymbol{w})\rangle \geqslant (1-2c\eta)\|\boldsymbol{w}-\boldsymbol{v}\|^2$ based on the updating rule. Thus, the proof is completed. □

We then prove Theorem 6.

Proof of Theorem 6. Denote by $R = \|\boldsymbol{w}_0 - \boldsymbol{v}\|$. The proof consists of several steps.

Step 1: the error of ψ decreases below a threshold fast. By the same arguments as those in the proof of Theorem 1, $\eta \in (0, 1/(12\pi))$ indicates $(\boldsymbol{w}_t, \psi_t) \in D$ for any $t \in \mathbb{N}$. Recalling the convergence of ψ in Eq. (12), we have $\psi_t \geqslant \pi/4$ when $t \geqslant \lceil 16\eta^{-1}(1-R^2)^{-1}\rceil$. From Eq. (9), one knows $\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \geqslant -6(\psi^*-\psi_t)$. Then we have

$$\langle \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t), \psi^* - \psi_t\rangle \geqslant -6(\psi^*-\psi_t)^2 \ .$$

Then we obtain from $\eta \in (0, 1/12)$ and Lemma 33 that

$$\psi^* - \psi_t \geqslant (1-12\eta)^{t/2}(\psi^*-\psi_0) \ . \tag{48}$$

Thus, one has

$$(1-12\eta)^{t/2}(\psi^*-\psi_0) \leqslant \psi^*-\psi_t \leqslant \frac{\pi}{4} \quad \text{with} \quad t \geqslant T_1 = 16\eta^{-1}(1-R^2)^{-1} \ .$$

Step 2: both errors of $\boldsymbol{w}$ and ψ decrease below small constants fast. According to Eq. (13), we have

$$\|\boldsymbol{w}_t - \boldsymbol{v}\| \leqslant \left(1-\frac{\eta}{48}\right)^{t-T_1} \quad \text{for} \quad t \geqslant T_1 \ , \tag{49}$$

which, together with Eqs. (12) and (48), implies that

$$\begin{aligned} &(1-12\eta)^{t/2}(\psi^*-\psi_0) \leqslant \psi^*-\psi_t \leqslant \frac{1}{384} \quad \text{and} \quad |w_2| \leqslant \|\boldsymbol{w}_t-\boldsymbol{v}\| \leqslant \frac{1}{384} \ , \\ &\text{with} \quad t \geqslant T_2 = \max\left\{T_1 + \frac{\ln 384}{\ln(1+\eta/48)}, \frac{3200\pi}{\eta(1-R^2)}\right\} \ . \end{aligned} \tag{50}$$

Step 3: $\boldsymbol{w}$ converges faster than ψ. For any $t \geqslant T_2$, Lemmas 13 and 14 imply

$$\langle \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t), \psi_t - \psi^*\rangle \leqslant 2(\psi^*-\psi_t)^3 + 2(\psi^*-\psi_t)^2|w_{t,2}| \leqslant \frac{1}{96}(\psi^*-\psi_t)^2 \ ,$$

where the second inequality holds based on Eq. (50). Then Lemma 33 indicates

$$\psi^* - \psi_{t+1} \geqslant \sqrt{1-\eta/48}(\psi^*-\psi_t) \quad \text{for} \quad t \geqslant T_2 \ ,$$

which, together with Eq. (49), indicates

$$|w_{w,t}| \leqslant \|\boldsymbol{w}_t - \boldsymbol{v}\| \leqslant \psi^*-\psi_t \quad \text{with} \quad t \geqslant T_3 = 2T_1 + \frac{T_2\ln(1-12\eta) + 2\ln(\psi^*-\psi_0)}{\ln(1-\eta/48)} \ . \tag{51}$$

Step 4: ψ converges with an inversely proportional rate. For any $t \geqslant T_3$, it is observed from Lemmas 13, 14, and Eq. (51) that $\nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \geqslant -4(\psi^* - \psi)^2$. Let $a_t = 4\eta(\psi^* - \psi_t)$. Then the updating rule implies $a_{t+1} \geqslant a_t(1 - a_t)$. Choosing $\eta \in (0, 1/(4\pi))$ guarantees $a_t \in [0, 1/2]$. Then Lemma 16 indicates

$$\psi^* - \psi_t \geqslant \frac{(1-12\eta)^{T_3/2}(\psi^* - \psi_0)}{t - T_3 + 1} \quad \text{for} \quad t \geqslant T_3 \ . \tag{52}$$

Step 5: the loss converges to 0 with a rate $O(t^{-1})$. Define non-negative quantities $\Delta_{\boldsymbol{w}} = \|\boldsymbol{w} - \boldsymbol{v}\|$ and $\Delta_\psi = \psi^* - \psi$. We provide a lower bound for L_{cr} by discussion.

1. Suppose $(\boldsymbol{w}, \psi) \in D_1$. Then we have

$$L_{\mathrm{cr}}(\boldsymbol{w}, \psi) \geqslant \frac{1}{4} - \frac{1}{8\pi}(4\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) = \frac{1}{8\pi}\Delta_\psi^3 + \frac{1}{8\pi}\Delta_{\boldsymbol{w}}^2(2\pi - \Delta_\psi^3) \geqslant \frac{1}{8\pi}\Delta_\psi^3 \ , \tag{53}$$

   where the first inequality holds based on $\sin(2\psi) + 2\psi = \sin(2\Delta_\psi) + 2\psi^* - 2\Delta_\psi \leqslant 2\psi^* - \Delta_\psi^3/2$ for any $\psi \in [0, \pi/2]$, and the second inequality holds from $\Delta_\psi \leqslant \pi/2$.

2. Suppose $(\boldsymbol{w}, \psi) \in D_2$. Recalling that $s_w = \sin(2\psi) + 2\psi$. The expected loss satisfies

$$\begin{aligned} L_{\mathrm{cr}}(\boldsymbol{w}, \psi) &= \frac{1}{4} - \frac{1}{4\pi} s_w (1 - \Delta_{\boldsymbol{w}}^2) + \frac{1}{4\pi}[(\cos(2\psi) - 1)|w_2| + (s_w + 2\theta - 2\psi^*)w_1] \\ &\geqslant \frac{1}{4} - \frac{1}{8\pi}(4\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) + \frac{1}{4\pi}[(\cos(2\psi) - 1)|w_2|] \\ &\geqslant \frac{1}{8\pi}\Delta_\psi^3 - \frac{1}{2\pi}\Delta_{\boldsymbol{w}} \ , \end{aligned} \tag{54}$$

   where the first inequality holds from $s_w \leqslant 2\psi^* - \Delta_\psi^3/2$ and $s_w + 2\theta - 2\psi^* \geqslant 0$, the second inequality holds based on $\cos(2\psi) - 1 \geqslant -2$ and $|w_2| \leqslant \Delta_{\boldsymbol{w}}$.

Combining Eqs. (53) and (54), for any $(\boldsymbol{w}_0, \psi_0) \in D$ and $t \geqslant T_3$, one has

$$L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi_t) \geqslant \frac{1}{8\pi}\Delta_{\psi,t}^3 - \frac{1}{2\pi}\Delta_{\boldsymbol{w},t} \geqslant \frac{(1-12\eta)^{3T_3/2}(\psi^* - \psi_0)^3}{8\pi(t - T_3 + 1)^3} - \frac{1}{2\pi}\left(1 - \frac{\eta}{48}\right)^{t - T_3} \ ,$$

where the second inequality holds by Eqs. (49) and (52). Thus, the proof is completed. □

Appendix F. Proof of Theorem 7

In the main part of this section, we present the detailed proof of Theorem 7. Subsection F.1 provides the optimization behaviors in a certain ideal region.

Proof of Theorem 7. For $R = 1/5$, define the ideal region as

$$D_\phi = \{(\boldsymbol{w}, \phi) \mid \|\boldsymbol{w} - \boldsymbol{v}\| \leqslant R, \phi \in [-r, r], \theta \in [0, \pi/2 - \psi(\phi)]\} \ . \tag{55}$$

We divide the proof into four steps.

Step 1: D_ϕ is closed under gradient descent. We first prove the maintenance of inclusion by mathematical induction, i.e., $(\boldsymbol{w}_0, \phi_0) \in D_\phi$ indicates $(\boldsymbol{w}_t, \phi_t) \in D_\phi$.

1. Base case. The conclusion holds for $t = 0$ from the condition.

2. Induction. Suppose that the conclusion holds for $t = k$ with $k \in \mathbb{N}$. According to $\eta < 4 < 3/(a_1^2\pi^2)$ and Lemma 34, we have

$$|\phi_{k+1}| \leqslant \left[1 - \frac{a_1^2(1-R^2)\pi^2}{192}\eta\right] |\phi_k| \leqslant |\phi_k| \leqslant r \ , \tag{56}$$

where the third inequality holds from the induction hypothesis. Meanwhile, Lemma 11 and $\eta < 4$ imply that

$$\|\boldsymbol{w}_{k+1} - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{4\pi}[\sin(2\psi(\phi_k)) + 2\psi(\phi_k)]\right) \|\boldsymbol{w}_k - \boldsymbol{v}\| \leqslant \|\boldsymbol{w}_k - \boldsymbol{v}\| \leqslant R \ , \tag{57}$$

where the third inequality holds from the induction hypothesis. Furthermore, we have

$$\begin{aligned} \theta_{k+1} &\leqslant \left[1 - \frac{2(1-R^2)\eta}{9}\left(1 - (2a_1 r)^{2/3}\right)\right]\theta_k \\ &\leqslant \left[1 - \frac{2(1-R^2)\eta}{9}\left(1 - (2a_1 r)^{2/3}\right)\right][\psi^* - \psi(\phi_k)] \\ &\leqslant \left[1 - \frac{2(1-R^2)\eta}{9}\left(1 - (2a_1 r)^{2/3}\right)\right]\left[1 - \frac{4a_1^2\pi^2\eta}{3}\right]^{-1}[\psi^* - \psi(\phi_{k+})] \, , \end{aligned}$$

where the first inequality holds based on Lemma 35 and $\eta < \min\{4, 4R^{-1}(1-R)\}$, the second inequality holds from the induction hypothesis, and the third inequality holds according to Lemma 36 and $\eta < 4 < 3/(4a_1^2\pi^2)$. It is observed that

$$\begin{aligned} &\left[1 - \frac{2(1-R^2)\eta}{9}\left(1 - (2a_1 r)^{2/3}\right)\right]\left[1 - \frac{4a_1^2\pi^2\eta}{3}\right]^{-1} - 1 \\ = \ &\frac{2\eta}{9}\left[1 - \frac{4a_1^2\pi^2\eta}{3}\right]^{-1}\left[(1-R^2)(2a_1 r)^{2/3} + 6\pi^2 a_1^2 - (1-R^2)\right] \\ \leqslant \ &\frac{2\eta(1-R^2)}{9}\left[1 - \frac{4a_1^2\pi^2\eta}{3}\right]^{-1}\left[\left(\frac{1}{5}\right)^{2/3} + \frac{6\pi^2}{100} - 1\right] \\ \leqslant \ &0 \ , \end{aligned}$$

where the first inequality holds from $r < 1/(10a_1)$ and $a_1 \leqslant 1/12 \leqslant \sqrt{1-R^2}/10$. Combining the above two inequalities, we obtain

$$\theta_{k+1} \leqslant \psi^* - \psi(\phi_{k+1}) \ . \tag{58}$$

Combining Eqs. (56), (57) and (58), the conclusion holds for $t = k+1$.

Therefore, mathematical induction implies $(\boldsymbol{w}_t, \psi_t) \in D_\phi$ when $(\boldsymbol{w}_0, \psi_0) \in D_\phi$.

Step 2: convergence of parameters in D_ϕ. According to $\psi(\phi_k) \in [0, \pi/2]$, one has

$$\sin(2\psi(\phi_k)) + 2\psi(\phi_k) \geqslant 2\psi(\phi_k) = \pi\left[1 - g(\phi_k)^{2/3}\right] \geqslant \pi\left[1 - (2a_1 r)^{2/3}\right] \geqslant 2 \ ,$$

where the first equality holds from the reparameterization, the second inequality holds based on Eq. (4) and $|\phi_k| \leqslant r$, and the third inequality holds because of $r < 1/(10a_1)$. Then one knows from Eq. (57) that

$$\|\boldsymbol{w}_t - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{2\pi}\right)\|\boldsymbol{w}_{t-1} - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{2\pi}\right)^t \|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant \left(1 - \frac{\eta}{2\pi}\right)^t , \tag{59}$$

where the third inequality holds from $\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant R \leqslant 1$. Meanwhile, Eq. (56) indicates

$$|\phi_k| \leqslant \left(1 - \frac{a_1^4\pi^2}{2}\eta\right)|\phi_{k-1}| \leqslant \left(1 - \frac{a_1^4\pi^2}{2}\eta\right)^t |\phi_0| \leqslant \frac{1}{10a_1}\left(1 - \frac{a_1^4\pi^2}{2}\eta\right)^t , \tag{60}$$

where the first inequality holds based on $a_1 \leqslant \sqrt{1-R^2}/10$, and the third inequality holds from $|\phi_0| \leqslant r \leqslant 1/(10a_1)$.

Step 3: the loss converges to 0 in D_ϕ. We estimate the convergence rate of the expected loss when $(\boldsymbol{w}_0, \phi_0) \in D_\phi$. For any $(\boldsymbol{w}, \phi) \in D_\phi$, define non-negative quantities $\Delta_{\boldsymbol{w}} = \|\boldsymbol{w} - \boldsymbol{v}\|$, $\Delta_\psi = \psi^* - \psi(\phi)$, and $\Delta_\phi = |\phi|$. Then we have

$$L_{\mathrm{cr}}(\boldsymbol{w}, \psi(\phi)) \leqslant \frac{1}{4} - \frac{1}{2\pi}(\psi^* - \Delta_\psi^3)(1 - \Delta_{\boldsymbol{w}}^2) \leqslant \frac{1}{2\pi}\Delta_\psi^3 + \frac{1}{4}\Delta_{\boldsymbol{w}}^2 \leqslant \frac{a_1^2\pi^2}{4}\Delta_\phi^2 + \frac{1}{4}\Delta_{\boldsymbol{w}}^2 , \tag{61}$$

where the first inequality holds based on $\sin(2\psi)+2\psi = \sin(2\Delta_\psi)+2\psi^* - 2\Delta_\psi \geqslant 2\psi^* - 2\Delta_\psi^3$, the second inequality holds from non-negative Δ_ψ, and the third inequality holds because of $\Delta_\psi = \psi^* g(\phi)^{2/3} \leqslant \psi^*(2a_1\Delta_\phi)^{2/3}$. Combining Eqs. (59), (60), and (61), we obtain

$$L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi(\phi_t)) \leqslant \frac{\pi^2}{400}\left(1 - \frac{a_1^4\pi^2}{2}\eta\right)^{2t} + \frac{1}{4}\left(1 - \frac{\eta}{2\pi}\right)^{2t} \leqslant \frac{1}{3}\left(1 - \frac{a_1^4\pi^2}{2}\eta\right)^{2t} , \tag{62}$$

where the second inequality holds from $a_1^4\pi^2/2 \leqslant 1/(2\pi)$.

Step 4: initialization falls into D_ϕ with constant probability. Denote by $p(x)$ the probability density function of $\mathcal{N}(0,1)$. Define $r_0 = (a_1\phi_0/2)^{2/3}$. Then one has

$$\begin{aligned}
\Pr[(\boldsymbol{w}_0, \phi_0) \in D_\phi] &= \Pr[\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant R \wedge \phi_0 \in [-r, r] \wedge \theta_0 \leqslant \psi^* - \psi(\phi_0)] \\
&\geqslant \Pr[\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant R \wedge \phi_0 \in [-r, r] \wedge \|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant \sin(\psi^* - \psi(\phi_0))] \\
&\geqslant \Pr\left[\|\boldsymbol{w}_0 - \boldsymbol{v}\| \leqslant (a_1\phi_0/2)^{2/3} \wedge \phi_0 \in [-r, r]\right] \\
&= \int_{\mathbb{R}^3} \mathbb{I}\left(\boldsymbol{w} \in B(\boldsymbol{v}, r)\right)\mathbb{I}(|\phi| \leqslant r) p(\boldsymbol{w}) p(\phi) \ \mathrm{d}\boldsymbol{w} \ \mathrm{d}\phi ,
\end{aligned}$$

where the second inequality holds from the reparameterization, Eq. (4), $\sin(x) \geqslant 2x/\pi$ for $x \in [0, \pi/2]$, and $(a_1\phi_0/2)^{2/3} \leqslant (a_1 r/2)^{2/3} \leqslant (1/20)^{2/3} \leqslant R$. It is observed that

$$\int_{\mathbb{R}^2} \mathbb{I}\left(\boldsymbol{w} \in B(\boldsymbol{v}, r)\right) p(\boldsymbol{w}) \ \mathrm{d}\boldsymbol{w} \geqslant \mu(B(\boldsymbol{v}, r)) \min_{\boldsymbol{w} \in B(\boldsymbol{v}, r)} p(\boldsymbol{w}) \geqslant \frac{r^2}{2\mathrm{e}^2} .$$

Define $\Omega = B(0, r) \backslash B(0, r/2)$. Combining the above two inequalities, we obtain

$$\Pr[(\boldsymbol{w}_0, \phi_0) \in D_\phi] \geqslant \mu(\Omega) \min_{\phi \in \Omega} \frac{(a_1\phi/2)^{4/3} p(\phi)}{2\mathrm{e}^2} \geqslant \frac{a_1^{4/3} r^{10/3} \mathrm{e}^{-r^2/2}}{100} . \tag{63}$$

Thus, we conclude from Eqs. (62) and (63) that

$$\Pr\left[L_{\mathrm{cr}}(\boldsymbol{w}_t, \psi(\phi_t)) \leqslant \frac{1}{3}\left(1 - \frac{a_1^4\pi^2}{2}\eta\right)^{2t}\right] \geqslant \frac{a_1^{4/3} r^{10/3} \mathrm{e}^{-r^2/2}}{100} ,$$

which completes the proof. □

F.1 Optimization Behaviors

The following lemma depicts the optimization behavior of ϕ in D_ϕ.

Lemma 34 *Let $\phi_{k+1} = \phi_k - \eta \nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k))$ for an integer $k \in \mathbb{N}$, and a_1 is the non-zero constant in the second requirement of the reparameterization in Eq. (3). If $(\boldsymbol{w}_k, \phi_k) \in D_\phi$ and $\eta \in (0, 3/(a_1^2\pi^2))$, then we have $\phi_k\phi_{k+1} \geqslant 0$ and*

$$\left[1 - 3^{-1}a_1^2\pi^2\eta\right]|\phi_k| \leqslant |\phi_{k+1}| \leqslant \left[1 - 192^{-1}a_1^2(1-R^2)\pi^2\eta\right]|\phi_k| \ .$$

Proof. Recalling the definition of D_1 in Eq. (8), one knows that $(\boldsymbol{w}, \psi(\phi)) \in D_1$ holds for any $(\boldsymbol{w}, \phi) \in D_\phi$. Then for any $(\boldsymbol{w}_k, \phi_k) \in D_\phi$, Lemma 13 implies

$$-\pi^{-1}[\psi(\phi^*) - \psi(\phi_k)]^2 \leqslant \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \leqslant -(4\pi)^{-1}(1-R^2)\left[\psi(\phi^*) - \psi(\phi_k)\right]^2 \leqslant 0 \ ,$$

where $\phi^* = 0$, and the third inequality holds from $R < 1$. Substituting the reparameterization $\psi(\phi) = \psi^*\left[1 - g(\phi)^{2/3}\right]$ into the above inequalities, we obtain

$$-\pi^{-1}(\psi^*)^2 g(\phi_k)^{4/3} \leqslant \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \leqslant -(4\pi)^{-1}(1-R^2)(\psi^*)^2 g(\phi_k)^{4/3} \leqslant 0 \ . \tag{64}$$

According to the reparameterization and the chain rule, we have

$$\begin{aligned}\nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) &= \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k))\psi'(\phi_k) \\ &= -\frac{2}{3}\psi^* g(\phi_k)^{-1/3} g'(\phi_k) \nabla_\psi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \ .\end{aligned} \tag{65}$$

We continue the calculation of $\nabla_\phi L_{\mathrm{cr}}$ by discussion.

- Suppose $\phi_k > 0$. Then one knows from Eq. (4) that $g(\phi_k)^{-1/3}g'(\phi_k) > 0$, which, together with Eqs. (64) and (65), implies

  $$12^{-1}\pi^2 g(\phi_k)g'(\phi_k) \geqslant \nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \geqslant 48^{-1}(1-R^2)\pi^2 g(\phi_k)g'(\phi_k) > 0 \ .$$

  Based on the second requirement of the reparameterization, one has $g(0) = 0$, which, together with Eq. (4) and the above inequalities, leads to

  $$3^{-1}a_1^2\pi^2\phi_k \geqslant \nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \geqslant 192^{-1}a_1^2(1-R^2)\pi^2\phi_k > 0 \ .$$

- Suppose $\phi_k < 0$. Similarly, the gradient with respect to ϕ satisfies

  $$3^{-1}a_1^2\pi^2\phi_k \leqslant \nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) \leqslant 192^{-1}a_1^2(1-R^2)\pi^2\phi_k < 0 \ .$$

- Suppose $\phi_k = 0$. By substituting the reparameterization into the loss L_{cr} in Eq. (7) and calculating the derivative of L_{cr} with respect to ϕ, we have

  $$\nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) = 0 \ .$$

Thus, the updating rule $\phi_{k+1} = \phi_k - \eta\nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k))$ with $\eta \in (0, 3/(a_1^2\pi^2))$ leads to $\phi_k\phi_{k+1} \geqslant 0$ and

$$\left[1 - 3^{-1}a_1^2\pi^2\eta\right]|\phi_k| \leqslant |\phi_{k+1}| \leqslant \left[1 - 192^{-1}a_1^2(1-R^2)\pi^2\eta\right]|\phi_k| \ , \tag{66}$$

which completes the proof. □

The following lemma describes the optimization behavior of $\theta_{\boldsymbol{w},\boldsymbol{v}}$ in D_ϕ.

Lemma 35 *Let $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k))$ for an integer $k \in \mathbb{N}$, $\theta_k = \theta_{\boldsymbol{w}_k, \boldsymbol{v}}$, and a_1 is the non-zero constant in the second requirement of the reparameterization in Eq. (3). If $(\boldsymbol{w}_k, \phi_k) \in D_\phi$ and $\eta \in (0, \min\{4, 4R^{-1}(1-R)\})$, then we have*

$$\theta_{k+1} \leqslant \left[1 - \frac{2(1-R^2)\eta}{9}\left(1 - (2a_1 r)^{2/3}\right)\right]\theta_k \ .$$

Proof. According to the definition of D_ϕ in Eq. (55), one has $\theta_k \leqslant \pi/2 - \psi(\phi_k)$. Then based on the expression of L_{cr} in Eq. (7), we have

$$\nabla_{\boldsymbol{w}} L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k)) = (4\pi)^{-1}[\sin(2\psi(\phi_k)) + 2\psi(\phi_k)](\boldsymbol{w} - \boldsymbol{v}) \ , \tag{67}$$

We then prove the conclusion by discussion.

1. Suppose $\theta_k = 0$. Let $\boldsymbol{w}_k = (w_{k,1}, w_{k,2})$. Then $\theta_k = \arctan|w_{k,2}/w_{k,1}|$ and $(\boldsymbol{w}_k, \phi_k) \in D_\phi$ imply $w_{k,2} = 0$ and $w_{k,1} > 0$. Based on Eq. (67) and $\eta \in (0,4)$, one has $w_{k+1,2} = 0$ and $w_{k+1,1} > 0$, which implies $\theta_{k+1} = 0$. Thus, the conclusion holds.

2. Suppose $\theta_k \neq 0$. Then one knows from $\theta_k = \arctan|w_{k,2}/w_{k,1}|$ that

   $$\theta_k^{-1}\theta_{k+1} = \arctan\left|\frac{w_{k,2} - c_k w_{k,2}}{w_{k,1} - c_k(w_{k,1}-1)}\right| \left[\arctan\left|\frac{w_{k,2}}{w_{k,1}}\right|\right]^{-1} ,$$

   where $c_k = \eta[\sin(2\psi(\phi_k)) + 2\psi(\phi_k)]/(4\pi)$ is irrelevant to $\boldsymbol{w}_k$. Let $g(x) = \arctan(x)$, $a = |(w_{k,2} - c_k w_{k,2})/(w_{k,1} - c_k(w_{k,1}-1))|$, and $b = |w_{k,2}/w_{k,1}|$. Then we have

   $$\theta_k^{-1}\theta_{k+1} = \frac{g(a)}{g(b)} \leqslant \frac{g(a) + [b - g(b)]}{g(b) + [b - g(b)]} = 1 - \frac{g'(\xi)(b-a)}{b} \leqslant 1 - \frac{g'(b)(b-a)}{b} \ ,$$

   where the first inequality holds from $g(a) \leqslant g(b)$ and $b - g(b) \geqslant 0$, the second equation holds based on Lagrange's mean value theorem with $\xi \in (a,b)$, and the second inequality holds since $g'(x)$ is monotonically increasing for $x > 0$. Substituting the definition of a, b, and g into the above inequality, we obtain

   $$\theta_k^{-1}\theta_{k+1} \leqslant 1 - \frac{w_{k,1}^2}{(w_{k,1}^2 + w_{k,2}^2)[w_{k,1} - c_k(w_{k,1}-1)]}c_k \leqslant 1 - \frac{w_{k,1}^2}{(w_{k,1}^2 + w_{k,2}^2)(w_{k,1}+1)}c_k \ ,$$

   where the first inequality holds since $\eta < 4R^{-1}(1-R)$ implies $|c_k(w_{k,1}-1)| \leqslant w_{k,1}$, and the second inequality holds because $\eta < 4$ leads to $-c_k(w_{k,1}-1) \leqslant c_k \leqslant 1$. According to $(w_{k,1}-1)^2 + w_{k,2}^2 \leqslant R^2$ for $(\boldsymbol{w}_k, \phi_k) \in D_\phi$, one has

   $$\theta_k^{-1}\theta_{k+1} \leqslant 1 - \frac{w_{k,1}^2}{[2w_{k,1} - (1-R^2)](w_{k,1}+1)}c_k \leqslant 1 - \frac{4(1-R^2)}{9}c_k \ , \tag{68}$$

   where the first inequality holds because the upper bound is monotonically increasing with respect to $w_{k,2}$, and the second inequality holds since the upper bound is maximized at $w_{k,1} = 2(1-R^2)/(1+R^2)$. It is observed that c_k satisfies

   $$c_k = (4\pi)^{-1}\eta[\sin(2\psi(\phi_k)) + 2\psi(\phi_k)] \geqslant (\pi)^{-1}\eta\psi(\phi_k) \geqslant 2^{-1}\eta\left[1 - (2a_1 r)^{2/3}\right] , \tag{69}$$

where the second inequality holds from Eqs. (3), (4), and (55). Substituting Eq. (69) into Eq. (68), we conclude

$$\theta_k^{-1}\theta_{k+1} \leqslant 1 - \frac{2(1-R^2)\eta}{9}\left(1-(2a_1 r)^{2/3}\right),$$

which completes the proof of this case.

Combining the cases above completes the proof. □

The following lemma characterizes the optimization behavior of $\psi(\phi)$ in D_ϕ.

Lemma 36 *Let $\phi_{k+1} = \phi_k - \eta\nabla_\phi L_{\mathrm{cr}}(\boldsymbol{w}_k, \psi(\phi_k))$ for an integer $k \in \mathbb{N}$, and a_1 is the non-zero constant in the second requirement of the reparameterization in Eq.* (3)*. If $(\boldsymbol{w}_k, \phi_k) \in D_\phi$ and $\eta \in (0, 3/(4a_1^2\pi^2))$, then we have*

$$\psi^* - \psi(\phi_{k+1}) \geqslant \left[1 - \frac{4a_1^2\pi^2\eta}{3}\right][\psi^* - \psi(\phi_k)].$$

Proof. We prove the conclusion by discussion.

1. Suppose $|g(\phi_k)| = 0$. Then the reparameterization in Eq. (3) implies $\psi(\phi_k) = \psi^*$. According to Lemma 34, one knows that $\psi(\phi_{k+1}) = \psi^*$, which indicates $|g(\phi_{k+1})| = 0$. Thus, the conclusion holds since $\psi^* - \psi(\phi_{k+1}) = \psi^* - \psi(\phi_k) = 0$.
2. Suppose $|g(\phi_k)| \neq 0$. According to Lemma 34 and $\eta < 3/(4a_1^2\pi^2) < 3/(a_1^2\pi^2)$, one knows that $\phi_k\phi_{k+1} \geqslant 0$ and $|\phi_k| \geqslant |\phi_{k+1}|$. Thus, Eq. (4) implies

$$\frac{|g(\phi_{k+1})|}{|g(\phi_k)|} = \frac{|g(\phi_k)| - |g(\phi_k) - g(\phi_{k+1})|}{|g(\phi_k)|} \geqslant 1 - \frac{2a_1|\phi_k - \phi_{k+1}|}{a_1|\phi_k|/2} \geqslant 1 - \frac{4a_1^2\pi^2\eta}{3},$$

where the first inequality holds from Lagrange's mean value theorem and Eq. (4), and the second inequality holds based on the lower bound of $|\phi_{k+1}|$ in Lemma 34. Recalling the parameterization $\psi(\phi) = \psi^*\left[1 - g(\phi)^{2/3}\right]$, one has

$$\frac{\psi^* - \psi(\phi_{k+1})}{\psi^* - \psi(\phi_k)} = \left[\frac{|g(\phi_{k+1})|}{|g(\phi_k)|}\right]^{2/3} \geqslant \left(1 - \frac{4a_1^2\pi^2\eta}{3}\right)^{2/3} \geqslant 1 - \frac{4a_1^2\pi^2\eta}{3},$$

where the second inequality holds from $\eta < 3/(4a_1^2\pi^2)$. Thus, the conclusion holds.

Combining the cases above completes the proof. □

Appendix G. Proof of Theorem 8

Proof. According to the reparameterization $\psi(\phi) = \psi^*[1-g(\phi)^{2/3}]$ and $g(\phi) = a_1\phi + O(\phi^2)$, the inverse function of $\psi(\cdot)$ is multiple-valued. Based on the strict monotonicity of g, we define $\psi^{-1}(\cdot)$ as the positive branch of the inverse function of $\psi(\cdot)$, i.e.,

$$\psi^{-1}(\psi) = \phi \quad \text{with} \quad \psi(\phi) = \psi \quad \text{and} \quad \phi \geqslant 0 .$$

Let $R = 1/25$, $\psi_l = \psi_v - 109R/100$, and $\psi_u = \psi_v + 109R/100$ as used in Lemma 21. According to $\eta < \bar{\eta} = R/120$, all conditions in Lemmasd 22-27 are satisfied. Define

$$D_{\phi,1} = \{(\boldsymbol{w}, \phi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \phi_w \in [\phi_l, \phi_u], \boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{v}} \in [0, |\psi(\phi_w) - \psi(\phi_v)|]\},$$
$$D_{\phi,2} = \{(\boldsymbol{w}, \phi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \phi_w \in [\phi_l, \phi_u], \boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{v}} \in [|\psi(\phi_w) - \psi(\phi_v)|, \psi(\phi_w) + \psi(\phi_v)]\},$$

where $\phi_l = \psi^{-1}(\psi_u)$ and $\phi_u = \psi^{-1}(\psi_l)$. Let $D_\phi = D_{\phi,1} \cup D_{\phi,2}$ denote the ideal region, i.e.,

$$D_\phi = \{(\boldsymbol{w}, \phi_w) \mid \|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R, \phi_w \in [\phi_l, \phi_u], \theta_{\boldsymbol{w},\boldsymbol{v}} \in [0, \psi(\phi_w) + \psi(\phi_v)]\} \ .$$

For any $\phi \in [\phi_l, \phi_u]$, one has $g'(\phi) \in [k, K]$ and

$$3/2 \leqslant (1 - \psi_l/\psi^*)^{-1/2} = g(\phi_u)^{-1/3} \leqslant g(\phi)^{-1/3} \leqslant g(\phi_l)^{-1/3} = (1 - \psi_u/\psi^*)^{-1/2} \leqslant 5/2 \ ,$$

where we use the monotonicity of g. Then the reparameterization implies

$$\psi'(\phi) = -3^{-1}\pi g(\phi)^{-1/3} g'(\phi) \in [-5\pi K/6, -\pi k/2] \quad \text{for} \quad \phi \in [\phi_l, \phi_u] \ . \tag{70}$$

We divide the rest of the proof into several steps.

Step 1: D_ϕ is closed under gradient descent. For any $(\boldsymbol{w}_t, \phi_{w,t}) \in D_\phi$ and $\psi_{w,t} = \psi(\phi_{w,t})$, we prove $(\boldsymbol{w}_{t+1}, \phi_{w,t+1}) \in D_\phi$ as follows.

1. It is observed that

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla_{\boldsymbol{w}} L_{\text{cc}}(\boldsymbol{w}_t, \psi(\phi_{w,t})) = \boldsymbol{w}_t - \eta_t \nabla_{\boldsymbol{w}} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \ ,$$

which is the same as the updating rule of $\boldsymbol{w}$ before reparameterization. Then based on Lemmas 22-25, we have $\|\boldsymbol{w} - \boldsymbol{v}\|_\infty \leqslant R$, and the first condition holds.

2. We prove the second condition by discussion.

   - Suppose $\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \geqslant 0$. Then one has

$$\phi_{w,t+1} = \phi_{w,t} - \eta_t \nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \psi'(\phi_{w,t}) \geqslant \phi_{w,t} \geqslant \phi_l \ , \tag{71}$$

   where the first equality holds from the chain rule, the first inequality holds based on $\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \geqslant 0$ and Eq. (70), and the second inequality holds according to $(\boldsymbol{w}_t, \phi_{w,t}) \in D_\phi$. Meanwhile, one knows from Lemmas 21, 26, and 27 that $\psi_t - \bar{\eta}\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \geqslant \psi_l$, which implies

$$\phi_{w,t+1} \leqslant \phi_{w,t} - \bar{\eta}^{-1}\eta_t \psi'(\phi_{w,t})(\psi_{w,t} - \psi_l) = \phi_{w,t} + \bar{\eta}^{-1}\eta_t \psi'(\phi_{w,t})\psi'(\xi)(\phi_u - \phi_{w,t}) \ ,$$

   where the first equality holds from Lagrange's mean value theorem with $\xi \in (\phi_{w,t}, \phi_u)$. Using Eq. (70) twice, we obtain

$$\phi_{w,t+1} \leqslant \phi_{w,t} + \bar{\eta}^{-1}\pi^2 \eta_t K^2 (\phi_u - \phi_{w,t}) \leqslant \phi_{w,t} + (\phi_u - \phi_{w,t}) \leqslant \phi_u \ , \tag{72}$$

   where the second inequality holds from $\eta_t \leqslant \bar{\eta}/(\pi^2 K^2)$. Combining Eqs. (71) and (72), we have $\phi_{w,t+1} \in [\phi_l, \phi_u]$.

   - Suppose $\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) < 0$. The proof of $\phi_{w,t+1} \in [\phi_l, \phi_u]$ is similar to the proof for $\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t}) \geqslant 0$.

   Combining the cases above, the second condition holds.

3. Let $\theta_k = \theta_{\boldsymbol{w}_k,\boldsymbol{v}}$ for an integer $k \in \mathbb{N}$. Then $\|\boldsymbol{w}_k - \boldsymbol{v}\|_\infty \leqslant R$ indicates

$$\theta_k = \arctan\left|\frac{w_{k,2}}{w_{k,1}}\right| \leqslant \left|\frac{w_{k,2}}{w_{k,1}}\right| \leqslant \frac{R}{1-R} = \frac{1}{24} \leqslant \psi(\phi_w) + \psi(\phi_v) \ ,$$

where the second equality hodls from the choice $R = 1/25$, and the third inequality hodls based on $\psi(\phi_v) \geqslant 7\pi/20$. Thus, the third condition holds from the first condition.

Combining the results above, we have $(\boldsymbol{w}_{t+1}, \phi_{w,t+1}) \in D_\phi$ for any $(\boldsymbol{w}_t, \phi_{w,t}) \in D_\phi$.
Step 2: w_2 converges to 0. For $(\boldsymbol{w}_0, \phi_{w,0}) \in D_\phi$, since reparameterization does not affect the gradient with respect to $w_{t,2}$, one knows from Eq. (32) that

$$|w_{t,2}| \leqslant \frac{c_2^2}{4\pi c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{73}$$

Step 3: ϕ_w converges to ϕ_v. Let $(\boldsymbol{w}_0, \phi_{w,0}) \in D_\phi$. For $(\boldsymbol{w}_0, \phi_{w,0}) \in D_{\phi,1}$, one has

$$\operatorname{sgn}(\phi_{w,t} - \phi_v)\nabla_{\phi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi(\phi_{w,t})) = -\operatorname{sgn}(\psi_{w,t} - \psi_v)\nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t})\psi'(\phi_{w,t}).$$

Then Lemma 26 and Eq. (70) indicate

$$\frac{k\cos^2\psi_u}{8} \leqslant \operatorname{sgn}(\phi_{w,t} - \phi_v)\nabla_{\phi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi(\phi_{w,t})) \leqslant \frac{5K}{2} \ . \tag{74}$$

For $(\boldsymbol{w}_0, \phi_{w,0}) \in D_{\phi,1}$, we have $\nabla_{\phi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi(\phi_{w,t})) = \nabla_{\psi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi_{w,t})\psi'(\phi_{w,t})$. Then Lemma 27 and Eq. (70) imply

$$|\nabla_{\phi_w} L_{\text{cc}}(\boldsymbol{w}_t, \psi(\phi_{w,t}))| \leqslant 24K \ . \tag{75}$$

Combining Eqs. (73), (74), and (75), Lemma 29 with $r_1 = r_2 = 109R/(100k)$, $a = c_2^2/(4\pi c_1)$, $g_l = k\cos^2\psi_u/8$, and $g_u = 24K$ indicates

$$|\psi_{w,t} - \psi_v| \leqslant \frac{(K+1)c_2^2}{c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{76}$$

Step 4: w_1 converges to 1. Let $(\boldsymbol{w}_0, \phi_{w,0}) \in D_\phi$. It is observed that the reparameterization does not affect the gradient with respect to $w_{t,1}$. Then the convergence of w_1 here is similar to that in Eq. (34), and the only difference is the choice of e, which is related to the upper bound of $|\psi_{w,t} - \psi_v|$. Thus, Lemma 30 with $d_l = 1/4$, $d_u = 1/2$, and $e = 2(K+1)c_2^2/(\pi c_1)$ leads to

$$|w_1 - 1| \leqslant \frac{2(K+1)c_2^3}{\pi c_1 t} \quad \text{for} \quad t \in \mathbb{N}^+ \ . \tag{77}$$

Step 5: the expected loss converges to 0. For any $(\boldsymbol{w}, \phi_w) \in D_\phi$, define non-negative quantities $\Delta_{\boldsymbol{w}} = \|\boldsymbol{w} - \boldsymbol{v}\|$, $\Delta_\psi = |\psi(\phi_w) - \psi(\phi_v)|$, and $\Delta_\phi = |\phi_w - \phi_v|$. According to Eq. (35), one knows that

$$L_{\text{cc}}(\boldsymbol{w}, \psi(\phi_w)) \leqslant 32\Delta_\psi + 5\Delta_{\boldsymbol{w}} \leqslant 27\pi K\Delta_\phi + 5\Delta_{\boldsymbol{w}} \ ,$$

where the second inequality holds from Lagrange's mean value theorem $\Delta_\psi = |\psi'(\xi)|\Delta_\phi$ with $\xi \in [\phi_l, \phi_u]$ and the upper bound of $|\phi'|$ in Eq. (70). Then for $(\boldsymbol{w}_0, \phi_{w,0}) \in D_\phi$, one knows from Eqs. (73), (76), and (77) that

$$L_{\text{cc}}(\boldsymbol{w}, \psi(\phi_w)) \leqslant \frac{27\pi K(K+1)c_2^2}{c_1 t} + \frac{5c_2^2}{4\pi c_1 t} + \frac{10(K+1)c_2^3}{\pi c_1 t} \leqslant \frac{100c_2^3(K+1)^2}{c_1 t} \ . \tag{78}$$

Step 6: initialization falls into D_ϕ with constant probability. Denote by $p(x)$ the probability density function of $\mathcal{N}(0,1)$. Then we have

$$\begin{aligned}
\Pr[(\boldsymbol{w}_0, \phi_{w,0}) \in D_\phi] &= \Pr[\|\boldsymbol{w}_0 - \boldsymbol{v}\|_\infty \leqslant R]\Pr[\phi_{w,0} \in [\phi_l, \phi_u]] \\
&\geqslant \mu(B_\infty(\boldsymbol{v}, R)) \min_{\boldsymbol{w}\in B_\infty(\boldsymbol{v},R)} p(\boldsymbol{w})\ \mu([\phi_l, \phi_u]) \min_{\phi\in[\phi_l,\phi_u]} p(\phi) \\
&\geqslant 4R^2 \cdot (2\pi)^{-1}\mathrm{e}^{-(1+2R+2R^2)/2} \cdot (10\pi K)^{-1} \cdot (2\pi)^{-1/2}\mathrm{e}^{-\psi_u^2/(2k^2)} \\
&\geqslant 10^{-6}K^{-1}\mathrm{e}^{-1/k^2} ,
\end{aligned} \tag{79}$$

where the second inequality holds from $\phi_u - \phi_l \geqslant 6(\psi_u - \psi_l)/(5\pi K) \geqslant (10\pi K)^{-1}$ and $\phi_u \leqslant \psi_u/k$. Combining Eqs. (78) and (79) completes the proof. □

Appendix H. Simulation Experiments

Experimental settings. A training set of size 7,000 and a test set of size 3,000 are generated by a randomly initialized target neuron (can be a real-valued or a complex-valued neuron). After random initialization, a real-valued neuron, a complex-valued neuron, and a reparameterized complex-valued neuron are trained by gradient descent with the empirical mean square loss and a learning rate of 0.1 for 100 epochs.

Notice that the empirical mean square loss is a piecewise constant function of the phase parameters no matter whether reparameterization is used. We use difference quotient, an approximation of the gradient of the expected mean square loss with respect to the phase parameter, as the gradient used in gradient descent, i.e., $\nabla_x L \approx [L(x+h) - L(x-h)]/(2h)$ with $h = 0.1$. If $x - h$ or $x + h$ is beyond the domain of phase, a forward or backward difference is used instead of the central difference.

Experimental results. Notice that complex-valued neurons cannot always learn a target neuron. From the theoretical aspect, our theories holds with a small constant probability. From the loss landscape, there exist constant pieces in the parameter space. Thus, we cannot expect complex-valued neurons to learn a target neuron all the time. In experiments, we train complex-valued neurons with several random initializations and find that our theoretical conclusions occur in experiments. This phenomenon verifies our theories and also motivates a novel learning algorithm for CVNNs, as discussed in the conclusion part.

References

Md Faijul Amin, Ramasamy Savitha, Muhammad Ilias Amin, and Kazuyuki Murase. Complex-valued functional link network design by orthogonal least squares method for function approximation problems. In *Proceedings of 2011 International Joint Conference on Neural Networks*, pages 1489–1496, 2011.

Paolo Arena, Luigi Fortuna, R Re, and Maria Gabriella Xibilia. Multilayer perceptrons to approximate complex valued functions. *International Journal of Neural Systems*, 6(4): 435–446, 1995.

Peter Auer, Mark Herbster, and Manfred K Warmuth. Exponentially many local minima for single neurons. *Advances in Neural Information Processing Systems 8*, pages 316–322, 1995.

Joshua Bassey, Lijun Qian, and Xianfang Li. A survey of complex-valued neural networks. *arXiv:2101.12249*, 2021.

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

Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R Klivans, and Mahdi Soltanolkotabi. Approximation schemes for ReLU regression. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pages 1452–1485, 2020.

Ilias Diakonikolas, Daniel Kane, Lisheng Ren, and Yuxin Sun. SQ lower bounds for learning single neurons with Massart noise. *Advances in Neural Information Processing Systems 35*, pages 24006–24018, 2022.

Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *Advances in Neural Information Processing Systems 33*, pages 5417–5428, 2020.

Quentin Gabot, Jérémy Fix, Joana Frontera-Pons, Chengfang Ren, and Jean-Philippe Ovarlez. Preserving polarimetric properties in PolSAR image reconstruction through complex-valued auto-encoders. In *2024 IEEE Radar Conference*, 2024.

Jingkun Gao, Bin Deng, Yuliang Qin, Hongqiang Wang, and Xiang Li. Enhanced radar imaging using a complex-valued convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 16(1):35–39, 2018.

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

Paul Geuchen, Thomas Jahn, and Hannes Matt. Universal approximation with complex-valued deep narrow neural networks. *arXiv:2305.16910*, 2023.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. In *Proceedings of the 30th Conference on Learning Theory*, pages 1004–1042, 2017.

Nitzan Guberman. On complex valued convolutional neural networks. *arXiv:1602.09046*, 2016.

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

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems 24*, pages 927–935, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. Fitting ReLUs via SGD and quantized SGD. In *Proceedings of 2019 IEEE International Symposium on Information Theory*, pages 2469–2473, 2019.

Taehwan Kim and Tülay Adali. Universal approximation of fully complex feed-forward neural networks. In *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 973–976, 2002.

ChiYan Lee, Hideyuki Hasegawa, and Shangce Gao. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1406–1426, 2022.

Henry Leung and Simon Haykin. The complex backpropagation algorithm. *IEEE Transactions on Signal Processing*, 39(9):2101–2104, 1991.

Timothy J Linhardt, Ananya Sen Gupta, and Ivars Kirsteins. Evaluating the generalization of complex-weight neural networks over simulated Lamb wave responses from hollow spheres. *The Journal of the Acoustical Society of America*, 157(4):2542–2555, 2025.

Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems 32*, pages 6426–6435, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. In *Proceedings of the 34th Conference on Learning Theory*, pages 3265–3295, 2021.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Tohru Nitta. Solving the XOR problem and the detection of symmetry using a single complex-valued neuron. *Neural Networks*, 16(8):1101–1105, 2003.

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

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4433–4441, 2018.

Izhak Shafran, Tom Bagby, and RJ Skerry-Ryan. Complex evolution recurrent neural networks (ceRNNs). In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5854–5858, 2018.

Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in Neural Information Processing Systems 30*, pages 2007–2017, 2017.

Zhi-Hao Tan, Yi Xie, Yuan Jiang, and Zhi-Hua Zhou. Real-valued backpropagation is unsuitable for complex-valued neural networks. *Advances in Neural Information Processing Systems 35*, pages 34052–34063, 2022.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Felix Voigtlaender. The universal approximation theorem for complex-valued neural networks. *Applied and Computational Harmonic Analysis*, 64:33–61, 2023.

Shanshan Wang, Huitao Cheng, Leslie Ying, Taohui Xiao, Ziwen Ke, Hairong Zheng, and Dong Liang. DeepcomplexMRI: Exploiting deep residual network for fast parallel MR imaging with complex convolution. *Magnetic Resonance Imaging*, 68:136–147, 2020.

Jin-Hui Wu, Shao-Qun Zhang, Yuan Jiang, and Zhi-Hua Zhou. Complex-valued neurons can learn more but slower than real-valued neurons via gradient descent. In *Advances in Neural Information Processing Systems 36*, pages 23714–23747, 2023.

Jin-Hui Wu, Shao-Qun Zhang, Yuan Jiang, and Zhi-Hua Zhou. Theoretical exploration of flexible transmitter model. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3674–3688, 2024.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Proccedings of the 33rd Conference on Learning Theory*, pages 3756–3786, 2020.

Junming Zhang and Yan Wu. A new method for automatic sleep stage classification. *IEEE Transactions on Biomedical Circuits and Systems*, 11(5):1097–1110, 2017.

Shao-Qun Zhang and Zhi-Hua Zhou. Flexible transmitter network. *Neural Computation*, 33(11):2951–2970, 2021.

Shao-Qun Zhang, Wei Gao, and Zhi-Hua Zhou. Towards understanding theoretical advantages of complex-reaction networks. *Neural Networks*, 151:80–93, 2022.

Zhimian Zhang, Haipeng Wang, Feng Xu, and Ya-Qiu Jin. Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7177–7188, 2017.

Bo Zhou and Qiankun Song. Boundedness and complete stability of complex-valued neural networks with time delay. *IEEE Transactions on Neural Networks and Learning Systems*, 24(8):1227–1238, 2013.

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