IS THE FREQUENCY PRINCIPLE ALWAYS VALID?

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the learning behaviors of shallow ReLU neural networks on the unit sphere $S^2 \subset \mathbb{R}^3$ in polar coordinates (τ, ϕ), comparing fixed versus trainable neuron weights $\{w_i\}$. Spherical harmonic analysis reveals that both cases exhibit low-frequency preference with coefficients decaying as $O(\ell^{5/2}/2^\ell)$ for fixed weights and $O(\ell^{7/2}/2^\ell)$ for trainable weights, typically leading to the Frequency Principle (FP) of lower-frequency-first learning. However, this principle can be violated under specific initial conditions or error distributions. Our numerical results show that trainable directions increase learning complexity while maintaining the low-frequency advantage, though they can also enable faster high-frequency emergence. This suggests the FP should be viewed as a tendency rather than a strict rule on curved domains like S^2.

1 INTRODUCTION

In recent years, deep learning implemented by deep neural networks (DNNs) has achieved great success in many applications such as computer vision Feng et al. (2019); Xie & Kim (2019); O'Mahony et al. (2020), speech recognition Pan et al. (2012); Deng et al. (2013); Seltzer et al. (2013), translation Singh et al. (2017); Zhang et al. (2015); Yang et al. (2020), and natural language processing Chowdhary & Chowdhary (2020); Nadkarni et al. (2011); Bird et al. (2009); Khurana et al. (2023); Liddy (2001); Eisenstein (2019). Meanwhile, it has become an indispensable method for solving various scientific problems. However, DNNs sometimes fail and generate many issues, such as poor interpretability and limited generalization ability. In theory, DNNs' research has been like exploring a black box for decades. Many researchers compare the theoretical study and practical application of DNNs to alchemy. Therefore, establishing a better theoretical understanding of DNNs has become important.

There has been some progress in the theoretical research of DNNs in recent years. However, we still need to clearly understand how these theoretical results provide key insights and guidance for the practical research of DNNs. Regarding Frequency Principle research, some literature has conducted numerical and theoretical studies on the Frequency Principle, DNNs often fit target functions from low to high frequencies during the training Hong et al. (2022); Xu et al. (2020); Luo et al. (2021); Xu et al. (2019); Rahaman et al. (2019); Ma et al. (2020); Zhang et al. (2023); Choraria et al. (2022); Xu et al. (2024). Through these findings and research, we have gained a basic understanding of the capabilities and limitations of deep learning, namely the difficulty in learning and achieving good generalization for high-frequency data. At the same time, it is relatively simple for low-frequency data. Based on the guiding principle of this frequency, many algorithms have been developed to utilize the low-frequency bias of DNN to fit smooth data well, or to design special techniques/architectures to alleviate the difficulties of DNN in fitting known highly oscillatory data Liu et al. (2020); Jagtap et al. (2020); Tancik et al. (2020); Cai & Xu (2019).

Previous theoretical analyses of the Frequency Principle have focused on neural networks with various activation functions in Euclidean spaces Xu et al. (2020); Luo et al. (2021); Rahaman et al. (2019); Hong et al. (2022). These studies generally support the principle that higher frequencies correspond to slower convergence.

However, our findings reveal that frequency alone cannot fully determine neural network convergence behavior. We demonstrate that the interplay between frequency and training error is critical:

- When training error is significant, high-frequency components can converge faster than low-frequency ones.
- When training error is minimal, low-frequency components typically dominate convergence.

This paper explores the Frequency Principle under two settings: ReLU networks with fixed weights and with trainable weights. We employ spherical harmonic expansions on the unit sphere to provide deeper theoretical insights, revealing conditions where the Frequency Principle fails to hold. Our numerical experiments use Fourier series analysis to illustrate how both frequency and error influence learning behavior.

The main contributions of this article are as follows:

- **Spherical Harmonic Analysis:** We re-examine the Frequency Principle through spherical harmonic expansions, uncovering failure scenarios.
- **Error-Frequency Interplay:** We show that the interaction between frequency and training error can invert learning order.
- **Interpretable Experiments:** We provide numerical evidence demonstrating when and why the Frequency Principle breaks down.

2 ReLU Shallow Neural Networks on the Unit Sphere

We consider a shallow neural network (SNN) with a single hidden layer of m neurons and ReLU activation on the unit sphere $S^2 \subset \mathbb{R}^3$ in polar coordinates (τ, ϕ). The network output is

$$u\big(\tau, \phi; \theta\big) \;=\; \sum_{i=1}^{m} a_i \,\mathrm{ReLU}\big(w_i^\top x(\tau, \phi)\big), \tag{1}$$

where $\theta = \big((a_1, \ldots, a_m)^\top, (w_1, \ldots, w_m)^\top\big)$ are the parameters, $a_i \in \mathbb{R}$ are output weights, and $w_i \in \mathbb{R}^3$ are weight vectors. We normalize $\|w_i\| = 1$ for theoretical simplicity.

The loss function is

$$L(\theta) \;=\; \frac{1}{2} \int_0^{2\pi} \int_0^{\pi} \Big(u\big(\tau, \phi; \theta\big) - h(\tau, \phi)\Big)^2 \,\sin\tau \, d\tau \, d\phi. \tag{2}$$

where $h(\tau, \phi)$ is the target function. Training follows gradient descent:

$$\frac{d\theta}{dt} = -\nabla_\theta L(\theta), \quad \theta(0) = \theta_0, \tag{3}$$

where θ_0 is the given initialization. The error function $D(\tau, \phi) = u(\tau, \phi; \theta) - h(\tau, \phi)$ evolves as

$$D(\tau, \phi) \;=\; u\big(\tau, \phi; \theta\big) \;-\; h(\tau, \phi), \tag{4}$$

and if we integrate over S^2, we have

$$\int_{S^2} D(x)\, dx \;=\; \int_0^{2\pi} \int_0^{\pi} D(\tau, \phi) \;\sin\tau \, d\tau \, d\phi,$$

which shows that the same error function appears in either coordinate system. Since $h(\tau, \phi)$ is fixed, the temporal evolution of $D(\tau, \phi)$ directly reflects changes in θ. By the chain rule,

$$\frac{\partial D(\tau, \phi)}{\partial t} \;=\; \frac{\partial u(\tau, \phi; \theta)}{\partial \theta} \;\cdot\; \frac{d\theta}{dt} \;=\; \frac{\partial u(\tau, \phi; \theta)}{\partial \theta} \;\cdot\; \frac{\partial L(\theta)}{\partial \theta}. \tag{5}$$

This formalizes how, under gradient descent, the error function $D(\tau, \phi)$ evolves over time t. In the following sections, we investigate these dynamics by examining the spherical harmonic expansions of the ReLU activation on the sphere, providing a detailed frequency-level perspective on how the SNN learns $h(\tau, \phi)$.

Then, we introduce the spherical harmonic framework. Spherical harmonics $Y_\ell^j(\tau,\phi)$ form a complete orthonormal basis for functions on S^2, where ℓ represents frequency degree and j represents azimuthal order. Any function can be expanded as $f(\tau,\phi) = \sum_{\ell=0}^{\infty}\sum_{j=-\ell}^{\ell} f_{\ell j}\, Y_\ell^j(\tau,\phi)$. The error evolution in spherical harmonics is $\frac{\partial D(\tau,\phi)}{\partial t} = \sum_{\ell=0}^{\infty}\sum_{j=-\ell}^{\ell} D_{\ell j}\, Y_\ell^j(\tau,\phi)$, where $D_{\ell j}$ are the harmonic coefficients. For detailed introduction of spherical harmonic, please refer to Appendix A.1.

At this point, we can define what the Frequency Principle is. For fixed azimuthal order j and frequency degrees $\ell_1 < \ell_2$, the network adheres to the Frequency Principle if both $D_{\ell_1 j}$ and $D_{\ell_2 j}$ are negative, and

$$|D_{\ell_1 j}| > |D_{\ell_2 j}|, \tag{6}$$

indicating faster error reduction in lower-frequency modes. A violation occurs when

$$|D_{\ell_1 j}| \leq |D_{\ell_2 j}|, \tag{7}$$

where at this point,it does not matter $D_{\ell_1 j}$ and $D_{\ell_2 j}$ are positive or negative, suggesting comparable or faster learning of higher-frequency components.

3 Frequency Analysis on the Unit Sphere for Fixed Weights

We analyze the frequency behavior of shallow neural networks on the unit sphere S^2 with fixed weight vectors w_i. This setting provides a baseline analysis of how ReLU activation inherently shapes frequency representation through spherical harmonics.

For a single neuron with fixed weights, the ReLU activation $\mathrm{ReLU}(w_i^\top x(\tau,\phi))$ is axisymmetric and expands in spherical harmonics as

$$\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \;=\; \sum_{\ell=0}^{\infty} c_\ell\, Y_\ell^0(\tau,\phi), \tag{8}$$

where the coefficients $c_\ell = \frac{\sqrt{\pi}}{24}\left(\frac{1}{2}\right)^\ell \sqrt{2\ell+1}\,\left(\ell^2+3\ell+2\right)$, while for $\ell=0$, $c_0=\frac{\pi}{2}$, and for $\ell=1$, $c_1=\frac{\sqrt{3\pi}}{3}$. We have $c_\ell = O\Big(\frac{\ell^{5/2}}{2^\ell}\Big)$ for large ℓ, indicating rapid high-frequency decay. See Appendix A.2 for detailed calculations.

For the full network, the harmonic coefficients become linear combinations of $\{c_\ell\}$, maintaining the low-frequency bias.

then

$$u_{\ell j}(\theta) \;=\; \int_{S^2} u(\tau,\phi;\theta)\,\overline{Y_\ell^j(\tau,\phi)}\; d\Omega \;=\; \sum_{i=1}^{m} a_i\; c_\ell\,\delta_{j0}.$$

Hence, the low-frequency bias in each single neuron's expansion (i.e., the decay of c_ℓ with ℓ) underpins the frequency behavior of the entire network on the sphere. Analyzing these coefficients $\{c_\ell\}$ thus reveals how the SNN distributes energy across spherical frequencies, providing insight into phenomena like the Frequency Principle.

When only output-layer coefficients $\theta = (a_1,\ldots,a_m)^\top$ are trained, the error evolution follows

$$\frac{\partial D(\tau,\phi)}{\partial t} \;=\; -\sum_{i=1}^{m} \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)\int_{S^2} D(\tau',\phi')\,\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\, d\Omega. \tag{9}$$

See Appendix A.3 for detailed derivation.

Hence equation 9 captures the interplay of the fixed directions, the ReLU activation, and the error distribution $D(\tau,\phi)$ in governing the time evolution of the error under gradient descent.

The error evolution in spherical harmonics becomes

$$\frac{\partial D(\tau,\phi)}{\partial t} \;=\; -\sum_{\ell=0}^{\infty}\left[m\Big(\tfrac{2\pi}{3}\sum_{k=1}^{m} a_k - C(h)\Big)\, c_\ell\right] Y_\ell^0(\tau,\phi), \tag{10}$$

where $C(h) = \int_0^{2\pi}\int_0^{\frac{\pi}{2}} h(\tau', \phi')\, \cos(\tau')\, \sin(\tau')\, d\tau'\, d\phi'$ and

$$D_\ell \;:=\; m\Big(\tfrac{2\pi}{3}\sum_{k=1}^{m} a_k - C(h)\Big)\, c_\ell. \tag{11}$$

Each mode ℓ is updated by D_ℓ, and the Frequency Principle holds when lower-frequency modes decay faster than higher ones.

Theorem 3.1 (Special Case for FP Violation on the Sphere) *Let $\{w_i\}_{i=1}^m \subset S^2$ be fixed unit vectors and align along the polar axis, and consider the shallow ReLU network equation 1. Let $D(\tau, \phi)$ defined in equation 4 be the error function on the sphere. Suppose that at any training time t, if*

$$\sum_{k=1}^{m} a_k = \frac{3}{2\pi} C(h)$$

with $C(h) = \int_0^{2\pi}\int_0^{\frac{\pi}{2}} h(\tau', \phi')\, \cos(\tau')\, \sin(\tau')\, d\tau'\, d\phi'$, then

$$D_\ell \;=\; 0 \quad \textit{for every } \ell.$$

Recalling that $\big|c_\ell\big| \approx O\big(\ell^{5/2}/2^\ell\big)$, we deduce that higher-frequency components of $\mathrm{ReLU}(w_i^\top x)$ decay exponentially in ℓ. Depending on how $\frac{2\pi}{3}\sum_{k=1}^m a_k - C(h)$ scales with ℓ, the SNN can either maintain or lose low-frequency dominance. Below are two theorems echoing that trade-off.

Theorem 3.2 (General Conditions for FP Violation on the Sphere) *Let $\{w_i\}_{i=1}^m \subset S^2$ be fixed unit vectors and align along the polar axis, and consider the shallow ReLU network equation 1. Let $D(\tau, \phi)$ defined in equation 4 be the error function on the sphere. Suppose that at any training time t, if*

$$\sum_{k=1}^{m} a_k < \frac{3}{2\pi} C(h)$$

with $C(h) = \int_0^{2\pi}\int_0^{\frac{\pi}{2}} h(\tau', \phi')\, \cos(\tau')\, \sin(\tau')\, d\tau'\, d\phi'$, then for all ℓ, the frequency principle can be violated*. In other words, the exponential decay of c_ℓ can be surmounted by such error terms at sufficiently low frequencies, enabling high-frequency components to persist.*

Theorem 3.3 (Conditions for FP Preservation on the Sphere) *Under the similar setting and notation as in Theorem 3.2, but*

$$\sum_{k=1}^{m} a_k > \frac{3}{2\pi} C(h),$$

the frequency principle holds *for all ℓ and the error terms $D\big(\tau_i, \phi_i\big)$ remain of the same symbolic order. That is, for higher modes ℓ, the exponential decay of c_ℓ cannot be balanced by constant-order (or moderately growing) error terms, thereby causing the high-frequency components to diminish naturally.*

The proof and some remarks of Theorem 3.2 and 3.3 is given in Appendix A.4.

Hence, in summary, once the exponential decay of c_ℓ is accounted for, only a sufficiently large or fast-growing error feedback $\big(\frac{2\pi}{3}\sum_{k=1}^m a_k - C(h)\big)$ can sustain high-frequency modes. Otherwise, the FP is preserved, manifesting as a bias toward learning low-frequency harmonics more rapidly than high-frequency ones.

3.1 Numerical Examples of Fixed Weights

We present numerical experiments to illustrate how shallow networks on the sphere adhere to or deviate from the Frequency Principle. We transform spherical coordinates (τ, ϕ) to Cartesian coordinates $\mathbf{x} = (\sin(\tau)\cos(\phi), \sin(\tau)\sin(\phi), \cos(\tau))$ for neural network compatibility.

The network architecture consists of: Input layer for spherical coordinates, hidden layer of size h with ReLU activation, and linear output layer producing scalar u_1.

The model is defined as:

$$\mathbf{z}_1 = \mathrm{ReLU}\big(W_1\,\mathbf{x}(\tau,\phi) + b_1\big), \quad u_1 = W_2\,\mathbf{z}_1,$$

where W_1 is fixed to 1, b_1 is fixed with uniform distribution on $[-1, 1]$, and W_2 is trainable.

We minimize a weighted MSE loss incorporating spherical geometry:

$$\mathrm{Loss} = \frac{1}{N}\sum_{i=1}^{N}\Big(u(\tau_i,\phi_i,\theta) - h(\tau_i,\phi_i)\Big)^2 \sin(\tau_i).$$

To analyze frequency component evolution, we compute spherical harmonic coefficients c_{lj} using:

$$c_{lj} = \int_0^{\pi}\int_0^{2\pi} h(\tau,\phi)\; Y_l^{j^*}(\tau,\phi)\;\sin(\tau)\;d\phi\,d\tau,$$

where Y_l^j are spherical harmonics and $Y_l^{j^*}$ denotes complex conjugate.

During training, we monitor: Training loss (global MSE over spherical dataset) and spherical harmonic coefficient errors: $|c_{lj,\mathrm{pred}} - c_{lj,\mathrm{true}}|$ for degrees up to $l_{\max}$.

3.1.1 $h(\tau,\phi) = 0$

Target function: $h(\tau,\phi) = 0$ on unit sphere $\mathbb{S}^2$. Network: 100 hidden neurons, ReLU activation. Training: 100 sampling points, SGD with learning rate $1e{-}3$, 10,000 epochs.

Case 1: Complete Adherence to Frequency Principle

The network strictly follows the frequency principle, learning low-frequency components before higher frequencies. Loss reaches 6.16e-6 with uniform error distribution across the sphere. See Figure 2 for training loss, Figure 3 for harmonic errors, and Figure 4 for target function, network output, and error distribution.

Case 2: Partial Adherence to Frequency Principle

Mixed learning pattern with loss 1.59e-5. Some higher-frequency components learned simultaneously with low-frequency ones, showing non-monotonic harmonic error patterns. See Figure 5 for training loss, Figure 6 for harmonic errors, and Figure 7 for target function, network output, and error distribution.

Case 3: Contradiction to Frequency Principle

With high-frequency initialization (Figure 8), the network contradicts the frequency principle, learning higher frequencies before lower ones. Loss reaches 6.79e-4 with significant error variations across the sphere. See Figure 9 for training loss, Figure 10 for harmonic errors, and Figure 11 for target function, network output, and error distribution.

Comparative Analysis The three cases demonstrate diverse learning behaviors. The frequency principle serves as a general tendency rather than a universal law, influenced by network initialization, optimization trajectory, and architecture. Detailed analysis and discussions are provided in Appendix A.5.

3.1.2 $h(\tau,\phi) = \sin(\tau)\cos(3\phi) + \sin(3\tau)\cos(5\phi)$

Target function combines multiple frequency components. Network: 100 hidden neurons, ReLU activation. Training: 100 sampling points, SGD with learning rate $1e{-}3$, 100,000 epochs.

Case 1: Partial Adherence to Frequency Principle

Complex learning dynamics with simultaneous learning of multiple frequency components. The natural coupling influences the network's learning trajectory, leading to partial adherence to the frequency principle. See Figure 12 for training loss, Figure 13 for harmonic errors, and Figure 14 for spatial comparison.

Case 2: Contradiction to Frequency Principle

With high-frequency initialization, the network achieves loss 9.1e-4 while contradicting the frequency principle. Higher-frequency components are learned before lower-frequency ones, demonstrating clear violation of traditional frequency learning patterns. See Figure 15 for training loss, Figure 16 for harmonic errors, and Figure 17 for spatial comparison.

Comparative Analysis The analysis reveals that coupled frequency components introduce complex learning dynamics that can disrupt the frequency principle. Network initialization plays a crucial role in determining learning behavior. The frequency principle may not fully capture learning dynamics for functions with intricate frequency interactions. Comprehensive analysis and theoretical implications are detailed in Appendix A.6.

4 Frequency Analysis on the Unit Sphere with Trained Weights

In this section, we analyze the frequency behavior of shallow neural networks (SNNs) defined on the unit sphere S^2 when both the weights w_i and the coefficients a_i can change during training. The unit sphere provides a natural domain for spherical harmonics, allowing us to investigate the frequency components of ReLU-activated functions.

We consider SNNs of the form equation 1, where $\theta = ((a_i)_{i=1}^m, (w_i)_{i=1}^m)$ represents both the output weights $a_i \in \mathbb{R}$ and the direction vectors $w_i \in S^2$. Unlike the fixed-weight scenario, here we allow w_i to evolve during training, introducing complexity as changes in w_i can alter the spherical harmonic representation dynamically.

When w_i varies during training, we fix a global coordinate system on the sphere, described by (τ, ϕ). Let R_{w_i} be the rotation that takes the global frame's north pole to align with w_i. For $w_i = (w_x, w_y, w_z)$ on the unit sphere S^2, we can explicitly construct R_{w_i} using Rodrigues' rotation formula: $R_{w_i} = I + \sin\theta K + (1 - \cos\theta)K^2$, where $\theta = \arccos(w_z)$ is the angle between w_i and the north pole $(0, 0, 1)$, and K is the skew-symmetric matrix formed from the normalized axis of rotation $k = (-w_y, w_x, 0)/\sqrt{w_x^2 + w_y^2}$.

Then any point $x(\tau, \phi)$ can be mapped into the local axis-aligned frame via $\tilde{x}(\tau, \phi) = R_{w_i}^{-1}\, x(\tau, \phi)$.

Via the Wigner D-matrices, we have $Y_\ell^0\big(\tilde{x}(\tau, \phi)\big) = \sum_{j=-\ell}^{\ell} D_{j0}^{(\ell)}\big(R_{w_i}\big)\, Y_\ell^j\big(x(\tau, \phi)\big)$, so we have

$$\mathrm{ReLU}\big(w_i^\top x(\tau, \phi)\big) = \sum_{\ell=0}^{\infty} \sum_{j=-\ell}^{\ell} \Big[c_\ell D_{j0}^{(\ell)}\big(R_{w_i}\big)\, Y_\ell^j(\tau, \phi) \Big]. \tag{12}$$

The training implications are threefold. **First, invariant local structure:** In the coordinate system where w_i defines the polar axis, the activation function exhibits axial symmetry, yielding a spherical harmonic expansion containing only $j = 0$ terms. **Second, global mode coupling:** When expressed in fixed global coordinates, the local $j = 0$ mode couples to multiple global angular momentum states. **Third, dynamic frequency control:** Through optimization of $w_i \in S^2$, the network repositions the axisymmetric ReLU activation pattern across the sphere, enabling selective amplification of specific global harmonics.

Consider the shallow neural network equation 1. Substituting the single-neuron expansions into $u(\tau, \phi; \theta)$, we obtain:

$$u\big(\tau, \phi; \theta\big) = \sum_{i=1}^{m} a_i \sum_{\ell=0}^{\infty} c_\ell \sum_{j=-\ell}^{\ell} D_{j0}^{(\ell)}\big(R_{w_i}\big)\, Y_\ell^j(\tau, \phi) = \sum_{\ell=0}^{\infty} \sum_{j=-\ell}^{\ell} \Big[\sum_{i=1}^{m} \big(a_i\, c_\ell\, D_{j0}^{(\ell)}(R_{w_i})\big) \Big]\, Y_\ell^j(\tau, \phi).$$

Hence, the coefficient in front of each global harmonic $Y_\ell^j(\tau, \phi)$ is $\sum_{i=1}^{m} a_i\, c_\ell\, D_{j0}^{(\ell)}(R_{w_i})$. Because $c_\ell \approx O(\ell^{5/2}/2^\ell)$ indicates rapid decay for large ℓ in the local axis frame, the network inherits a natural bias toward lower-frequency modes. However, the rotation terms can evolve during training, altering how that local bias manifests globally.

When each w_i is free to move on S^2 during training, the local axis alignment changes over time. This dynamic remapping operates through two complementary mechanisms. **Each neuron maintains**

its intrinsic low-frequency bias: In any local coordinate system aligned with w_i, the expansion heavily favors small ℓ values. **However, collective rotational effects enable global frequency redistribution:** By updating each w_i, the network can redirect how that low-frequency local pattern is projected onto the global $\{\ell, j\}$ modes.

We now analyze the case in which both the output-layer coefficients $\{a_i\}$ and the directions $\{w_i\} \subset S^2$ evolve under gradient descent. To track how $D(\tau, \phi)$ changes over time, note that $\frac{\partial D(\tau,\phi)}{\partial t} = \frac{\partial}{\partial t}\big[u(\tau, \phi; \theta) - h(\tau, \phi)\big] = \frac{\partial u(\tau,\phi;\theta)}{\partial \theta} \cdot \frac{d\theta}{dt}$.

Since θ now includes both a_i and w_i, we partition:

$$\frac{\partial D(\tau,\phi)}{\partial t} = \sum_{i=1}^{m} \Big[\underbrace{\frac{\partial u(\tau,\phi;\theta)}{\partial a_i} \, \frac{\partial L(\theta)}{\partial a_i}}_{\text{update in } a_i} + \underbrace{\frac{\partial u(\tau,\phi;\theta)}{\partial w_i} \, \frac{\partial L(\theta)}{\partial w_i}}_{\text{update in } w_i} \Big].$$

Recall $\frac{\partial u(\tau,\phi;\theta)}{\partial a_i} = \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$, $\frac{\partial u(\tau,\phi;\theta)}{\partial w_i} = a_i \, \nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$. The loss gradients are $\frac{\partial L(\theta)}{\partial a_i} = -\int_0^{2\pi}\!\int_0^{\pi} D(\tau',\phi') \, \mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big) \, \sin\tau' \, d\tau' \, d\phi'$, and $\frac{\partial L(\theta)}{\partial w_i} = -\int_0^{2\pi}\!\int_0^{\pi} D(\tau',\phi') \, a_i \, \nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big) \, \sin\tau' \, d\tau' \, d\phi'$.

Hence,

$$\begin{aligned}\frac{\partial D(\tau,\phi)}{\partial t} = -\sum_{i=1}^{m} \Big\{ & \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \int_{S^2} D(\tau',\phi') \, \mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big) \, d\Omega \\ & + \, a_i^2 \, \nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \int_{S^2} D(\tau',\phi') \, \nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big) \, d\Omega \Big\}.\end{aligned}$$

Observe that

$$\nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x\big) = \begin{cases} x(\tau,\phi), & \text{if } w_i^\top x(\tau,\phi) > 0, \\ 0, & \text{if } w_i^\top x(\tau,\phi) \le 0. \end{cases}$$

The term $\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$ affects how a_i updates, whereas $\nabla_{w_i} \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$ affects how w_i itself rotates or shifts.

When both the output coefficients $\{a_i\}$ and the directions $\{w_i\} \subset S^2$ evolve under gradient descent, the time evolution of the error becomes more intricate due to the additional directional updates. From Section 3, the partial derivative of $D(\tau, \phi)$ under gradient descent is

$$\begin{aligned}\frac{\partial D(\tau,\phi)}{\partial t} = -\sum_{i=1}^{m} \Big(& \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \; \underbrace{\pi \, D\big(\tau_i^*, \phi_i^*\big) \, \sin\big(2\tau_i^*\big)}_{\text{from } \int D(y) \, \mathrm{ReLU}(w_i^\top y) \, dy} \\ & + \, a_i^2 \, \nabla_{w_i} \, \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \quad \underbrace{\pi \, D\big(\tau_i^\dagger, \phi_i^\dagger\big) \, \bar{y}_i}_{\text{from } \int D(y) \, \nabla_{w_i} \mathrm{ReLU}(w_i^\top y) \, dy} \Big),\end{aligned}$$

where the points $\big(\tau_i^*, \phi_i^*\big)$ and $\big(\tau_i^\dagger, \phi_i^\dagger\big)$ arise from a 2D Mean Value Theorem argument over the hemisphere $\{(\tau', \phi') : w_i^\top x(\tau', \phi') \ge 0\}$.

Recalling the global expansion of each neuron's ReLU term from equation 12, we substitute this into the gradient-based update for $\frac{\partial D}{\partial t}$. Consequently, we collect the resulting terms into spherical harmonic modes $Y_\ell^j(\tau, \phi)$. Define the direct neuron-output contribution:

$$C_\ell^j := \sum_{i=1}^{m} \Big[\pi \, D\big(\tau_i^*, \phi_i^*\big) \, \sin\big(2\,\tau_i^*\big) \Big] \, c_\ell \, D_{j0}^{(\ell)}\big(R_{w_i}\big), \tag{13}$$

and the weight rotation contribution:

$$G_\ell^j := \sum_{i=1}^{m} a_i^2 \Big[\pi \, D\big(\tau_i^\dagger, \phi_i^\dagger\big) \, \bar{y}_i \Big] \cdot \nabla_{w_i} \Big(c_\ell \, D_{j0}^{(\ell)}\big(R_{w_i}\big) \Big). \tag{14}$$

The detailed calculation of $\nabla_{w_i}\left(c_\ell D_{j0}^{(\ell)}(R_{w_i})\right)$ involves complex derivatives with respect to rotation angles and is provided in Appendix A.7. The key result is that this gradient term introduces an additional polynomial factor in the frequency decay.

Finally, collecting all terms, the error's spherical harmonic components evolve according to

$$\frac{\partial D(x)}{\partial t} = -\sum_{\ell=0}^{\infty}\sum_{j=-\ell}^{\ell}\left(C_\ell^j + G_\ell^j\right) Y_\ell^j(x). \tag{15}$$

This shows how both the direct amplitude updates (via C_ℓ^j) and the rotation effects (via G_ℓ^j) shape the frequency distribution in the trained-weights regime.

Unlike the fixed-weight setting, each coefficient G_ℓ^j now provides a mechanism for rotating w_i to activate different regions of S^2. As w_i changes, the rotation R_{w_i} continuously re-mixes the local axisymmetric modes across the global j modes. This adds substantial flexibility, allowing higher-frequency components to be selectively enhanced or suppressed based on the error distribution $D(\tau,\phi)$ over time.

Having established the updated expressions for C_ℓ^j and G_ℓ^j, we now turn to analyze how they shape the Frequency Principle in the trained-weights setting. Below, we present a series of theorems and corollaries that clarify when and why low-frequency dominance is preserved, and under what circumstances it may fail.

Lemma 4.1 (Decay Rates of Error Evolution Terms) *Consider the spherical harmonic coefficients C_ℓ^j and G_ℓ^j in the error evolution equation equation 15, where C_ℓ^j and G_ℓ^j are defined by equation 13 and equation 14, respectively. Under the assumption that the geometric/error factors remain bound, the following estimates hold:*

$$\left|C_\ell^j\right| = O\Big(\frac{\ell^{5/2}}{2^\ell}\Big), \quad \left|G_\ell^j\right| = O\Big(\frac{\ell^{7/2}}{2^\ell}\Big).$$

Theorem 4.1 (Immediate FP Violation in the Trained-Weights Scenario) *Let $\{w_i\}_{i=1}^m \subset S^2$ be directions updated by gradient descent and $\{a_i\}$ be scalar output weights. Consider the shallow ReLU network $u(\tau,\phi;\theta) = \sum_{i=1}^m a_i \,\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$ on the sphere, and define the error function $D(\tau,\phi) = u(\tau,\phi;\theta) - h(\tau,\phi)$. Let (τ_i^*,ϕ_i^*) and $(\tau_i^\dagger,\phi_i^\dagger)$ be the representative points for the integrals. From equation 13 and equation 14,* if at any training time t,

$$D\big(\tau_i^*,\phi_i^*\big) = 0 \quad \textit{and} \quad D\big(\tau_i^\dagger,\phi_i^\dagger\big) = 0 \quad \forall\, i = 1,\dots,m,$$

then

$$C_\ell^j = 0 \quad \textit{and} \quad G_\ell^j = 0 \quad \textit{for every } \ell \in \{0,1,2,\dots\},\ j = -\ell,\dots,\ell.$$

The proof of Lemma 4.1 and some remarks of Theorem 4.1 is given in the Appendix A.8.

In summary, these results clarify how weight rotation adds a polynomial factor ℓ to the exponential decay but does not remove it. For very large ℓ, both C_ℓ^j and G_ℓ^j (shown in Figure 18) remain exponentially small in ℓ, so the FP is preserved. However, at lower or moderate frequencies, geometric effects in G_ℓ^j can become significant, potentially allowing certain modes to increase faster, leading to FP violations. This framework provides a detailed account of when (and how) trained weights alter frequency-dependent learning dynamics on the sphere.

4.1 Numerical Examples of Trained Weights

We employ a shallow neural network with 100 hidden neurons to approximate functions on $\mathbb{S}^2$. The network is trained using SGD with learning rate $1e{-}3$ over 10,000 epochs for $u(\tau,\phi) = 0$ and 100,000 epochs for $u(\tau,\phi) = \sin(\tau)\cos(3\phi) + \sin(3\tau)\cos(5\phi)$. Training data consists of 100 uniformly distributed sampling points on the sphere.

4.1.1 $h(\tau, \phi) = 0$

Case 1: All Frequencies follow the Frequency Principle See Figure 19 for loss evolution and Figures 20-21 for harmonic errors. Figure 22 shows target function, network output, and error distribution.

In this case, trained weights exhibit direct correlation with the frequency principle. The weight distribution shows hierarchical structure favoring lower-frequency components, with systematic decay in weight magnitudes as frequency increases. This results in optimal approximation properties with uniform error distribution across the sphere.

Case 2: Partial Adherence to Frequency Principle See Figure 23 for loss evolution and Figures 24-25 for harmonic errors. Figure 26 shows target function, network output, and error distribution.

This case presents mixed frequency characteristics with non-monotonic weight distribution and simultaneous activation of low and high-frequency modes. This demonstrates partial adherence to the frequency principle.

Case 3: All Frequencies violate the Frequency Principle With high-frequency initialization (Figure 8), the network contradicts the frequency principle, learning higher frequencies before lower ones. Loss reaches 4.45e-4 with significant error variations. See Figure 27 for loss evolution and Figures 28-29 for harmonic errors. Figure 30 shows target function, network output, and error distribution.

This case presents that initial weights significantly influence the final frequency learning pattern.

4.1.2 $h(\tau, \phi) = \sin(\tau)\cos(3\phi) + \sin(3\tau)\cos(5\phi)$

We examine learning dynamics when both weights and biases are trained simultaneously. The network architecture employs 100 hidden neurons with ReLU activation, trained using SGD with learning rate $1e{-}3$ over 10,000 epochs.

Case 1: Partial Adherence to Frequency Principle See Figure 31 for loss evolution and Figures 32-33 for harmonic errors. Figure 34 shows spatial comparison.

With trainable weights, the network achieves final loss of 3.2e-3, showing improved convergence. Harmonic error analysis reveals interleaved frequency learning where high-frequency components are learned concurrently with low-frequency counterparts. The spatial error distribution exhibits more uniform structure, indicating better balance in frequency component approximation.

Case 2: Contradiction to Frequency Principle See Figure 35 for loss evolution and Figures 36-37 for harmonic errors. Figure 38 shows spatial comparison.

With high-frequency initialization, the network shows pronounced contradiction to the frequency principle, reaching final loss of 8.4e-4. The network prioritizes learning high-frequency components before addressing lower-frequency terms. Spatial error distribution shows concentrated errors in low-frequency dominated regions while high-frequency features are captured accurately.

5 Conclusion and Future Work

In this paper, we examined shallow ReLU networks defined on the unit sphere using spherical harmonic expansions. In fixed-weight settings, ReLU neurons exhibit axisymmetry around weight vectors, explaining exponential decay in spherical harmonic coefficients and natural prioritization of low-frequency modes. But there also a caesthat high-frequency components can converge rapidly when substantial errors exist in those regions. For fully trainable networks, we discovered *weight rotation* as an additional mechanism. Changing neuron directions reorients local axisymmetry globally, allowing selective amplification or suppression of higher frequencies based on gradient flows and error distribution. Our experiments confirm that both full and partial violations of the Frequency Principle can occur, especially with deliberate high-frequency perturbations or initialization favoring higher-frequency modes.

Future work may explore deeper architectures, other activation functions, and strategies for mitigating adverse high-frequency effects.

ACKNOWLEDGMENTS

Authors would like to express their sincere gratitude to Professor Jinchao Xu for his valuable suggestions and kind invitation to KAUST, which made this research possible. We are also grateful to Research Assistant Jongho Park for insightful discussions that contributed significantly to the development of this work. We acknowledge the use of ChatGPT for assistance in manuscript preparation and editing.

REFERENCES

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Wei Cai and Zhi-Qin John Xu. Multi-scale deep neural networks for solving high dimensional pdes. *arXiv preprint arXiv:1910.11710*, 2019.

Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks. *arXiv preprint arXiv:2202.13473*, 2022.

KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8599–8603. IEEE, 2013.

Jacob Eisenstein. *Introduction to natural language processing*. MIT press, 2019.

Xin Feng, Youni Jiang, Xuejiao Yang, Ming Du, and Xin Li. Computer vision algorithms and hardware implementations: A survey. *Integration*, 69:309–320, 2019.

Qingguo Hong, Jonathan W. Siegel, Qinyang Tan, and Jinchao Xu. On the activation function dependence of the spectral bias of neural networks. *arXiv preprint arXiv:2208.04924*, 2022.

Ameya D Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

Elizabeth D Liddy. Natural language processing. 2001.

Ziqi Liu, Wei Cai, and Zhi-Qin John Xu. Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains. *arXiv preprint arXiv:2007.11207*, 2020.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *CSIAM Trans. Appl. Math.*, 2(3):484–507, 2021.

Chao Ma, Lei Wu, et al. Machine learning from a continuous viewpoint, i. *Science China Mathematics*, 63(11):2233–2266, 2020.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 128–144. Springer, 2020.

Jia Pan, Cong Liu, Zhiguo Wang, Yu Hu, and Hui Jiang. Investigation of deep neural networks (dnn) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic modeling. In *2012 8th International Symposium on Chinese Spoken Language Processing*, pp. 301–305. IEEE, 2012.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019.

Michael L Seltzer, Dong Yu, and Yongqiang Wang. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 7398–7402. IEEE, 2013.

Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pp. 162–167. IEEE, 2017.

Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.

Xiufeng Xie and Kyu-Han Kim. Source compression with bounded dnn perception loss for iot edge computer vision. In *The 25th Annual International Conference on Mobile Computing and Networking*, pp. 1–16, 2019.

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing. ICONIP 2019.*, volume 11953 of *Lecture Notes in Computer Science*, pp. 264–274. Springer, Cham, 2019.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Commun. Comput. Phys.*, 28(5):1746–1767, 2020.

Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, pp. 1–38, 2024.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020.

Jiajun Zhang, Chengqing Zong, et al. Deep neural networks in machine translation: An overview. *IEEE Intell. Syst.*, 30(5):16–25, 2015.

Shijun Zhang, Hongkai Zhao, Yimin Zhong, and Haomin Zhou. Why shallow networks struggle with approximating and learning high frequency: A numerical study. *arXiv preprint arXiv:2306.17301*, 2023.

A Appendix

A.1 Detailed Spherical Harmonic Definitions

The spherical harmonics $Y_\ell^j(\tau,\phi)$ in polar coordinates are defined by

$$Y_\ell^j(\tau,\phi) \;=\; N_\ell^j\, P_\ell^j\big(\cos\tau\big)\, e^{\,i\,j\,\phi}, \quad \ell \geq 0, \quad -\ell \leq j \leq \ell, \tag{16}$$

where

$$N_\ell^j \;=\; \sqrt{\frac{(2\ell+1)\,(\ell-j)!}{4\pi\,(\ell+j)!}} \tag{17}$$

is the normalization constant, and

$$P_\ell^j(x) \;=\; (-1)^j\,(1-x^2)^{\frac{j}{2}}\,\frac{d^j}{dx^j}\,P_\ell(x), \quad x\in\mathbb{R}$$

are the associated Legendre polynomials.

The spherical harmonics form a complete orthonormal basis with:

1. *Orthogonality:* For any $\ell,\ell'\geq 0$ and $-\ell\leq j\leq \ell$, $-\ell'\leq j'\leq \ell'$,

$$\int_0^{2\pi}\!\!\int_0^{\pi} Y_\ell^j(\tau,\phi)\,\overline{Y_{\ell'}^{j'}(\tau,\phi)}\;\sin\tau\,d\tau\,d\phi \;=\; \delta_{\ell\ell'}\,\delta_{jj'}, \tag{18}$$

2. *Completeness:* Any square-integrable function admits an expansion with coefficients computed by

$$f_{\ell j} \;=\; \int_0^{2\pi}\!\!\int_0^{\pi} f(\tau,\phi)\,\overline{Y_\ell^j(\tau,\phi)}\;\sin\tau\,d\tau\,d\phi. \tag{19}$$

A.2 Detailed Single Neuron Expansion

Having established the basic spherical harmonic machinery, we now analyze the simpler case of a *single neuron* with ReLU activation:

$$\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big),$$

where $w_i\in\mathbb{R}^3$ is fixed (i.e., not trained), and $x(\tau,\phi)\in S^2$ is the input in polar coordinates. To study its spherical harmonic expansion, we first interpret $w_i^\top x(\tau,\phi)$ geometrically.

Since $\|x(\tau,\phi)\|=1$ and $\|w_i\|=1$, the quantity $w_i^\top x(\tau,\phi)$ is

$$w_i^\top x(\tau,\phi) \;=\; \cos(\tau'),$$

where $\tau'\in[0,\pi]$ is the angle between w_i and $x(\tau,\phi)$. By choosing a rotated coordinate system so that w_i coincides with the north pole (0,0,1), this angle τ' can be identified with the polar angle. Thus

$$\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \;=\; \max\big\{0,\,\cos(\tau')\big\}. \tag{20}$$

In particular, $\cos(\tau')$ is nonnegative only if $\tau'\in[0,\frac{\pi}{2}]$. This means the neuron output is positive (active) in a polar cap region about w_i and zero elsewhere, partitioning S^2 into *active* and *inactive* zones.

Because $\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$ depends only on τ' (and not on the azimuth ϕ), it is an *axisymmetric* function on the sphere. Consequently, its spherical harmonic expansion will involve only the $j=0$ modes:

$$\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big) \;=\; \sum_{\ell=0}^{\infty} c_\ell\,Y_\ell^0(\tau,\phi), \tag{21}$$

where each coefficient c_ℓ is given by an inner product with $Y_\ell^0(\tau,\phi)$. In polar coordinates,

$$Y_\ell^0(\tau,\phi) \;=\; \sqrt{\frac{2\ell+1}{4\pi}}\;P_\ell\big(\cos\tau\big).$$

Since $\mathrm{ReLU}(\cos\tau')$ is nonzero only for $\tau'\in[0,\frac{\pi}{2}]$, we may express

$$c_\ell \;=\; \int_{S^2}\mathrm{ReLU}\big(\cos\tau'\big)\,\overline{Y_\ell^0(\tau',\phi')}\;d\Omega \;=\; 2\pi\sqrt{\frac{2\ell+1}{4\pi}}\int_0^1 t\,P_\ell(t)\,dt,$$

where $t=\cos\tau'\in[0,1]$ and $d\Omega=\sin(\tau')\,d(\tau')\,d\phi'$.

Carrying out the integral carefully, one finds for $\ell\geq 2$Bach (2017):

$$c_\ell \;=\; \frac{\sqrt{\pi}}{24}\big(\tfrac{1}{2}\big)^\ell\sqrt{2\ell+1}\,\big(\ell^2+3\ell+2\big),$$

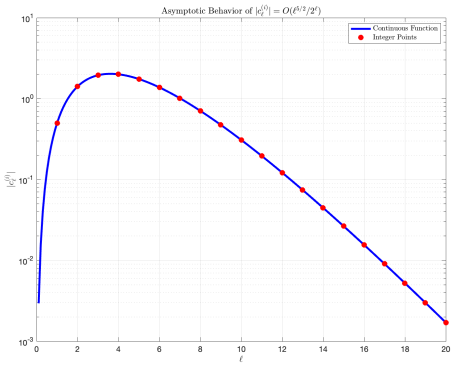


Figure 1: Asymptotic behavior of $|c_\ell|$. Note the polynomial factor $\ell^{5/2}$ is dominated by the exponential decay $2^{-\ell}$.

while for $\ell = 0$, $c_0 = \frac{\pi}{2}$, and for $\ell = 1$, $c_1 = \frac{\sqrt{3\pi}}{3}$. In general, these coefficients satisfy

$$c_\ell \;=\; O\Big(\frac{\ell^{5/2}}{2^\ell}\Big) \quad \text{as } \ell \to \infty,$$

indicating that $\mathrm{ReLU}(\cos\tau')$ has significant low-frequency components but rapidly decaying high-frequency components.

For an SNN of the form equation 1, each $\mathrm{ReLU}(w_i^\top x(\tau,\phi))$ can be expanded similarly. The overall harmonic coefficient in the $j = 0$ mode becomes a linear combination of $\{c_\ell\}$. Specifically, if

$$u\big(\tau,\phi;\theta\big) \;=\; \sum_{i=1}^{m} a_i \,\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big),$$

then

$$u_{\ell j}(\theta) \;=\; \int_{S^2} u(\tau,\phi;\theta)\,\overline{Y_\ell^j(\tau,\phi)}\,d\Omega \;=\; \sum_{i=1}^{m} a_i\, c_\ell\, \delta_{j0}.$$

Hence, the low-frequency bias in each single neuron's expansion (i.e., the decay of c_ℓ with ℓ) underpins the frequency behavior of the entire network on the sphere. Analyzing these coefficients $\{c_\ell\}$ thus reveals how the SNN distributes energy across spherical frequencies, providing insight into phenomena like the Frequency Principle.

A.3 Detailed Error Dynamics

We now analyze the case where only the *output-layer* coefficients $\theta = (a_1,\ldots,a_m)^\top$ are trained, while the directions $w_i \in S^2$ remain *fixed*. We define the *error function* in equation 4 and measure this error via the L^2-type loss equation 2 and we train θ by gradient descent on $L(\theta)$.

Observe that

$$\frac{\partial u(\tau,\phi;\theta)}{\partial a_i} \;=\; \mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big),$$

so

$$\frac{\partial D(\tau,\phi)}{\partial t} \;=\; \sum_{i=1}^{m} \frac{\partial u(\tau,\phi;\theta)}{\partial a_i}\,\frac{\partial L(\theta)}{\partial a_i}.$$

Moreover, for the reason that

$$\frac{\partial L(\theta)}{\partial a_i} \;=\; -\int_0^{2\pi}\!\!\int_0^{\pi} \Big[u(\tau',\phi';\theta) - h(\tau',\phi')\Big]\,\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\,\sin\tau'\,d\tau'\,d\phi',$$

hence

$$\frac{\partial L(\theta)}{\partial a_i} \;=\; -\int_0^{2\pi}\!\!\int_0^{\pi} D(\tau',\phi')\,\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\,\sin\tau'\,d\tau'\,d\phi'.$$

Consequently,

$$\begin{aligned}\frac{\partial D(\tau,\phi)}{\partial t} \;&=\\ -\sum_{i=1}^{m}\Big\{&\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)\int_{S^2} D(\tau',\phi')\,\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\,d\Omega\\ +\; a_i^2\,&\nabla_{w_i}\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)\int_{S^2} D(\tau',\phi')\,\nabla_{w_i}\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\,d\Omega\Big\}.\end{aligned} \tag{22}$$

If all w_i align along the polar axis (e.g., $w_i^\top x(\tau,\phi) = \cos\tau'$), then $\mathrm{ReLU}(\cos\tau')$ is $\cos\tau'$ on $\tau' \in [0, \frac{\pi}{2}]$ and 0 otherwise. In that special case,

$$u(\tau,\phi;\theta) \;=\; \Big(\sum_{i=1}^{m} a_i\Big)\,\cos(\tau'), \quad D(\tau,\phi) \;=\; \Big(\sum_{i=1}^{m} a_i\Big)\cos(\tau') \;-\; h(\tau,\phi).$$

Then, for each i, $i = 1, 2, \ldots, m$

$$\begin{aligned}&\int_0^{2\pi}\!\!\int_0^{\pi} D(\tau',\phi')\,\mathrm{ReLU}\big(w_i^\top x(\tau',\phi')\big)\,\sin\tau'\,d\tau'\,d\phi'\\ &= \int_0^{2\pi}\!\!\int_0^{\frac{\pi}{2}} \Big(\sum_{i=1}^{m} a_i\cos(\tau') - h(\tau',\phi')\Big)\cos(\tau')\,\sin(\tau')\,d\tau'\,d\phi'.\end{aligned}$$

Since $\int_0^{2\pi}\int_0^{\frac{\pi}{2}} \cos^2(\tau')\sin(\tau')\,d\tau'\,d\phi' = \frac{2\pi}{3}$, the integral becomes

$$\frac{2\pi}{3}\sum_{i=1}^{m} a_i \;-\; \int_0^{2\pi}\!\!\int_0^{\frac{\pi}{2}} h(\tau',\phi')\,\cos(\tau')\,\sin(\tau')\,d\tau'\,d\phi' \;=\; \frac{2\pi}{3}\sum_{i=1}^{m} a_i \;-\; C(h),$$

where $C(h) = \int_0^{2\pi}\!\int_0^{\frac{\pi}{2}} h(\tau',\phi')\,\cos(\tau')\,\sin(\tau')\,d\tau'\,d\phi'$ denotes the contribution from $h(\tau',\phi')$, independent of the learnable parameters $\{a_i\}$. Substituting back into equation 9 provides a closed-form expression for $\frac{\partial D(\tau,\phi)}{\partial t}$.

In general, if the directions $\{w_i\}$ differ (so $w_i^\top x(\tau',\phi')$ defines different upper hemispheres), one can still decompose each ReLU integral piecewise over $\tau' \in [0, \frac{\pi}{2}]$ but must account for the orientation of each w_i. While more involved, the integrals remain straightforward to evaluate numerically or in closed form (for certain symmetric weight configurations). Thus, even without invoking higher-dimensional mean-value arguments, one sees explicitly how each "cap" $\{\tau' \le \frac{\pi}{2}\}$ relative to w_i contributes to the gradient signal.

Hence equation 9 captures the interplay of the fixed directions, the ReLU activation, and the error distribution $D(\tau,\phi)$ in governing the time evolution of the error under gradient descent.

A.4 Proof of Theorem 3.2 and 3.3

Proof A.1 (Proof of Theorem 3.2) *This proof can be directly calculated from equation 11 that c_ℓ is positive for all ℓ and $\Big(\frac{2\pi}{3}\sum_{k=1}^{m} a_k - C(h)\Big)$ is negative. So $\frac{\partial D_\ell}{\partial t} > 0$, which means the neural network is not convergent.*

Proof A.2 (Proof of Theroem 3.3) *The proof is similar to the proof of Theorem 3.2.*

Remark A.1 *Intuitively, if the error $D(\tau,\phi)$ is zero on the regions of the sphere (where $\mathrm{ReLU}(w_i^\top x)$ is active) that contribute to the integral, then each mode sees no gradient at initialization. Thus no low-frequency advantage emerges, leading to immediate FP violation.*

Remark A.2 *These results highlight that high-frequency spherical harmonic components, whose amplitudes decay on the order of $\ell^{5/2}/2^{\ell}$, are difficult to sustain in the trained network unless the factors $D(\tau_i, \phi_i)\ \sin(2\,\tau_i)$ grow at least as fast as $2^{\ell}/\ell^{5/2}$. Consequently,* violating *the frequency principle at large ℓ requires significantly amplified error terms, while for moderate or slowly varying $D(\tau_i, \phi_i)\ \sin(2\,\tau_i)$, the exponential decay ensures that the learned solution exhibits low-frequency dominance in accordance with the FP.*

A.5 Figures of FirstNumerical Experiments in Section 3

A.5.1 Case 1: Complete Adherence to Frequency Principle

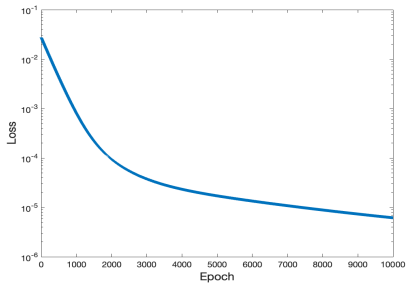
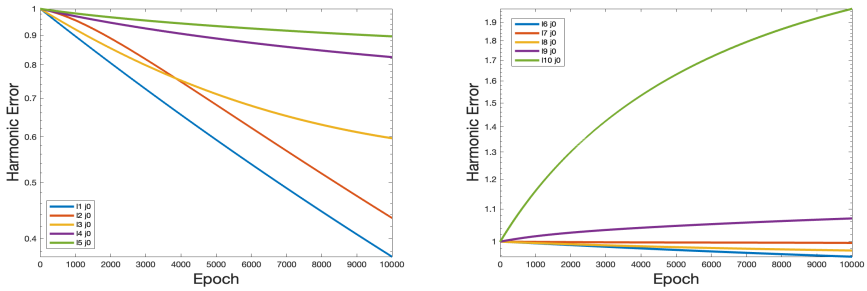


Figure 2: Training loss (Best: 6.16e-6).



(a) $l = 1, 2, \ldots, 5,\ j = 0$ (b) $l = 6, 7, \ldots, 10,\ j = 0$

Figure 3: Harmonic errors for $l = 1, 2, \ldots, 10,\ j = 0$.

A.5.2 Case 2: Partial Adherence to Frequency Principle

A.5.3 Case 3: Contradiction to Frequency Principle

A.6 Figures of Second Numerical Experiments in Section 3

A.6.1 Case 1: Partial Adherence to Frequency Principle

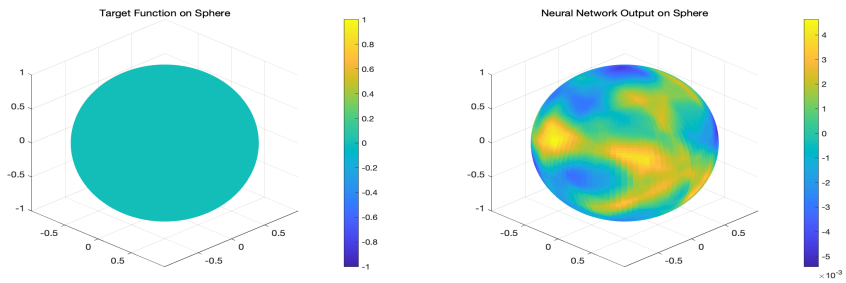
A.6.2 Case 2: Contradiction to Frequency Principle

A.7 Gradient Calculation Details

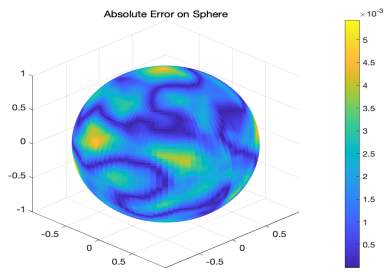
This appendix provides the detailed calculation of the gradient term $\nabla_{w_i}\Big(c_\ell D_{j0}^{(\ell)}(R_{w_i})\Big)$ that was referenced in Section 4.

To make G_ℓ^j explicit, we need to calculate

$$\nabla_{w_i}\Big(c_\ell D_{j0}^{(\ell)}(R_{w_i})\Big),$$



(a) Target function (b) Network output



(c) Absolute error

Figure 4: Target function, network output, and error distribution.

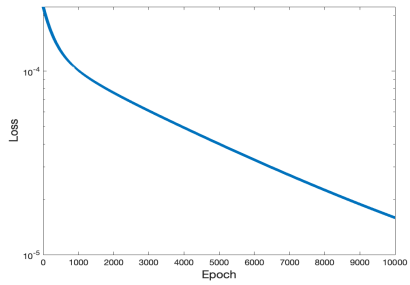


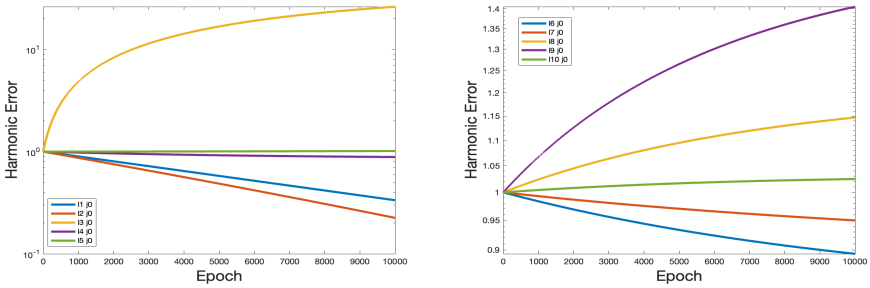
Figure 5: Training loss (Best: 1.59e-5).

which expands as

$$\nabla_{w_i}\Big(c_\ell D_{j0}^{(\ell)}(R_{w_i})\Big) = c_\ell\,\nabla_{w_i}\Big(D_{j0}^{(\ell)}(R_{w_i})\Big) = c_\ell\Big(\frac{\partial D_{j0}^{(\ell)}(R_{w_i})}{\partial R_{w_i}}\,\frac{\partial R_{w_i}}{\partial w_i}\Big).$$

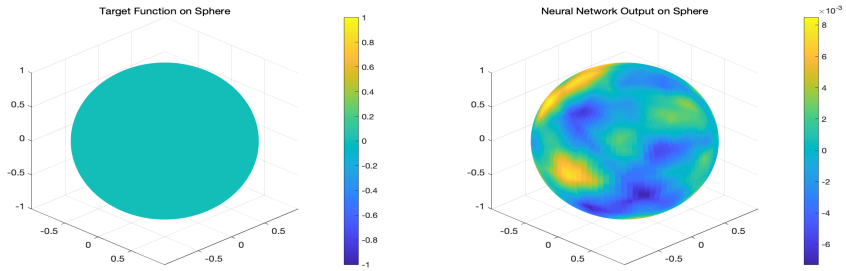
The rotation R_{w_i} depends on (x_i, y_i, z_i) through angles τ (polar) and ϕ (azimuthal), with $\cos(\tau) = z_i/\|w_i\|$ and $\tan(\phi) = y_i/x_i$. The Wigner D-matrix elements $D_{j0}^{(\ell)}(R_{w_i})$ can be written as $\sqrt{\frac{4\pi}{2\ell+1}}\,Y_\ell^j(\tau,\phi)$. Consequently,

$$\nabla_{w_i}\Big(c_\ell D_{j0}^{(\ell)}(R_{w_i})\Big) = c_\ell\sqrt{\frac{4\pi}{2\ell+1}}\left(\frac{\partial Y_\ell^j(\tau,\phi)}{\partial\tau}\,\frac{\partial\tau}{\partial w_i} + \frac{\partial Y_\ell^j(\tau,\phi)}{\partial\phi}\,\frac{\partial\phi}{\partial w_i}\right)$$

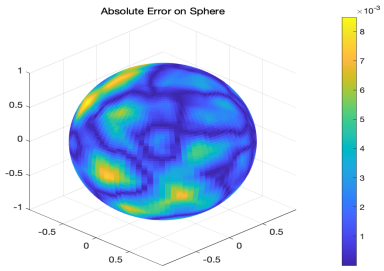


(a) $l = 1, 2, \ldots, 5$, $j = 0$ (b) $l = 6, 7, \ldots, 10$, $j = 0$

Figure 6: Harmonic errors for $l = 1, 2, \ldots, 10$, $j = 0$.



(a) Target function (b) Network output



(c) Absolute error

Figure 7: Target function, network output, and error distribution.

with

$$\frac{\partial \tau}{\partial w_i} = \begin{pmatrix} -\dfrac{\cos\tau\sin\tau\cos\phi}{\|w_i\|^2\sqrt{1-\frac{\cos^2\tau}{\|w_i\|^2}}} \\ -\dfrac{\cos\tau\sin\tau\sin\phi}{\|w_i\|^2\sqrt{1-\frac{\cos^2\tau}{\|w_i\|^2}}} \\ \dfrac{\sin\tau}{\|w_i\|^2\sqrt{1-\frac{\cos^2\tau}{\|w_i\|^2}}} \end{pmatrix}, \quad \frac{\partial \phi}{\partial w_i} = \begin{pmatrix} -\frac{\sin\phi}{\sin\tau}, \\ \frac{\cos\phi}{\sin\tau}, \\ 0 \end{pmatrix}.$$

Recall that

$$Y_\ell^j(\tau, \phi) = N_\ell^j \; P_\ell^j\big(\cos\tau\big)\, e^{ij\phi},$$

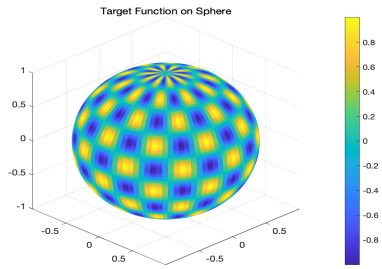


Figure 8: High-frequency initialization $\sin(10\tau)\cos(10\phi)$.

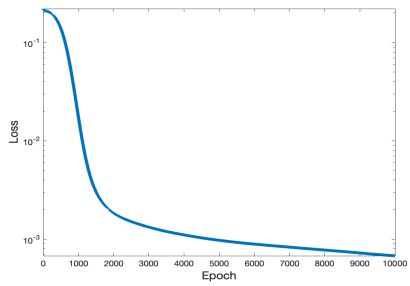


Figure 9: Training loss (Best: 6.79e-4).

where $N_\ell^j = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-j)!}{(\ell+j)!}}$ is the normalization and P_ℓ^j are the associated Legendre polynomials. Then we have

$$\begin{aligned}\frac{\partial Y_\ell^j(\tau,\phi)}{\partial\tau} &= N_\ell^j e^{ij\phi}(-\sin\tau)\left[\frac{\ell\cos\tau\, P_\ell^j(\cos\tau)-(\ell+j)\, P_{\ell-1}^j(\cos\tau)}{\sin^2\tau}\right],\\ \frac{\partial Y_\ell^j(\tau,\phi)}{\partial\phi} &= ij\, Y_\ell^j(\tau,\phi).\end{aligned}$$

Substituting these derivatives back yields the explicit form of G_ℓ^j as shown in the main text.

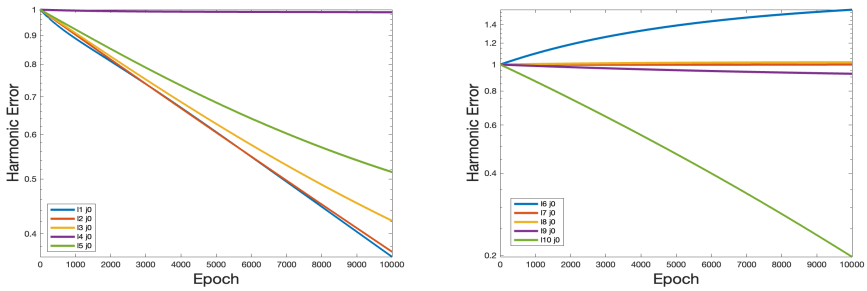
A.8 Proof of Lemma 4.1 and some Remarks

Proof A.3 (Proof of Lemma 4.1) ***Estimate for*** $\left|C_\ell^j\right|$. *From equation 13, each coefficient C_ℓ^j is a linear combination of the single-neuron coefficients c_ℓ, multiplied by bounded factors. Since $\left|c_\ell\right|$ decays on the order of $O(\ell^{5/2}/2^\ell)$ and all other terms are $O(1)$, we deduce $\left|C_\ell^j\right| = O\left(\frac{\ell^{5/2}}{2^\ell}\right)$.*

Estimate for $\left|G_\ell^j\right|$. *From equation 14, the term G_ℓ^j involves an additional derivative $\nabla_{w_i}\left(c_\ell D_{j0}^{(\ell)}(R_{w_i})\right)$. The partial derivatives with respect to the angles τ and ϕ introduce an extra factor of $O(\ell)$. Hence multiplying this additional factor by the intrinsic $\left|c_\ell\right| \approx O(\ell^{5/2}/2^\ell)$ yields $\left|G_\ell^j\right| = O\left(\frac{\ell^{7/2}}{2^\ell}\right)$.*

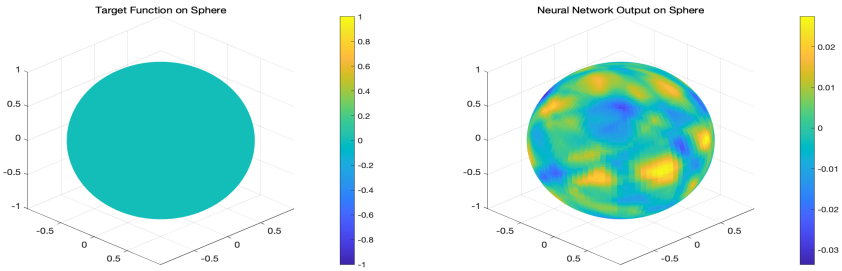
Remark A.3 *If $D\left(\tau_i^*,\phi_i^*\right) = D\left(\tau_i^\dagger,\phi_i^\dagger\right) = 0$ at precisely those mean-value points driving ReLU and rotation signals, the initialization kills all updates at every ℓ, j. Thus no preference for low frequencies arises at training onset, violating the FP in a degenerate scenario.*

Remark A.4 *Let $u(\tau,\phi;\theta) = \sum_{i=1}^m a_i\,\mathrm{ReLU}\big(w_i^\top x(\tau,\phi)\big)$ be a shallow ReLU network on the unit sphere S^2, where both $\{a_i\}$ and $\{w_i\}$ are updated by gradient descent. Suppose:*

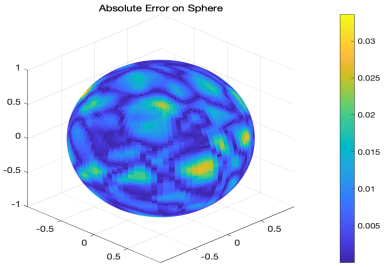


(a) $l = 1, 2, \ldots, 5$, $j = 0$ (b) $l = 6, 7, \ldots, 10$, $j = 0$

Figure 10: Harmonic errors for $l = 1, 2, \ldots, 10$, $j = 0$.



(a) Target function (b) Network output



(c) Absolute error

Figure 11: Target function, network output, and error distribution.

- *For each i, the terms $D(\tau_i^*, \phi_i^*)$ and $D(\tau_i^\dagger, \phi_i^\dagger)$ remain bounded and do not oscillate in sign over training.*
- *The coefficients $|c_\ell|$ and their derivatives follow the exponential decay property in Lemma 4.1.*

Then for sufficiently large ℓ, the Frequency Principle holds: namely, lower-frequency modes (smaller ℓ) are learned faster than higher-frequency modes ($\ell \gg 1$).

Remark A.5 *Under the same setting and notation as Corollary 4.1, assume:*

- *The sign of $D(\tau_i^*, \phi_i^*)$ and $D(\tau_i^\dagger, \phi_i^\dagger)$ remains consistent over training, but*
- *For certain lower ℓ, the rotation-driven term $|G_\ell^j|$ can exceed (or be comparable to) $|C_\ell^j|$, and can grow in a way that competes with or surpasses some higher ℓ'.*

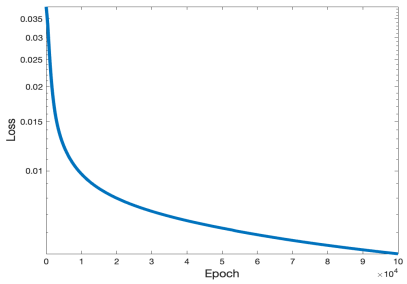
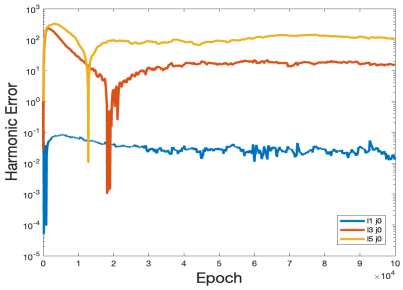
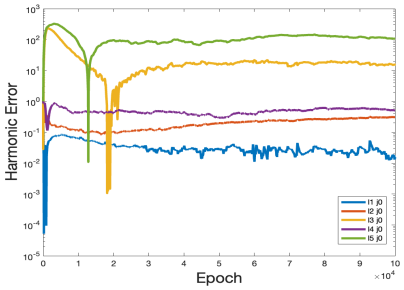


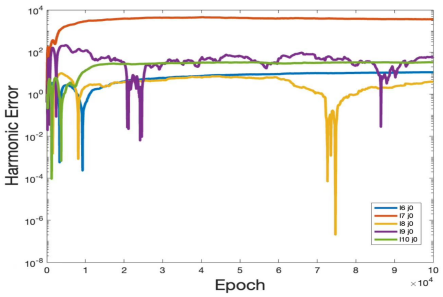
Figure 12: Training loss (Best: 5.7e-3).



(a) $l = 1, 3, 5, j = 0$



(b) $l = 1, 2, \ldots, 5, j = 0$



(c) $l = 6, 7, \ldots, 10, j = 0$

Figure 13: Harmonic errors showing mixed learning patterns.

Then there exists a regime in which the error update for ℓ-modes is at least as large as that for certain higher ℓ'-modes, thereby violating the usual low-frequency dominance. In other words, the Frequency Principle can fail for those modes.

A.9 Detailed Figures for $h(\tau, \phi) = 0$

A.9.1 Case 1: All Frequencies follow the Frequency Principle

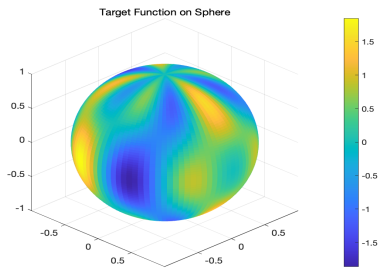
A.9.2 Case 2: Partial Adherence to Frequency Principle

A.9.3 Case 3: All Frequencies violate the Frequency Principle

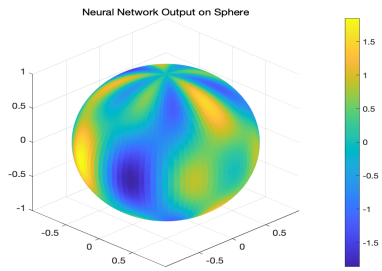
A.10 Detailed Figures for $h(\tau, \phi) = \sin(\tau)\cos(3\phi) + \sin(3\tau)\cos(5\phi)$

A.10.1 Case 1: Partial Adherence to Frequency Principle

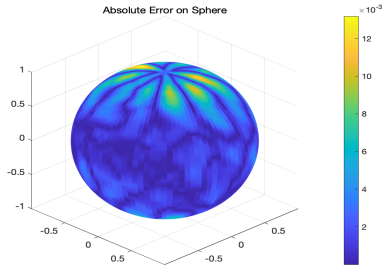
A.10.2 Case 2: Contradiction to Frequency Principle



(a) Target function



(b) Neural network output



(c) Absolute error distribution

Figure 14: Spatial comparison revealing partial frequency learning.

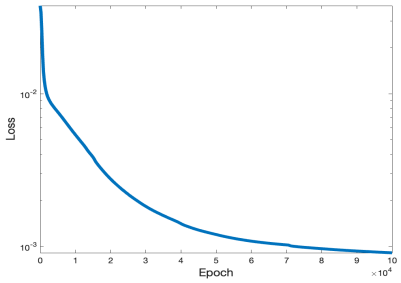
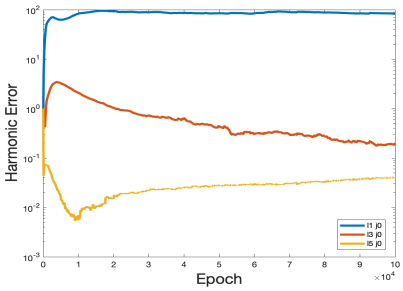
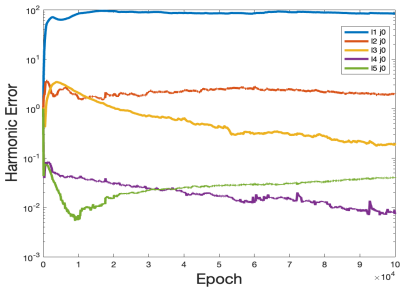


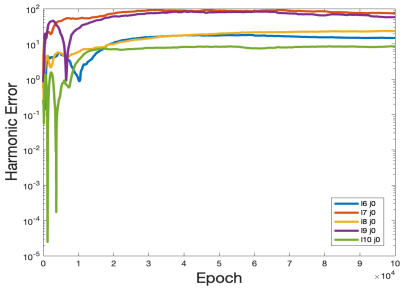
Figure 15: Training loss (Best: 9.1e-4).



(a) $l = 1, 3, 5, j = 0$

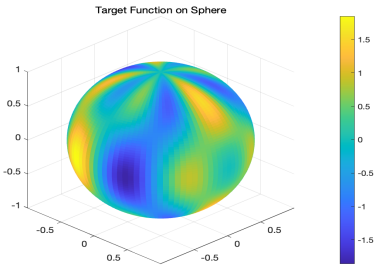


(b) $l = 1, 2, \ldots, 5, j = 0$

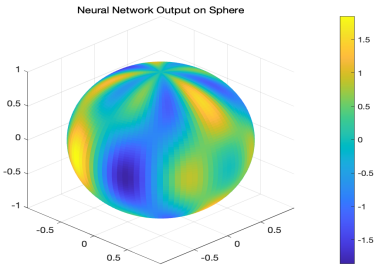


(c) $l = 6, 7, \ldots, 10, j = 0$

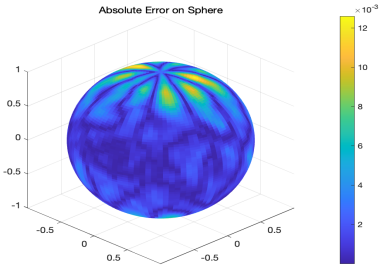
Figure 16: Harmonic errors demonstrating inverse frequency learning



(a) Target function

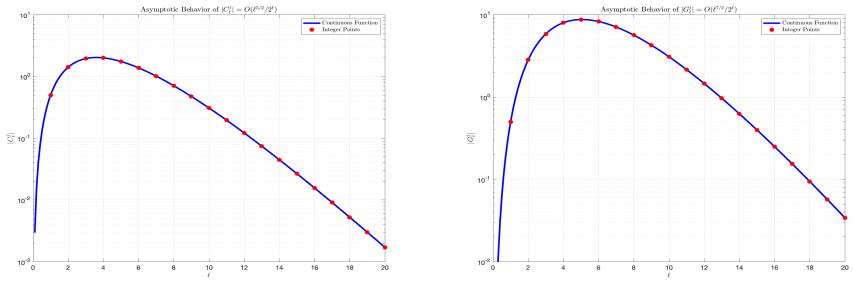


(b) Neural network output



(c) Absolute error distribution

Figure 17: Spatial comparison showing inverse frequency learning patterns



(a) $|C_\ell^j|$ (b) $|G_\ell^j|$

Figure 18: Asymptotic Behavior of $|C_\ell^j|$ and $|G_\ell^j|$

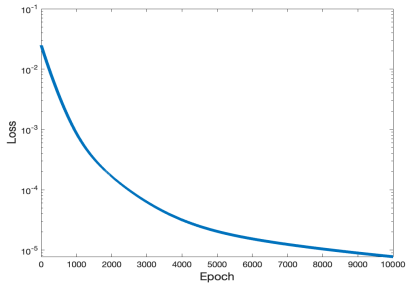
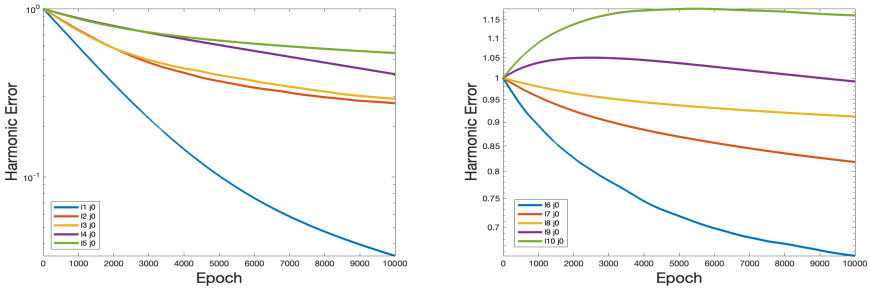
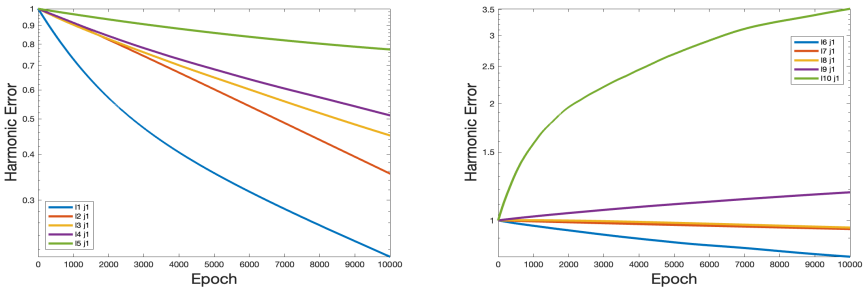


Figure 19: Image of Loss of the sixth test (Best Loss: 7.76e-6).



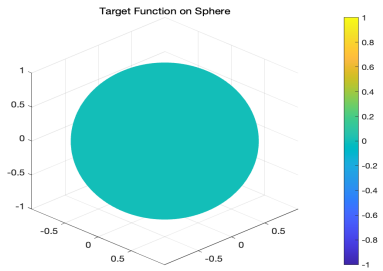
(a) $l = 1, 2, \ldots, 5$, $j = 0$ (b) $l = 6, 7, \ldots, 10$, $j = 0$

Figure 20: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10$, $j = 0$.

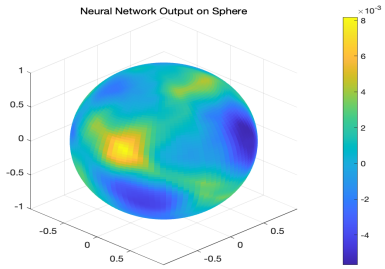


(a) $l = 1, 2, \ldots, 5$, $j = 1$ (b) $l = 6, 7, \ldots, 10$, $j = 1$

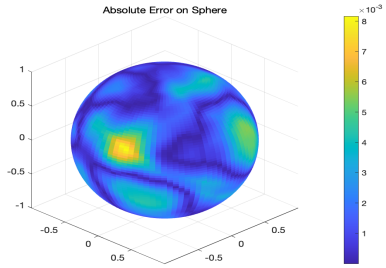
Figure 21: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10$, $j = 1$.



(a) Image of the target function



(b) Image of the SNN output



(c) Image of the error between the target function and SNN

Figure 22: The target function, SNN output, and error of the sixth test.

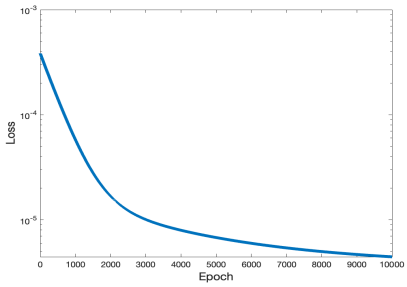
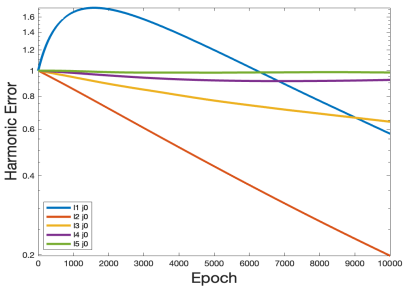
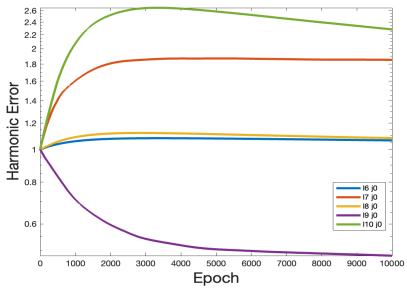


Figure 23: Image of Loss of the seventh test (Best Loss: 4.45e-6).

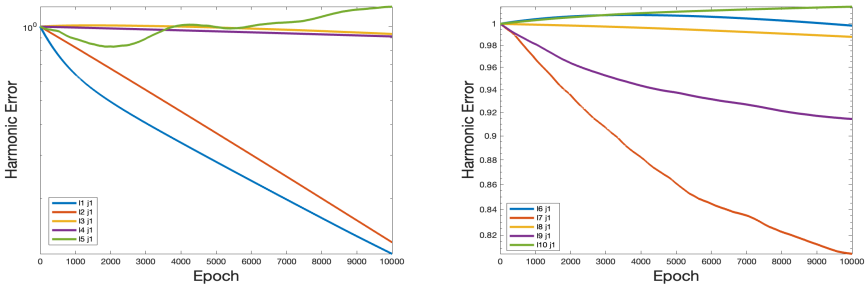


(a) $l = 1, 2, \ldots, 5, j = 0$



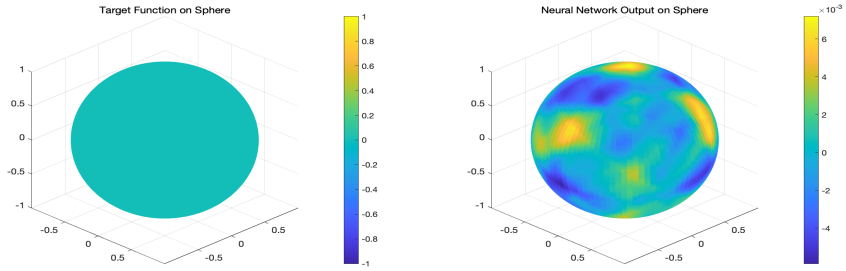
(b) $l = 6, 7, \ldots, 10, j = 0$

Figure 24: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10, j = 0$.

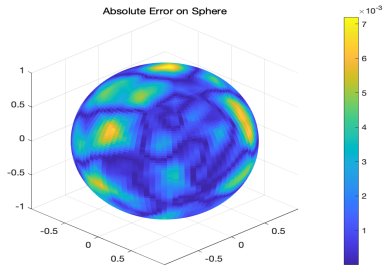


(a) $l = 1, 2, \ldots, 5$, $j = 1$ (b) $l = 6, 7, \ldots, 10$, $j = 1$

Figure 25: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10$, $j = 1$.



(a) Image of the target function (b) Image of the SNN output



(c) Image of the error between the target function and SNN

Figure 26: The target function, SNN output, and error of the seventh test.

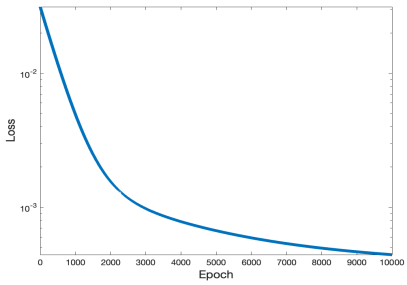
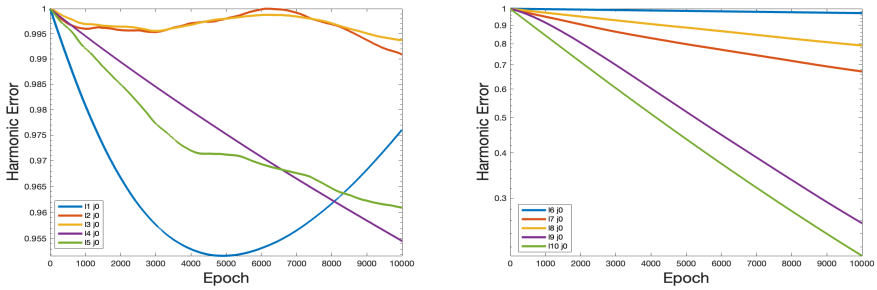
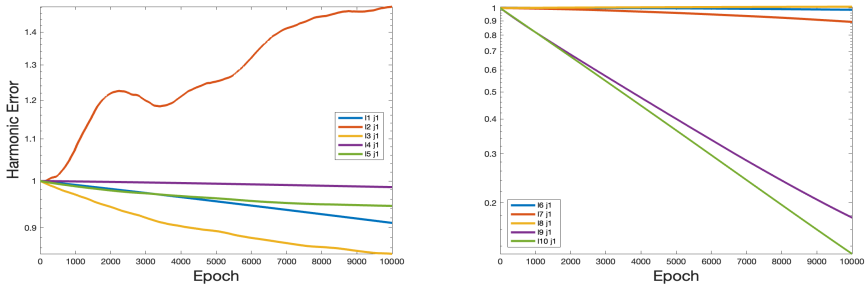


Figure 27: Image of Loss of the eighth test (Best Loss: 4.45e-4).



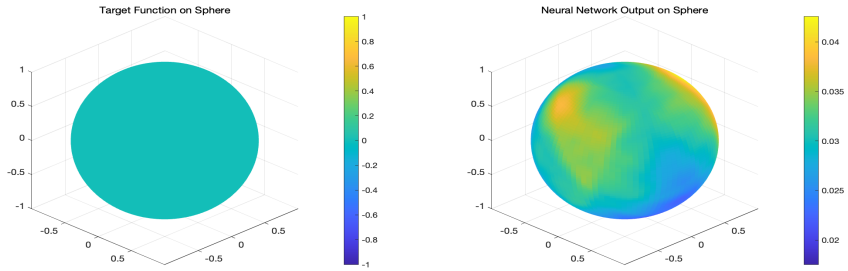
(a) $l = 1, 2, \ldots, 5$, $j = 0$ (b) $l = 6, 7, \ldots, 10$, $j = 0$

Figure 28: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10$, $j = 0$.

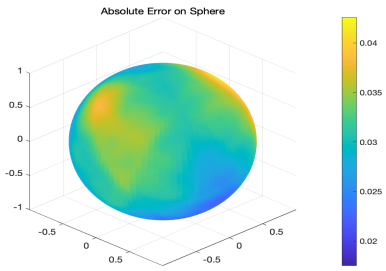


(a) $l = 1, 2, \ldots, 5$, $j = 1$ (b) $l = 6, 7, \ldots, 10$, $j = 1$

Figure 29: Harmonic Errors of c_{lj} When $l = 1, 2, \ldots, 10$, $j = 1$.



(a) Image of the target function (b) Image of the SNN output



(c) Image of the error between the target function and SNN

Figure 30: The target function, SNN output, and error of the eighth test.

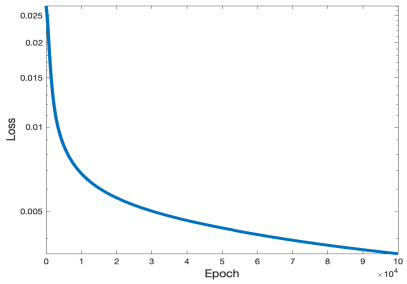
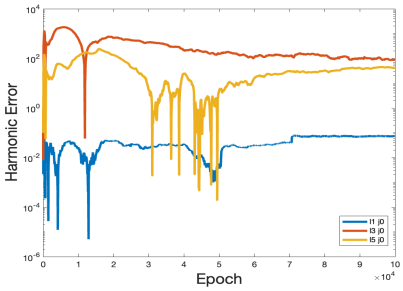
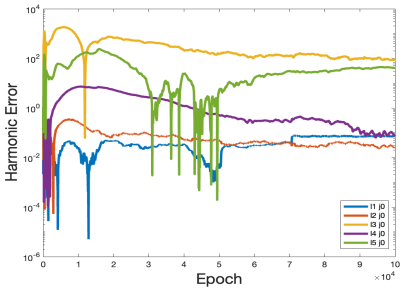


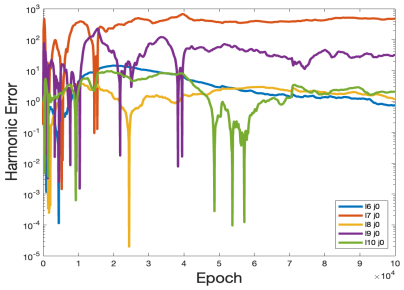
Figure 31: Training loss evolution for trainable weights showing mixed convergence patterns



(a) $l = 1, 3, 5, j = 0$

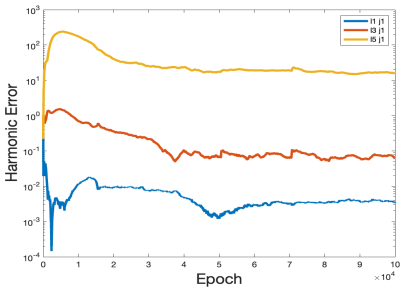


(b) $l = 1, 2, \ldots, 5, j = 0$

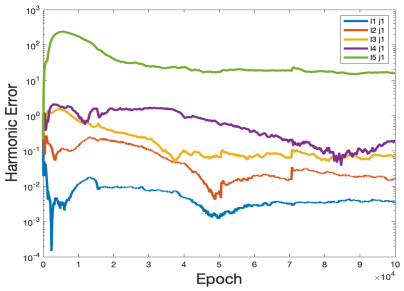


(c) $l = 6, 7, \ldots, 10, j = 0$

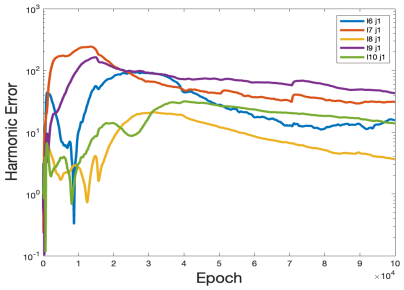
Figure 32: Harmonic Errors Revealing Interleaved Frequency Learning



(a) $l = 1, 3, 5, j = 1$

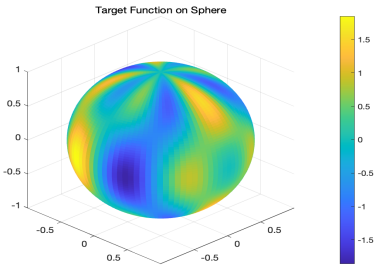


(b) $l = 1, 2, \ldots, 5, j = 1$

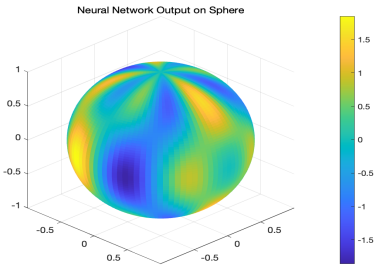


(c) $l = 6, 7, \ldots, 10, j = 1$

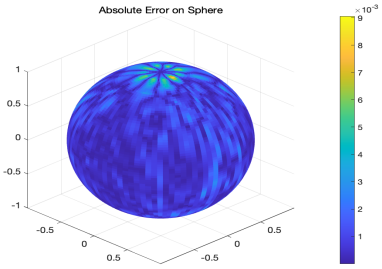
Figure 33: Harmonic Errors Revealing Interleaved Frequency Learning



(a) Target function



(b) Neural network output



(c) Absolute error distribution

Figure 34: Spatial comparison showing mixed-frequency approximation

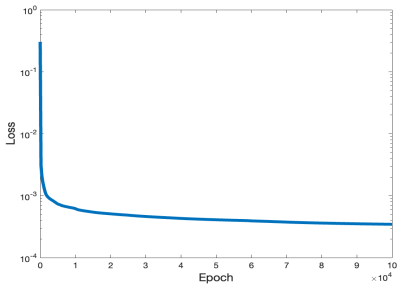
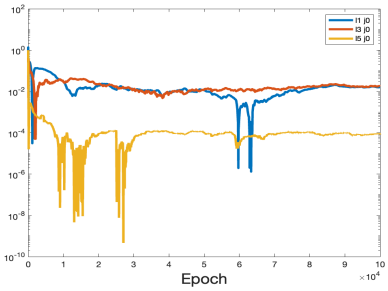
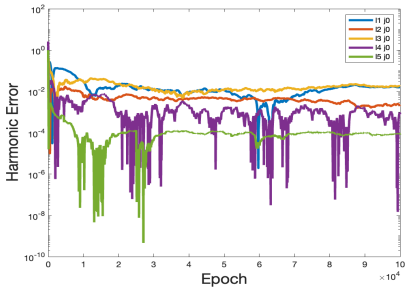


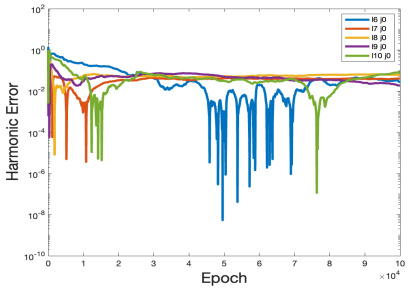
Figure 35: Training loss evolution showing inverse frequency learning



(a) $l = 1, 3, 5, j = 0$

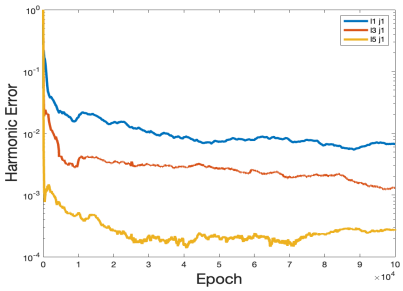


(b) $l = 1, 2, \ldots, 5, j = 0$

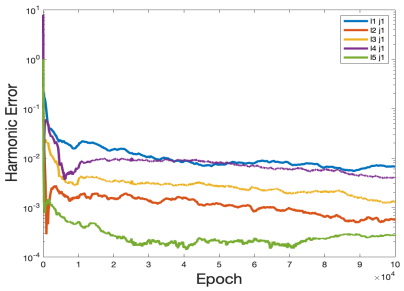


(c) $l = 6, 7, \ldots, 10, j = 0$

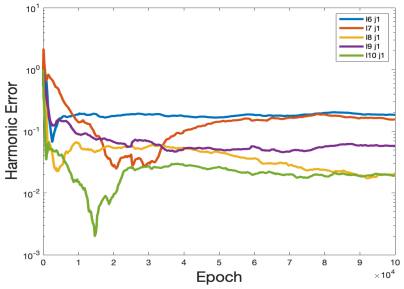
Figure 36: Harmonic Errors showing prioritized high-frequency learning



(a) $l = 1, 3, 5, j = 1$

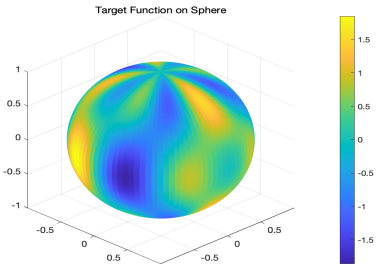


(b) $l = 1, 2, \ldots, 5, j = 1$

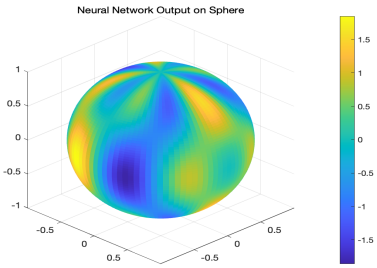


(c) $l = 6, 7, \ldots, 10, j = 1$

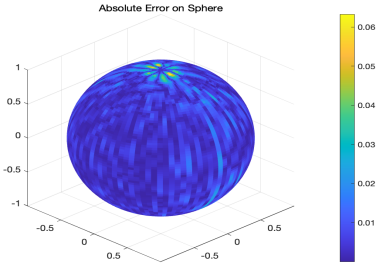
Figure 37: Harmonic Errors showing prioritized high-frequency learning



(a) Target function



(b) Neural network output



(c) Absolute error distribution

Figure 38: Spatial comparison revealing high-frequency dominance