SIAM J. MATH. DATA SCI.
Vol. 4, No. 4, pp. 1272–1292

On the Exact Computation of Linear Frequency Principle Dynamics and Its Generalization*

Tao Luo†, Zheng Ma‡, Zhi-Qin John Xu§, and Yaoyu Zhang¶

Abstract. Recent works show the intriguing phenomenon of the frequency principle (F-Principle) that deep neural networks (DNNs) fit the target function from low to high frequency during training, which provides insight into the training and generalization behavior of DNNs in complex tasks. In this paper, through analysis of an infinite-width two-layer NN in the neural tangent kernel regime, we derive the exact differential equation, namely the linear frequency-principle (LFP) model, governing the evolution of NN output function in the frequency domain during training. Our exact computation applies for general activation functions with no assumption on size and distribution of training data. This LFP model unravels that higher frequencies evolve polynomially or exponentially slower than lower frequencies depending on the smoothness/regularity of the activation function. We further bridge the gap between training dynamics and generalization by proving that the LFP model implicitly minimizes a frequency-principle norm (FP-norm) of the learned function, by which higher frequencies are more severely penalized depending on the inverse of their evolution rate. Finally, we derive an a priori generalization error bound controlled by the FP-norm of the target function, which provides a theoretical justification for the empirical results that DNNs often generalize well for low-frequency functions.

Key words. deep learning, frequency principle, Fourier analysis, two-layer neural network, neural tangent kernel, optimization

* Received by the editors September 7, 2021; accepted for publication (in revised form) June 28, 2022; published electronically November 21, 2022.

https://doi.org/10.1137/21M1444400

Funding: This research was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102 and the HPC School of Mathematical Sciences and the Student Innovation Center at Shanghai Jiao Tong University. The first author was supported by National Natural Science Foundation of China grant 12101401. The second author was supported by National Key R&D Program of China grant 2020YFA0712000 and National Natural Science Foundation of China grant 12031013. The third author was supported by National Key R&D Program of China grant 2019YFA0709503, the Shanghai Sailing Program, Natural Science Foundation of Shanghai grant 20ZR1429000, and National Natural Science Foundation of China grant 62002221. The fourth author was supported by the National Natural Science Foundation of China grant 12101402, Shanghai Municipal of Science and Technology Project grant 20JC1419500, Shanghai Municipal of Science and Technology Major Project 2021SHZDZX0102, the Lingang Laboratory grant LG-QS-202202-08.

† Corresponding author. School of Mathematical Sciences, CMA-Shanghai, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China (luotao41@sjtu.edu.cn).

‡ School of Mathematical Sciences, CMA-Shanghai, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China (zhengma@sjtu.edu.cn).

§ School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China (xuzhiqin@sjtu.edu.cn).

¶ School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, and Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, 200240, People's Republic of China (zhyy.sjtu@sjtu.edu.cn).

MSC codes. 68T05, 42B35

DOI. 10.1137/21M1444400

1. Introduction. Recently, an intriguing phenomenon, known as the frequency principle (F-Principle), has shed light on understanding the success and failure of deep neural networks (DNNs). It was discovered that, in various settings, DNNs fit the target function from low to high frequency during training [20, 25, 27]. The F-Principle implies that DNNs are biased toward a low-frequency fitting of the training data, which provides hints to the generalization of DNNs in practice [16, 25]. The F-Principle has provided valuable guidance in designing DNN-based algorithms [5, 7, 13, 14]. The convergence behavior from low to high frequency is also consistent with other empirical studies showing that DNNs increase the complexity of the output function during the training process quantified by various complexity measures [2, 17, 19, 22]. An overview of the F-Principle can be found in [24].

Despite the rich practical implications of the F-Principle, the gap between F-Principle training dynamics and success or failure of DNNs (i.e., generalization performance) remains a key theoretical challenge. Bridging this gap requires an exact characterization of the F-Principle accounting for the conditions of overparameterization and finite training data in practice, which is not provided by existing theories [4, 6, 8, 12].

In this work, based on mean-field analysis of an infinite-width two-layer NN in the neural tangent kernel (NTK) regime, we derive the exact differential equation, namely the linear frequency-principle (LFP) model, governing the evolution of the NN output function in the frequency domain during training. Our exact computation applies for general activation functions with no assumption on size and distribution of training data. Our LFP model rigorously characterizes the F-Principle and unravels that higher frequencies evolve polynomially or exponentially slower than lower frequencies depending on the smoothness/regularity of the activation function. We further prove that LFP dynamics implicitly minimizes a frequency-principle norm (FP-norm), by which higher frequencies are more severely penalized depending on the inverse of their evolution rate. Specifically, for one-dimensional (1-d) regression problems, this optimization yields a linear spline, a cubic spline, or their combination depending on parameter initialization for ReLU activation. Finally, we derive an a priori generalization error bound controlled by the FP-norm of the target function, which provides a unified qualitative explanation to the success and failure of DNNs. These three results are demonstrated by Theorems 1, 2 , and 3, respectively. For a better understanding of how we arrive at the three theorems, we depict a sketch of the proof for each theorem in Figure 1 and all proofs can be found in the supplementary materials (supplement.pdf [local/web 585KB]).

The structure of the paper is as follows. We review related works in section 2. Before we present our results, we introduce some preliminaries in section 3. Then, we show the exact computation of the LFP model in section 4. In section 5, we explicitize the implicit bias of the F-Principle by proving the equivalence between the LFP model and an optimization problem with an explicit penalty function. Further, we estimate an a priori generalization error bound for the LFP model in section 6. In section 7, we use experiments to validate the effectiveness

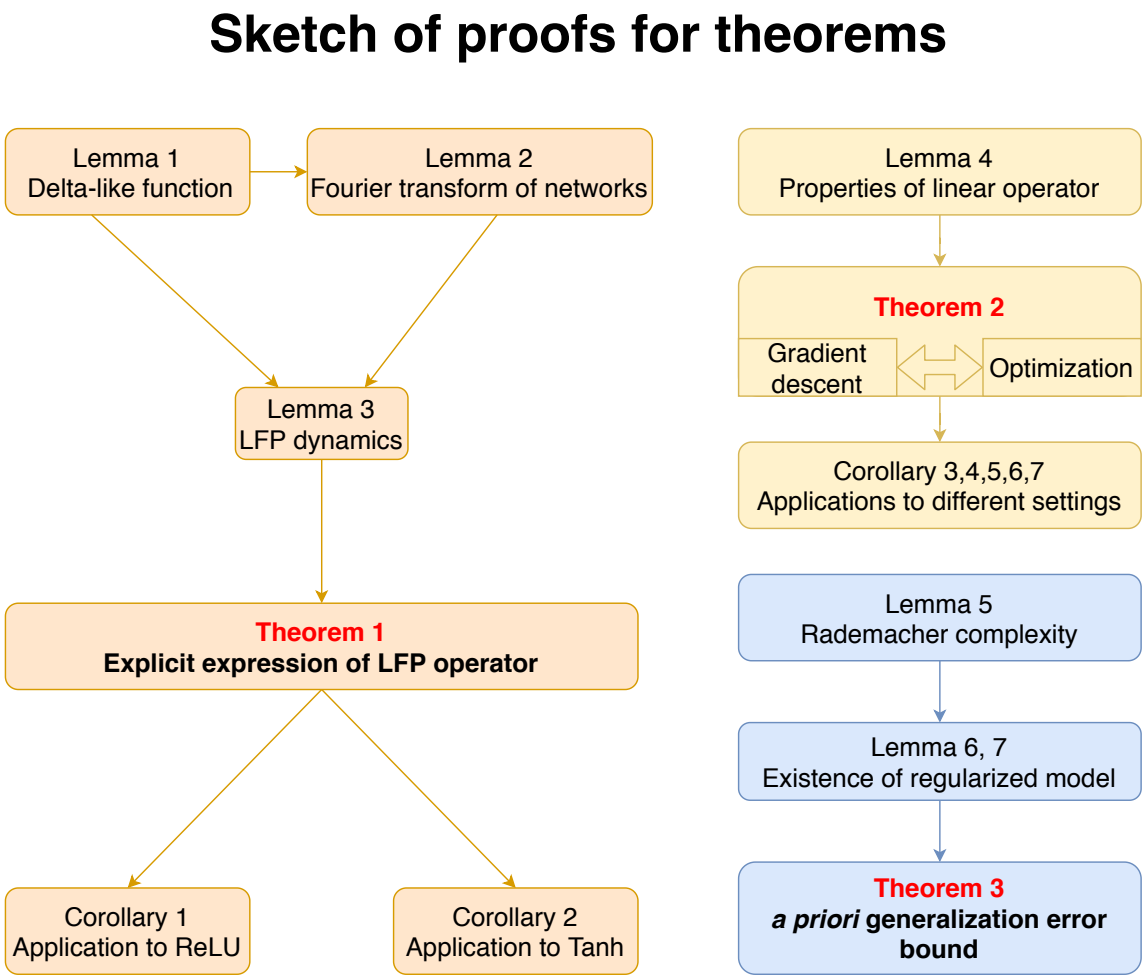


Figure 1. *Main theoretical results and sketch of proofs.*

of the LFP model for ReLU and tanh activation functions. Finally, we present conclusions and a discussion in section 8.

2. Related works. A series of works have been devoted to revealing underlying mechanisms of the F-Principle. [23] and [25] show that the gradient of low-frequency loss exponentially dominates that of high-frequency ones when parameters are small for DNNs with tanh activation. A key mechanism of the F-Principle has pointed out that the low-frequency-dominant gradient is a consequence of the smoothness of the activation function. [20] later extended the framework of the tanh activation function to the ReLU activation function. [15] estimate the dynamics of different frequency components of the loss function for arbitrary data distribution with mild regularity assumption and sufficiently large size of training data.

At the same time as our work, several parallel works have also analyzed the F-Principle (or spectral bias) in the NTK regime. [4] and [8] estimate the convergence speed of each frequency for two-layer wide ReLU networks in the NTK regime with the assumption of a

sufficiently large size of training data uniformly distributed on a hyper-sphere. [3] relax the assumption on data distribution to a nonuniform one, which is restricted to a 2-d sphere, and they derive a similar frequency bias for two-layer wide ReLU networks in the NTK regime. [6] study the dependence of the spectral bias on the sample size. Several other works also focus on studying the spectral of the Gram matrix in the NTK regime [1, 28].

In this work, our exact derivation of LFP dynamics makes no assumption about the distribution and size of training data. It is the first NN-derived quantitative model that not only shows the origin of the F-Principle but also can be used to analyze both its training and generalization consequence.[1]

3. Preliminaries.

We provide some preliminary results in this section.

3.1. Fourier transforms.

The Fourier transform of a function g is denoted by $\hat{g}$ or $\mathcal{F}[g]$. The 1-d Fourier transform and its inverse transform are defined by

$$\mathcal{F}[g](\xi) = \mathcal{F}_{x\to\xi}[g](\xi) = \int_{\mathbb{R}} g(x)\mathrm{e}^{-2\pi \mathrm{i}\xi x}\mathrm{d}x, \tag{3.1}$$

$$\mathcal{F}^{-1}[g](x) = \mathcal{F}^{-1}_{\xi\to x}[g](x) = \int_{\mathbb{R}} g(\xi)\mathrm{e}^{2\pi \mathrm{i}\xi x}\mathrm{d}\xi. \tag{3.2}$$

Based on these, we define the high-dimensional Fourier transform and its inverse transform:

$$\mathcal{F}[g](\boldsymbol{\xi}) = \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g](\boldsymbol{\xi}) = \int_{\mathbb{R}^d} g(\boldsymbol{x})\mathrm{e}^{-2\pi \mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x}, \tag{3.3}$$

$$\mathcal{F}^{-1}[g](\boldsymbol{x}) = \mathcal{F}^{-1}_{\boldsymbol{\xi}\to\boldsymbol{x}}[g](\boldsymbol{x}) = \int_{\mathbb{R}^d} g(\boldsymbol{\xi})\mathrm{e}^{2\pi \mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}. \tag{3.4}$$

Here and later, the vector $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{x}^\perp = \boldsymbol{x} - (\boldsymbol{x}\cdot\hat{\boldsymbol{w}})\hat{\boldsymbol{w}}$ for a given $\boldsymbol{w} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\}$ with $\hat{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$. We list some useful and well-known results for 1-d as well as high-dimensional Fourier transforms in Appendix SM1. To compute rigorously, we work in the theory of tempered distributions. Let $\mathcal{S}(\mathbb{R}^d)$ be the Schwartz space on $\mathbb{R}^d$ and $\mathcal{S}'(\mathbb{R}^d) := (\mathcal{S}(\mathbb{R}^d))'$ is the space of tempered distributions. For any Schwartz function $\phi \in \mathcal{S}(\mathbb{R}^d)$ and any tempered distribution $\psi \in \mathcal{S}'(\mathbb{R}^d)$, we write the pairing $\langle\psi,\phi\rangle := \langle\psi,\phi\rangle_{\mathcal{S}'(\mathbb{R}^d),\mathcal{S}(\mathbb{R}^d)} = \psi(\phi)$, and then the Fourier transform of ψ is defined by

$$\langle\mathcal{F}[\psi],\phi\rangle = \langle\psi,\mathcal{F}[\phi]\rangle. \tag{3.5}$$

3.2. High-dimensional delta-like function.

Here we introduce a delta-like function, which is essential for the calculation of the Fourier transform of the NNs due to the affine structure. In particular, to calculate $\mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{\nu}^\intercal\boldsymbol{x})](\boldsymbol{\xi})$, a delta-like function emerges naturally, i.e., $\mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{\nu}^\intercal\boldsymbol{x})](\boldsymbol{\xi}) = \delta_{\boldsymbol{\nu}}(\boldsymbol{\xi})\mathcal{F}[g](\boldsymbol{\xi}^\intercal\boldsymbol{\nu})$. In the following, we give a rigorous definition and provide two lemmas, which are essential to the proof of Lemma 3.

[1] A previous incomplete version of this work was released at arXiv [30].

Definition 1. *Given a nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$, *we define the delta-like function* $\delta_{\boldsymbol{w}} : \mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$ *such that for any* $\phi \in \mathcal{S}(\mathbb{R}^d)$,

$$\langle \delta_{\boldsymbol{w}}, \phi \rangle = \int_{\mathbb{R}} \phi(y\boldsymbol{w}) \mathrm{d}y. \tag{3.6}$$

Lemma 1 (scaling property of delta-like function). *Given any nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$ *with* $\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, *we have*

$$\frac{1}{\|\boldsymbol{w}\|^d} \delta_{\hat{\boldsymbol{w}}} \left(\frac{\boldsymbol{x}}{\|\boldsymbol{w}\|} \right) = \delta_{\boldsymbol{w}}(\boldsymbol{x}). \tag{3.7}$$

Lemma 2 (Fourier transforms of network functions). *For any unit vector* $\boldsymbol{\nu} \in \mathbb{R}^d$, *any nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$ *with* $\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, *and* $g \in \mathcal{S}'(\mathbb{R})$ *with* $\mathcal{F}[g] \in C(\mathbb{R})$, *we have, in the sense of distribution,*

$$\text{(a)} \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{x})](\boldsymbol{\xi}) = \delta_{\boldsymbol{\nu}}(\boldsymbol{\xi}) \mathcal{F}[g](\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{\nu}), \tag{3.8}$$

$$\text{(b)} \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b)](\boldsymbol{\xi}) = \delta_{\boldsymbol{w}}(\boldsymbol{\xi}) \mathcal{F}[g] \left(\frac{\boldsymbol{\xi}^{\mathsf{T}}\hat{\boldsymbol{w}}}{\|\boldsymbol{w}\|} \right) \mathrm{e}^{2\pi \mathrm{i} \frac{b}{\|\boldsymbol{w}\|} \boldsymbol{\xi}^{\mathsf{T}}\hat{\boldsymbol{w}}}, \tag{3.9}$$

$$\text{(c)} \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[\boldsymbol{x} g(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b)](\boldsymbol{\xi}) = \frac{\mathrm{i}}{2\pi} \nabla_{\boldsymbol{\xi}} \left[\delta_{\boldsymbol{w}}(\boldsymbol{\xi}) \mathcal{F}[g] \left(\frac{\boldsymbol{\xi}^{\mathsf{T}}\hat{\boldsymbol{w}}}{\|\boldsymbol{w}\|} \right) \mathrm{e}^{2\pi \mathrm{i} \frac{b}{\|\boldsymbol{w}\|} \boldsymbol{\xi}^{\mathsf{T}}\hat{\boldsymbol{w}}} \right]. \tag{3.10}$$

4. Exact derivation of LFP model.

In this section, we first present the general form of the LFP model for two-layer NNs. Then, we exactly compute the LFP model in the Fourier domain and derive the expressions for two commonly used activation functions, i.e., $\mathrm{ReLU}(x) := \max(x, 0)$ and $\tanh(x)$.

For any positive integer N, we denote the set $\{1, 2, \ldots, N\}$ by $[N]$. The training dataset $S = \{(\boldsymbol{x}_i, y_i)\})_{i=1}^n$, where $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent and identically distributed sampled from unknown distribution $\mathcal{D}$ on a domain $\Omega \subset \mathbb{R}^d$ and $y_i = f(\boldsymbol{x}_i)$, $i \in [n]$, for some unknown function f.

4.1. Mean-field kernel dynamics in frequency domain.

We suppose that $f \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and that the activation function is locally H^1 and grows polynomially, i.e., $|\sigma(z)| \le C|z|^p$ for some $p > 0$.

We consider the following gradient descent dynamics of the empirical risk R_S of a network function $f(\cdot, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$:

$$\begin{cases} \dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}), \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \end{cases} \tag{4.1}$$

where

$$R_S(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n (f(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i)^2. \tag{4.2}$$

Then the training dynamics of output function $f(\cdot, \boldsymbol{\theta})$ is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}, \boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \dot{\boldsymbol{\theta}} \\ &= -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) \\ &= -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i, \boldsymbol{\theta})(f(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i) \\ &= -\sum_{i=1}^{n} K_m(\boldsymbol{x}, \boldsymbol{x}_i)(f(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i),\end{aligned}$$

where for time t the NTK evaluated at $(\boldsymbol{x}, \boldsymbol{x}') \in \Omega \times \Omega$ reads as

$$K_m(\boldsymbol{x}, \boldsymbol{x}')(t) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}(t)) \cdot \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}', \boldsymbol{\theta}(t)). \tag{4.3}$$

The gradient descent of the model thus becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(f(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f(\boldsymbol{x})\right) = -\sum_{i=1}^{n} K_m(\boldsymbol{x}, \boldsymbol{x}_i)(t) \left(f(\boldsymbol{x}_i, \boldsymbol{\theta}(t)) - f(\boldsymbol{x}_i)\right). \tag{4.4}$$

Define the residual $\boldsymbol{u}(\boldsymbol{x}, t) = f(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f(\boldsymbol{x})$ and the empirical density $\rho(\boldsymbol{x}) = \sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$. We further denote $u_\rho(\boldsymbol{x}) = u(\boldsymbol{x})\rho(\boldsymbol{x})$. Therefore the dynamics for u becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} u(\boldsymbol{x}, t) = -\int_{\mathbb{R}^d} K_m(\boldsymbol{x}, \boldsymbol{x}')(t) u_\rho(\boldsymbol{x}', t) \mathrm{d}\boldsymbol{x}'. \tag{4.5}$$

From now on, we consider the two-layer NN

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x} + b_j) \tag{4.6}$$

$$= \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sigma^*(\boldsymbol{x}, \boldsymbol{q}_j), \tag{4.7}$$

where the vector of all parameters $\boldsymbol{\theta} = \mathrm{vec}(\{\boldsymbol{q}_j\}_{j=1}^m)$ is formed of the parameters for each neuron $\boldsymbol{q}_j = (a_j, \boldsymbol{w}_j^\intercal, b_j)^\intercal \in \mathbb{R}^{d+2}$ and $\sigma^*(\boldsymbol{x}, \boldsymbol{q}_j) = a_j \sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x} + b_j)$ for $j \in [m]$. We consider the kernel regime that $m \gg 1$ and assume that $b \sim \mathcal{N}(0, \sigma_b^2)$ with $\sigma_b \gg 1$. For the two-layer network, its NTK can be calculated as follows:

$$K_m(\boldsymbol{x}, \boldsymbol{x}')(t) = \frac{1}{m} \sum_{j=1}^{m} \nabla_{\boldsymbol{q}_j} \sigma^*(\boldsymbol{x}, \boldsymbol{q}_j(t)) \cdot \sigma^*(\boldsymbol{x}', \boldsymbol{q}_j(t)), \tag{4.8}$$

where the parameters $\boldsymbol{q}_j$ are evaluated at time t. Under some weak condition and for sufficiently large m, [11] proved that the dynamics (4.5), with a high probability, converges to the following dynamics for any $t \in \mathbb{R}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} u(\boldsymbol{x}, t) = -\int_{\mathbb{R}^d} K(\boldsymbol{x}, \boldsymbol{x}') u_\rho(\boldsymbol{x}', t) \mathrm{d}\boldsymbol{x}', \tag{4.9}$$

where the kernel only depends on the initial distribution of parameters and reads as

$$K(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{q}}\nabla_{\boldsymbol{q}}\sigma^*(\boldsymbol{x},\boldsymbol{q})\cdot\sigma^*(\boldsymbol{x}',\boldsymbol{q}) \tag{4.10}$$

$$\begin{aligned} &= \mathbb{E}_{\boldsymbol{q}}(\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}'+b) + a^2\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}'+b)\boldsymbol{x}^\intercal\boldsymbol{x}' \\ &\quad + a^2\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}'+b)). \end{aligned} \tag{4.11}$$

Intuitively, this is because $K_m(\boldsymbol{x},\boldsymbol{x}')(t) = K(\boldsymbol{x},\boldsymbol{x}') + O(\frac{1}{\sqrt{m}})$ according to the law of large numbers. In the following, we analyze (4.9) and calculate its formulation in the frequency domain.

We start with the following lemma.

Lemma 3 (LFP dynamics for general DNNs). *The dynamics* (4.9) *has the following expression in the frequency domain for all* $\phi \in \mathcal{S}(\mathbb{R}^d)$:

$$\langle \partial_t \mathcal{F}[u], \phi\rangle = -\langle \mathcal{L}[\mathcal{F}[u_\rho]], \phi\rangle, \tag{4.12}$$

where $\mathcal{L}[\cdot]$ *is the LFP operator given by*

$$\mathcal{L}[\mathcal{F}[u_\rho]] = \int_{\mathbb{R}^d} \hat{K}(\boldsymbol{\xi},\boldsymbol{\xi}')\mathcal{F}[u_\rho](\boldsymbol{\xi}')\mathrm{d}\boldsymbol{\xi}',$$

and

$$\hat{K}(\boldsymbol{\xi},\boldsymbol{\xi}') := \mathbb{E}_{\boldsymbol{q}}\hat{K}_{\boldsymbol{q}}(\boldsymbol{\xi},\boldsymbol{\xi}') := \mathbb{E}_{\boldsymbol{q}}\mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[\nabla_{\boldsymbol{q}}\sigma^*(\boldsymbol{x},\boldsymbol{q})]\cdot\overline{\mathcal{F}_{\boldsymbol{x}'\to\boldsymbol{\xi}'}[\nabla_{\boldsymbol{q}}\sigma^*(\boldsymbol{x}',\boldsymbol{q})]}. \tag{4.13}$$

The expectation $\mathbb{E}_{\boldsymbol{q}}$ *is taken w.r.t. initial distribution of parameters.*

4.2. LFP dynamics derived for two-layer networks. In this section, we derive the LFP dynamics for two-layer networks with general activation function. The key difficulty comes from the repeated integral representation of the operator. By using the Laplace method in a proper way, we overcome this difficulty and arrive at a simpler expression for the dynamics.

To simplify the notation, we define $\boldsymbol{g}_1(z) := (\sigma(z), a\sigma'(z))^\intercal$ and $g_2(z) := a\sigma'(z)$ for $z \in \mathbb{R}$. Then

$$\boldsymbol{g}_1(\boldsymbol{w}^\intercal\boldsymbol{x}+b) = \begin{pmatrix} \sigma(\boldsymbol{w}^\intercal\boldsymbol{x}+b) \\ a\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\end{pmatrix} = \begin{pmatrix} \partial_a[a\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}+b)] \\ \partial_b[a\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}+b)]\end{pmatrix}, \tag{4.14}$$

$$g_2(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\boldsymbol{x} = \nabla_{\boldsymbol{w}}[a\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}+b)] = a\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}+b)\boldsymbol{x}. \tag{4.15}$$

The following theorem is the key to the exact expression of LFP dynamics for two-layer networks.

Assumption 1. We assume that the initial distribution of $\boldsymbol{q} = (a, \boldsymbol{w}^\intercal, b)^\intercal$ satisfies the following conditions:

(i) independence of $a, \boldsymbol{w}, b$: $\rho_{\boldsymbol{q}}(\boldsymbol{q}) = \rho_a(a)\rho_{\boldsymbol{w}}(\boldsymbol{w})\rho_b(b)$;
(ii) zero mean and finite variance of b: $\mathbb{E}_b b = 0$ and $\mathbb{E}_b b^2 = \sigma_b^2 < \infty$;
(iii) radial symmetry of $\boldsymbol{w}$: $\rho_{\boldsymbol{w}}(\boldsymbol{w}) = \rho_{\boldsymbol{w}}(\|\boldsymbol{w}\|\boldsymbol{e}_1)$ where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\intercal$.

Theorem 1 (main result: explicit expression of LFP operator for two-layer networks). *Suppose that Assumption 1 holds. If $\sigma_b \gg 1$, then the dynamics (4.9) has the following expression:*

$$\langle \partial_t \mathcal{F}[u], \phi \rangle = -\langle \mathcal{L}[\mathcal{F}[u_\rho]], \phi \rangle + O(\sigma_b^{-3}), \tag{4.16}$$

where $\phi \in \mathcal{S}(\mathbb{R}^d)$ is a test function and the LFP operator is given by

$$\begin{aligned}\mathcal{L}[\mathcal{F}[u_\rho]] = &\frac{\Gamma(d/2)}{2\sqrt{2}\pi^{(d+1)/2}\sigma_b\|\boldsymbol{\xi}\|^{d-1}}\mathbb{E}_{a,r}\left[\frac{1}{r}\mathcal{F}[\boldsymbol{g}_1]\left(\frac{\|\boldsymbol{\xi}\|}{r}\right)\cdot\mathcal{F}[\boldsymbol{g}_1]\left(\frac{-\|\boldsymbol{\xi}\|}{r}\right)\right]\mathcal{F}[u_\rho](\boldsymbol{\xi})\\ &-\frac{\Gamma(d/2)}{2\sqrt{2}\pi^{(d+1)/2}\sigma_b}\nabla\cdot\left(\mathbb{E}_{a,r}\left[\frac{1}{r\|\boldsymbol{\xi}\|^{d-1}}\mathcal{F}[g_2]\left(\frac{\|\boldsymbol{\xi}\|}{r}\right)\mathcal{F}[g_2]\left(-\frac{\|\boldsymbol{\xi}\|}{r}\right)\right]\nabla\mathcal{F}[u_\rho](\boldsymbol{\xi})\right),\end{aligned} \tag{4.17}$$

where $\Gamma(\cdot)$ is the gamma function. The expectations are taken w.r.t. initial parameter distribution. Here $r = \|\boldsymbol{w}\|$ with the probability density $\rho_r(r) := \frac{2\pi^{d/2}}{\Gamma(d/2)}\rho_{\boldsymbol{w}}(r\boldsymbol{e}_1)r^{d-1}$, $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\intercal$.

Remark 1. The operator $\mathcal{L}$ presents a unified framework for general activation functions.

Remark 2. The derivatives of most activation functions decay in the Fourier domain, e.g., ReLU, tanh, and sigmoid. Hence, the dynamics in (4.16) for a higher-frequency component is slower, i.e., the F-Principle.

Remark 3. The last term in (4.17) arising from the evolution of $\boldsymbol{w}$ is much more complicated, without which our experiments show that the LFP model can still predict the learning results of two-layer wide NNs.

4.3. Exact LFP model for common activation functions. Based on (4.17), we derive the exact LFP dynamics for the cases where the activation function is ReLU or tanh.

Corollary 1 (LFP operator for ReLU activation function). *Suppose that Assumption 1 holds. If $\sigma_b \gg 1$ and $\sigma = \mathrm{ReLU}$, then the dynamics (4.9) has the following expression:*

$$\langle \partial_t \mathcal{F}[u], \phi \rangle = -\langle \mathcal{L}[\mathcal{F}[u_\rho]], \phi \rangle + O(\sigma_b^{-3}), \tag{4.18}$$

where $\phi \in \mathcal{S}(\mathbb{R}^d)$ is a test function and the LFP operator reads as

$$\begin{aligned}\mathcal{L}[\mathcal{F}[u_\rho]] = &\frac{\Gamma(d/2)}{2\sqrt{2}\pi^{(d+1)/2}\sigma_b}\mathbb{E}_{a,r}\left[\frac{r^3}{16\pi^4\|\boldsymbol{\xi}\|^{d+3}}+\frac{a^2 r}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}}\right]\mathcal{F}[u_\rho](\boldsymbol{\xi})\\ &-\frac{\Gamma(d/2)}{2\sqrt{2}\pi^{(d+1)/2}\sigma_b}\nabla\cdot\left(\mathbb{E}_{a,r}\left[\frac{a^2 r}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}}\right]\nabla\mathcal{F}[u_\rho](\boldsymbol{\xi})\right).\end{aligned} \tag{4.19}$$

The expectations are taken w.r.t. initial parameter distribution. Here $r = \|\boldsymbol{w}\|$ with the probability density $\rho_r(r) := \frac{2\pi^{d/2}}{\Gamma(d/2)}\rho_{\boldsymbol{w}}(r\boldsymbol{e}_1)r^{d-1}$, $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\intercal$.

Corollary 2 (LFP operator for tanh activation function). *Suppose that Assumption 1 holds. If $\sigma_b \gg 1$ and $\sigma = \tanh$, then the dynamics (4.9) has the following expression:*

$$\langle \partial_t \mathcal{F}[u], \phi \rangle = -\langle \mathcal{L}[\mathcal{F}[u_\rho]], \phi \rangle + O(\sigma_b^{-3}), \tag{4.20}$$

where $\phi \in \mathcal{S}(\mathbb{R}^d)$ is a test function and the LFP operator reads as

$$\mathcal{L}[\mathcal{F}[u_\rho]] = \frac{\Gamma(d/2)}{2\sqrt{2}\pi^{\frac{d+1}{2}}\sigma_b\|\boldsymbol{\xi}\|^{d-1}}\mathbb{E}_{a,r}\left[\frac{\pi^2}{r}\operatorname{csch}^2\left(\frac{\pi^2\|\boldsymbol{\xi}\|}{r}\right)+\frac{4\pi^4a^2\|\boldsymbol{\xi}\|^2}{r^3}\operatorname{csch}^2\left(\frac{\pi^2\|\boldsymbol{\xi}\|}{r}\right)\right]\mathcal{F}[u_\rho](\boldsymbol{\xi})$$
$$-\frac{\Gamma(d/2)}{2\sqrt{2}\pi^{\frac{d+1}{2}}\sigma_b}\nabla\cdot\left(\mathbb{E}_{a,r}\left[\frac{4\pi^4a^2}{r^3\|\boldsymbol{\xi}\|^{d-3}}\operatorname{csch}^2\left(\frac{\pi^2\|\boldsymbol{\xi}\|}{r}\right)\right]\nabla\mathcal{F}[u_\rho](\boldsymbol{\xi})\right). \tag{4.21}$$

The expectations are taken w.r.t. initial parameter distribution. Here $r = \|\boldsymbol{w}\|$ with the probability density $\rho_r(r) := \frac{2\pi^{d/2}}{\Gamma(d/2)}\rho_{\boldsymbol{w}}(r\boldsymbol{e}_1)r^{d-1}$, $\boldsymbol{e}_1 = (1,0,\dots,0)^\intercal$.

5. Explicitizing the implicit bias of the F-Principle. For overparameterized NNs, it has been widely observed in experiments that a gradient descent dynamics with proper initialization implicitly biases the training toward a low-frequency interpolation among infinite minimizers achieving 0 training loss [25, 27]. To understand this implicit bias quantitatively, one approach, namely explicitizing the implicit bias, is to find an explicit penalty function, the minimizer of which exactly recovers the solution of the gradient descent/flow dynamics. In this section, we prove that the implicit bias of the F-Principle indeed can be explicitized for the LFP dynamics, based on which we provide a quantitative analysis of the consequence of the F-Principle.

We first analyze a simplified LFP model with ReLU activation function in (4.19) as follows:

$$\partial_t\mathcal{F}[u] = -\mathbb{E}_{a,r}\left[\frac{r^3}{16\pi^4\|\boldsymbol{\xi}\|^{d+3}}+\frac{a^2r}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}}\right]\mathcal{F}[u_\rho](\boldsymbol{\xi}). \tag{5.1}$$

We discard the last term in (4.19) arising from the evolution of $\boldsymbol{w}$. The reason is twofold. First, experiments show that (5.1) is accurate enough to predict the wide two-layer NN output after training. Second, the last term in (4.19) is too complicated to analyze for now.

In the LFP model, the solution is implicitly regularized by a decaying coefficient for different frequencies of $\mathcal{F}[u]$ throughout the training. For a quantitative analysis of this solution, we explicitize such an implicit dynamical regularization by a constrained optimization problem as follows.

5.1. An equivalent optimization problem to the gradient flow dynamics. First, we present a general theorem that the long-time limit solution of a gradient flow dynamics is equivalent to the solution of a constrained optimization problem.

Let H_1 and H_2 be two separable Hilbert spaces and $\mathcal{P}: H_1 \to H_2$ be a bounded linear operator. Let $\mathcal{P}^*: H_2 \to H_1$ be the adjoint operator of $\mathcal{P}$, defined by

$$\langle\mathcal{P}\phi_1,\phi_2\rangle_{H_2} = \langle\phi_1,\mathcal{P}^*\phi_2\rangle_{H_1} \quad \text{for all} \quad \phi_1\in H_1,\phi_2\in H_2. \tag{5.2}$$

Lemma 4. *Suppose that H_1 and H_2 are two separable Hilbert spaces and $\mathcal{P}: H_1 \to H_2$ and $\mathcal{P}^*: H_2 \to H_1$ is the adjoint of $\mathcal{P}$. Then all eigenvalues of $\mathcal{P}^*\mathcal{P}$ and $\mathcal{P}\mathcal{P}^*$ are nonnegative. Moreover, they have the same positive spectrum. If in particular we assume that the operator $\mathcal{P}\mathcal{P}^*$ is surjective, then the operator $\mathcal{P}\mathcal{P}^*$ is invertible.*

Remark 4. For the finite dimensional case $H_2 = \mathbb{R}^n$, conditions for the operator $\mathcal{P}$ in Lemma 4 are reduced to that the matrix $\boldsymbol{P}$ has rank n (full rank).

Given $g \in H_2$, we consider the following two problems.

(i) The initial value problem

$$\begin{cases} \dfrac{\mathrm{d}\phi}{\mathrm{d}t} &= \mathcal{P}^*(g - \mathcal{P}\phi), \\ \phi(0) &= \phi_{\mathrm{ini}}. \end{cases}$$

Since this equation is linear and with nonpositive eigenvalues on the right-hand side, there exists a unique global-in-time solution $\phi(t)$ for all $t \in [0, +\infty)$ satisfying the initial condition. Moreover, the long-time limit $\lim_{t\to+\infty} \phi(t)$ exists and will be denoted as ϕ_∞.

(ii) The minimization problem

$$\min_{\phi-\phi_{\mathrm{ini}}\in H_1} \|\phi - \phi_{\mathrm{ini}}\|_{H_1}$$

$$\text{s.t.} \quad \mathcal{P}\phi = g.$$

In the following, we will show it has a unique minimizer which is denoted as $h_{\min}$.

Now we show the following equivalent theorem. We remark that, in the NTK regime, the equivalence between the optimization problem and the gradient flow dynamics is known because of the linearized dynamics. For completeness, we present a unified and general result (see Theorem 2), as well as its weighted version (see Corollary 4), for this kind of equivalence. Then we apply Theorem 2 to the linearized DNN dynamics in the parameter space (see Corollary 3). We also apply Corollary 4 to the linearized DNN dynamics in the frequency domain (see Corollary 5) and its discretized version (see Corollary 6).

Theorem 2 (equivalence between gradient descent and optimization problems). *Suppose that $\mathcal{P}\mathcal{P}^*$ is surjective. The above problems* (i) *and* (ii) *are equivalent in the sense that $\phi_\infty = \phi_{\min}$. More precisely, we have*

$$\phi_\infty = h_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}(g - \mathcal{P}\phi_{\mathrm{ini}}) + \phi_{\mathrm{ini}}. \tag{5.3}$$

The following corollaries are obtained directly from Theorem 2.

Corollary 3. *Let ϕ be the parameter vector $\boldsymbol{\theta}$ in $H_1 = \mathbb{R}^m$, g be the outputs of the training data $\boldsymbol{Y}$, and $\boldsymbol{P}$ be a full rank matrix in the linear DNN model. Then the following two problems are equivalent in the sense that $\boldsymbol{\theta}_\infty = \boldsymbol{\theta}_{\min}$.*

(A1) *The initial value problem*

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{P}^*(\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{\theta}), \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_{\mathrm{ini}}. \end{cases}$$

(A2) *The minimization problem*

$$\min_{\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{ini}}\in\mathbb{R}^m} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ini}}\|_2$$

$$\text{s.t.} \quad \boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{Y}.$$

The next corollary is a weighted version of Theorem 2.

Corollary 4. *Let H_1 and H_2 be two separable Hilbert spaces and $\Gamma : H_1 \to H_1$ be an injective operator. Define the Hilbert space $H_\Gamma := \mathrm{Im}(\Gamma)$. Let $g \in H_2$ and $\mathcal{P} : H_\Gamma \to H_2$ be an operator such that $\mathcal{P}\mathcal{P}^* : H_2 \to H_2$ is surjective. Then $\Gamma^{-1} : H_\Gamma \to H_1$ exists and H_Γ is a Hilbert space with norm $\|\phi\|_{H_\Gamma} := \|\Gamma^{-1}\phi\|_{H_1}$. Moreover, the following two problems are equivalent in the sense that $\phi_\infty = \phi_{\min}$.*

(B1) *The initial value problem*

$$\begin{cases} \dfrac{\mathrm{d}\phi}{\mathrm{d}t} = \Gamma\Gamma^*\mathcal{P}^*(g - \mathcal{P}\phi), \\ \phi(0) = \phi_{\mathrm{ini}}. \end{cases}$$

(B2) *The minimization problem*

$$\min_{\phi-\phi_0 \in H_\Gamma} \|\phi - \phi_{\mathrm{ini}}\|_{H_\Gamma}$$

$$\text{s.t.} \quad \mathcal{P}\phi = g.$$

Corollary 5. *Let $\gamma : \mathbb{R}^d \to \mathbb{R}^+$ be a positive function, h be a function in $L^2(\mathbb{R}^d)$, and $\phi = \mathcal{F}[h]$. The operator $\Gamma : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is defined by $[\Gamma\phi](\boldsymbol{\xi}) = \gamma(\boldsymbol{\xi})\phi(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \mathbb{R}^d$. Define the Hilbert space $H_\Gamma := \mathrm{Im}(\Gamma)$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal \in \mathbb{R}^{n\times d}$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$, and $\mathcal{P} : H_\Gamma \to \mathbb{R}^n$ be a surjective operator*

$$\mathcal{P} : \phi \mapsto \left(\int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_1^\intercal \boldsymbol{\xi}} \mathrm{d}\boldsymbol{\xi}, \ldots, \int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_n^\intercal \boldsymbol{\xi}} \mathrm{d}\boldsymbol{\xi} \right)^\intercal = (h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_n))^\intercal. \tag{5.4}$$

Then the following two problems are equivalent in the sense that $\phi_\infty = \phi_{\min}$.

(C1) *The initial value problem*

$$\begin{cases} \dfrac{\mathrm{d}\phi(\boldsymbol{\xi})}{\mathrm{d}t} = (\gamma(\boldsymbol{\xi}))^2 \sum_{i=1}^n \left(y_i \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}} - \left[\phi * \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\cdot)} \right] (\boldsymbol{\xi}) \right), \\ \phi(0) = \phi_{\mathrm{ini}}. \end{cases}$$

(C2) *The minimization problem*

$$\min_{\phi-\phi_{\mathrm{ini}} \in H_\Gamma} \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2} |\phi(\boldsymbol{\xi}) - \phi_{\mathrm{ini}}(\boldsymbol{\xi})|^2 \mathrm{d}\boldsymbol{\xi}$$

$$\text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \ldots, n.$$

We remark that $\mathcal{P}^*\mathcal{P}\phi = \sum_{i=1}^n \mathcal{F}[h\delta_{\boldsymbol{x}_i}]$, where $\delta_{\boldsymbol{x}_i}(\cdot) = \delta(\cdot - \boldsymbol{x}_i)$, $i = 1, \ldots, n$. Therefore problem (C1) can also be written as

$$\begin{cases} \dfrac{\mathrm{d}\mathcal{F}[h]}{\mathrm{d}t} = \gamma^2 \sum_{i=1}^n (y_i \mathcal{F}[\delta_{\boldsymbol{x}_i}] - \mathcal{F}[h\delta_{\boldsymbol{x}_i}]), \\ \mathcal{F}[h](\mathbf{0}) = \mathcal{F}[h]_{\mathrm{ini}}. \end{cases}$$

In the following, we study the discretized version of this dynamics-optimization problem ((C1) and (C2)).

Corollary 6. *Let $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$ be a positive function defined on lattice $\mathbb{Z}^d$ and $\phi = \mathcal{F}[h]$. The operator $\Gamma : \ell^2(\mathbb{Z}^d) \to \ell^2(\mathbb{Z}^d)$ is defined by $[\Gamma\phi](\boldsymbol{k}) = \gamma(\boldsymbol{k})\phi(\boldsymbol{k})$, $\boldsymbol{k} \in \mathbb{Z}^d$. Here $\ell^2(\mathbb{Z}^d)$ is set of square summable functions on the lattice $\mathbb{Z}^d$. Define the Hilbert space $H_\Gamma := \mathrm{Im}(\Gamma)$. Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal \in \mathbb{T}^{n\times d}$, $Y = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$, and $\mathcal{P} : H_\Gamma \to \mathbb{R}^n$ be a surjective operator such as*

$$P : \phi \mapsto \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} \phi(\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_1^\intercal \boldsymbol{k}}, \ldots, \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \phi(\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_n^\intercal \boldsymbol{k}} \right)^\intercal. \tag{5.5}$$

Then the following two problems are equivalent in the sense that $\phi_\infty = \phi_{\min}$.

(D1) *The initial value problem*

$$\begin{cases} \dfrac{\mathrm{d}\phi(\boldsymbol{k})}{\mathrm{d}t} = (\gamma(\boldsymbol{k}))^2 \sum_{i=1}^n \left(y_i \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{k}} - \left[\phi * \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\cdot)} \right] (\boldsymbol{k}) \right), \\ \phi(\boldsymbol{0}) = \phi_{\mathrm{ini}}. \end{cases}$$

(D2) *The minimization problem*

$$\min_{\phi - \phi_{\mathrm{ini}} \in H_\Gamma} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\phi(\boldsymbol{k}) - \phi_{\mathrm{ini}}(\boldsymbol{k})|^2$$

$$\text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \ldots, n.$$

5.2. Example: Explicitizing the implicit bias for two-layer ReLU NNs. As an example, by Corollary 5, we derive the following constrained optimization problem explicitly minimizing an FP-norm (see the next section), whose solution is the same as the long-time limit solution of the simplified LFP model (5.1), that is,

$$\min_{h - h_{\mathrm{ini}} \in F_\gamma} \int_{\mathbb{R}^d} \left(\mathbb{E}_{a,r} \left[\frac{r^3}{16\pi^4 \|\boldsymbol{\xi}\|^{d+3}} + \frac{a^2 r}{4\pi^2 \|\boldsymbol{\xi}\|^{d+1}} \right] \right)^{-1} |\mathcal{F}[h](\boldsymbol{\xi}) - \mathcal{F}[h_{\mathrm{ini}}](\boldsymbol{\xi})|^2 \mathrm{d}\boldsymbol{\xi}, \tag{5.6}$$

subject to constraints $h(\boldsymbol{x}_i) = y_i$ for $i = 1, \ldots, n$. The F_γ is defined in the next section. This explicit penalty indicates that the learning of DNN is biased toward functions with more power at low frequencies, which was speculated in previous works [25, 27]. For 1-d problems ($d = 1$), when the $1/\xi^2$ term dominates, the minimization of the FP-norm is equivalent to the spatial domain minimization problem $\min_{h(x)} \int_{\mathbb{R}} |h'(x) - h_{\mathrm{ini}}{}'(x)|^2 \mathrm{d}x$ subject to constraints $h(x_i) = y_i$ for $i = 1, \ldots, n$, which yields a linear spline interpolation for $h_{\mathrm{ini}}(x) = 0$ [9]. Similarly, when $1/\xi^4$ dominates, the minimization of the FP-norm is equivalent to the spatial domain minimization problem $\min_{h(x)} \int_{\mathbb{R}} |h''(x) - h_{\mathrm{ini}}{}''(x)|^2 \mathrm{d}x$ subject to constraints $h(x_i) = y_i$ for $i = 1, \ldots, n$, which yields a cubic spline interpolation for $h_{\mathrm{ini}}(x) = 0$ [9]. In general, the two power laws of frequency usually coexist, thus, leading to a specific mixture of linear and cubic splines. For high dimensional problems, the minimization problem is difficult to interpret by a specific interpolation because the order of differentiation depends on d and can be fractal.

6. FP-norm and an a priori generalization error bound. The equivalent explicit optimization problem (5.6) provides a way to analyze the generalization of sufficiently wide two-layer NNs. We consider the Fourier domain with discretized frequencies. Then, we begin with the definition of an FP-norm, which naturally induces an FP-space containing all possible solutions of a target NN, whose Rademacher complexity can be controlled by the FP-norm of the target function. Thus we obtain an a priori estimate of the generalization error of NN by the theory of Rademacher complexity. Our a priori estimate follows the Monte Carlo error rates with respect to the sample size. Importantly, our estimate unravels how frequency components of the target function affect the generalization performance of DNNs.

6.1. Problem setup. We focus on the regression problem. Assume the target function $f : \Omega := [0,1]^d \to \mathbb{R}$. Let the training set be $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i$'s are independently sampled from an underlying distribution $\mathcal{D}(\boldsymbol{x})$ and $y_i = f(\boldsymbol{x}_i)$. We consider the square loss

$$\ell(h, \boldsymbol{x}, y) = |h(\boldsymbol{x}) - y|^2 \tag{6.1}$$

with population risk

$$R_{\mathcal{D}}(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\ell(h, \boldsymbol{x}, f(\boldsymbol{x})) \tag{6.2}$$

and empirical risk

$$R_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, \boldsymbol{x}_i, y_i). \tag{6.3}$$

6.2. FP-space. The quantity in the minimization problem motivates a definition of FP-norm. This FP-norm then leads to the definition of the function space where the solution of the minimization problem lies. We denote $\mathbb{Z}^{d*} := \mathbb{Z}^d \backslash \{\mathbf{0}\}$. Given a frequency weight function $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$ or $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ satisfying

$$\|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} (\gamma(\boldsymbol{k}))^2\right)^{\frac{1}{2}} < +\infty \quad \text{or} \quad \|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2\right)^{\frac{1}{2}} < +\infty, \tag{6.4}$$

we define the FP-norm for all functions $h \in L^2(\Omega)$:

$$\|h\|_\gamma := \|\mathcal{F}[h]\|_{H_\Gamma} = \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h](\boldsymbol{k})|^2\right)^{\frac{1}{2}}. \tag{6.5}$$

If $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ is not defined at $\boldsymbol{\xi} = \mathbf{0}$, we set $(\gamma(\mathbf{0}))^{-1} := 0$ in the above definition and $\|\cdot\|_\gamma$ is only a seminorm of h.

Then we define the FP-space

$$\mathcal{F}_\gamma(\Omega) = \{h \in L^2(\Omega) : \|h\|_\gamma < \infty\}. \tag{6.6}$$

Clearly, for any γ, the FP-space is a subspace of $L^2(\Omega)$. In addition, if $\gamma : \boldsymbol{k} \mapsto \|\boldsymbol{k}\|^{-r}$ for $\boldsymbol{k} \in \mathbb{Z}^{d*}$, then functions in the FP-space with $\mathcal{F}[h](\mathbf{0}) = \int_\Omega h(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ form the Sobolev space $H^r(\Omega)$. Note that in the case of DNN, according to the F-Principle, $(\gamma(\boldsymbol{k}))^{-2}$ increases with the frequency. Thus, the contribution of high frequency to the FP-norm is more significant than that of low frequency.

6.3. A priori generalization error bound. Next, we show the upper bound of the FP-norm of a function leads to an upper bound of the Rademacher complexity of the function space. The Rademacher complexity is defined as

$$\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\tau_i h(\boldsymbol{x}_i)\right] \tag{6.7}$$

for the function space $\mathcal{H}$ and dataset $S = \{\boldsymbol{x}_i, h(\boldsymbol{x}_i)\}_{i=1}^n$.

Lemma 5. (i) *For* $\mathcal{H}_Q = \{h : \|h\|_\gamma \le Q\}$ *with* $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, *we have*

$$\mathrm{Rad}_S(\mathcal{H}_Q) \le \frac{1}{\sqrt{n}} Q\|\gamma\|_{\ell^2}. \tag{6.8}$$

(ii) *For* $\mathcal{H}_Q{}' = \{h : \|h\|_\gamma \le Q, |\mathcal{F}[h](\mathbf{0})| \le c_0\}$ *with* $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ *and* $\gamma^{-1}(\mathbf{0}) := 0$, *we have*

$$\mathrm{Rad}_S(\mathcal{H}_Q{}') \le \frac{c_0}{\sqrt{n}} + \frac{1}{\sqrt{n}} Q\|\gamma\|_{\ell^2}. \tag{6.9}$$

Then, we prove that the target function can be used to bound the FP-norm of the solution of the minimization problem.

Lemma 6. *Suppose that the real-valued target function* $f \in \mathcal{F}_\gamma(\Omega)$ *and that the training dataset* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ *satisfies* $y_i = f(\boldsymbol{x}_i)$, $i = 1, \dots, n$. *If* $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, *then there exists a unique solution* h_n *to the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma \quad \text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \dots, n. \tag{6.10}$$

Moreover, we have

$$\|h_n - h_{\mathrm{ini}}\|_\gamma \le \|f - h_{\mathrm{ini}}\|_\gamma. \tag{6.11}$$

Lemma 7. *Suppose that the real-valued target function* $f \in \mathcal{F}_\gamma(\Omega)$ *and that the training dataset* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ *satisfies* $y_i = f(\boldsymbol{x}_i)$, $i = 1, \dots, n$. *If* $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ *with* $\gamma^{-1}(\mathbf{0}) := 0$, *then there exists a solution* h_n *to the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma \quad \text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \dots, n. \tag{6.12}$$

Moreover, we have

$$|\mathcal{F}[h_n - h_{\mathrm{ini}}](\mathbf{0})| \le \|f - h_{\mathrm{ini}}\|_\infty + \|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2}. \tag{6.13}$$

Based on the above analysis, we derive an a priori generalization error bound of the minimization problem.

Theorem 3 (a priori generalization error bound). *Suppose that the real-valued target function* $f \in \mathcal{F}_\gamma(\Omega)$, *the training dataset* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ *satisfies* $y_i = f(\boldsymbol{x}_i)$, $i = 1, \dots, n$, *and* h_n *is the solution of the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma \quad \text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \dots, n. \tag{6.14}$$

Then we have

(i) *given* $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, *for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$ *over the random training sample, the population risk has the bound*

$$R_{\mathcal{D}}(h_n) \le \|f - h_{\rm ini}\|_\gamma \|\gamma\|_{\ell^2} \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right), \tag{6.15}$$

(ii) *given* $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ *with* $\gamma(\mathbf{0})^{-1} := 0$, *for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$ *over the random training sample, the population risk has the bound*

$$R_{\mathcal{D}}(h_n) \le (\|f - h_{\rm ini}\|_\infty + 2\|f - h_{\rm ini}\|_\gamma \|\gamma\|_{\ell^2}) \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right). \tag{6.16}$$

Remark 5. By the assumption in the theorem, the target function f belongs to $\mathcal{F}_\gamma(\Omega)$, which is a subspace of $L^2(\Omega)$. In most applications, f is also a continuous function. In any case, f can be well approximated by a large NN due to the universal approximation theory [10].

Our a priori generalization error bound in Theorem 3 is large if the target function possesses significant high-frequency components. Thus, it explains the failure of DNNs in generalization for learning the parity function [21], whose power concentrates at high frequencies. In the following, We use experiments to illustrate that, as predicted by our a priori generalization error bound, a larger FP-norm of the target function indicates a larger generalization error.

7. Numerical experiments.

In this section, we conduct numerical experiments to validate the effectiveness of the LFP model for two-layer ReLU and tanh networks. In addition, we show that, with sufficient samples, the test error still increases as the frequency of the target function increases.

7.1. Numerically solve the LFP optimization problem.

Numerically, we solve the LFP model (D2) by solving the following problem:[2]

$$\min_{a_n,b_n} \sum_{i=1}^{M} \left(\sum_{j\in I} \left[a_j \sin\left(2\pi \frac{j}{L'} x_i\right) + b_j \cos\left(2\pi \frac{j}{L'} x_i\right) \right] - y_i \right)^2 + \varepsilon \sum_{j\in I} \gamma\left(2\pi \frac{j}{L'}\right)^{-2} \left(a_j^2 + b_j^2\right), \tag{7.1}$$

where we set $I = \left\{0, \dots, \frac{L'}{L}K - 1\right\}$, $L' = 10L$, L is the range of the training inputs, $K = 200$ which is much larger than the number of training samples, and $\varepsilon = 10^{-6}$. Denote $M_I = \frac{L'}{L}K - 1$. We can rewrite the above problem into the vector form

$$\min_{\boldsymbol{a}} (\boldsymbol{E}\boldsymbol{a} - \boldsymbol{Y})^{\intercal} (\boldsymbol{E}\boldsymbol{a} - \boldsymbol{Y}) + \varepsilon \boldsymbol{a}^{\intercal} \boldsymbol{W}^{-1} \boldsymbol{a}, \tag{7.2}$$

[2]The code can be found at https://github.com/xuzhiqin1990/LFP.

where

$$\boldsymbol{a} = [a_0, \dots, a_{M_I}, b_0, \dots, b_{M_I}]^\intercal,$$
$$\boldsymbol{E} = \left[\sin\left(2\pi\frac{0}{L'}\boldsymbol{X}\right), \dots, \sin\left(\frac{2\pi}{L'}M_I\boldsymbol{X}\right), \cos\left(2\pi\frac{0}{L'}\boldsymbol{X}\right), \dots, \cos\left(\frac{2\pi}{L'}M_I\boldsymbol{X}\right)\right],$$
$$\boldsymbol{X} = [x_1, \dots, x_M]^\intercal, \quad \boldsymbol{Y} = [y_1, \dots, y_M]^\intercal,$$
$$\boldsymbol{W}^{-1} = \operatorname{diag}\left\{\gamma\left(2\pi\frac{0}{L'}\right)^{-2}, \dots, \gamma\left(2\pi\frac{1}{L'}\left(\frac{L'}{L}K-1\right)\right)^{-2}\right\}.$$

The solution of the above problem satisfies

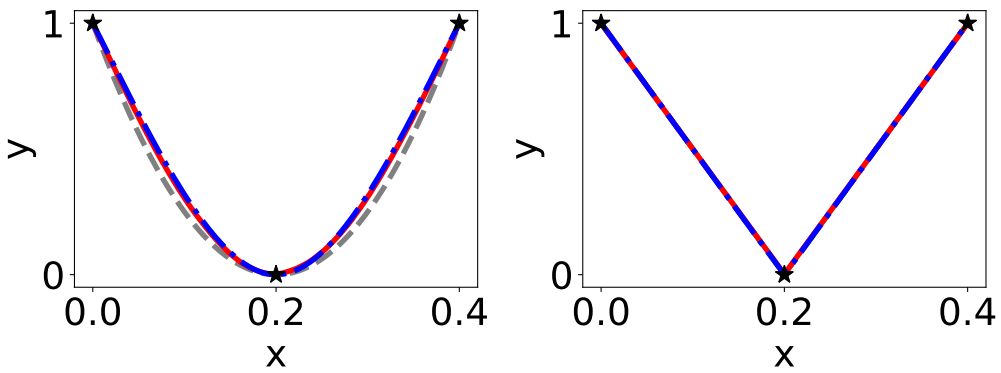
$$\boldsymbol{E}^\intercal(\boldsymbol{E}\boldsymbol{a}-\boldsymbol{Y}) + \varepsilon\boldsymbol{W}^{-1}\boldsymbol{a} = 0. \tag{7.3}$$

Then $\boldsymbol{a}$ is solved as

$$\boldsymbol{a} = \left[\boldsymbol{E}^\intercal\boldsymbol{E} + \varepsilon\boldsymbol{W}^{-1}\right]^{-1}\boldsymbol{E}^\intercal\boldsymbol{Y}. \tag{7.4}$$

7.2. The effectiveness of LFP model. Without the last term in (4.19) arising from the evolution of $\boldsymbol{w}$, we show that the simplified LFP model in (5.1) can still predict the learning results of two-layer wide NNs, trained by full-batch gradient descent with learning rate 10^{-5}.

For the 1-d input example, we use a ReLU NN with 10000 hidden neurons and different initializations. Three data points are used as the training dataset, shown in Figure 2(a). The initialization of all parameters is sampled from uniform distributions with zero mean. Denote the sampling interval by $[-U, U]$. To make the term of $1/\xi^4$ dominate, we set U for initializing $\boldsymbol{w}$ as 3, for initializing a as 0.01, and for initializing the bias term as 3. Note that the bias terms are not initialized by too large values. As shown in Figure 2(a), the NN



(a) (b)

Figure 2. *f_{NN} (red solid) versus f_{LFP} (blue dashed dot) versus splines (gray dashed, cubic spline for* (a) *and linear spline for* (b)*) for a 1-d problem. All curves nearly overlap with one other. Two-layer ReLU NN of* 10000 *hidden neurons is initialized with* (a) $\langle r^2\rangle_r \gg \langle a^2\rangle_a$, *and* (b) $\langle r^2\rangle_r \ll \langle a^2\rangle_a$. *Black stars indicate training data.*

interpolates training data by a smooth function (denoted by f_{NN}, red solid), which nearly overlaps with the prediction of the LFP model (denoted by f_{LFP}, blue dashed) and the cubic spline interpolation (gray dashed). On the contrary, to make the term of $1/\xi^2$ dominate, we set U for initializing $\boldsymbol{w}$ as 0.1, for initializing a as 2, and for initializing the bias term as 2. As shown in Figure 2(b), the NN interpolates training data by a function, which nearly overlaps with the prediction of the LFP model and the linear spline interpolation.

Theorem 1 shows that the two-layer wide NN in the linear regime can be characterized by an LFP dynamics, whose long-time solution is equivalent to the solution of an LFP optimization problem. These numerical experiments show that solutions obtained by the long-time training NN are very close to the ones obtained by numerically solving the LFP optimization problem. Therefore, these results are consistent with the above analysis. To show the numerical results are rather general, we show another eight examples in Figure SM1 in the supplement, where the parameters and labels are randomly generated. These results show that the solutions obtained by the LFP model are very consistent with the ones obtained by training two-layer NNs. We then perform numerical experiments for 2-d input and the tanh activation function.

For the 2-d input example, we use a ReLU NN of 8000 hidden neurons to solve the XOR problem, which cannot be solved by one-layer NNs [18]. The training samples consist of four points represented by black stars in Figure 3(a). Similarly, we use uniform distribution for initialization and set U for initializing $\boldsymbol{w}$ as 0.8, for initializing a as 0.2, and for initializing the bias term as 1. The NN output, which can fit the training data well, over $[-1,1]^2$ is shown in Figure 3(a). Our LFP model accurately predicts outputs of the well-trained NN over the input domain $[-1,1]^2$ as shown in Figure 3(b).

For the two-layer tanh NN, we use the same setting as the cases in Figure 2 except for the tanh activation function. In this case, the weight coefficient decays exponentially w.r.t. the frequency no matter which part dominates; thus, the NN always learns the training data by a smooth function, as shown in Figure 4.

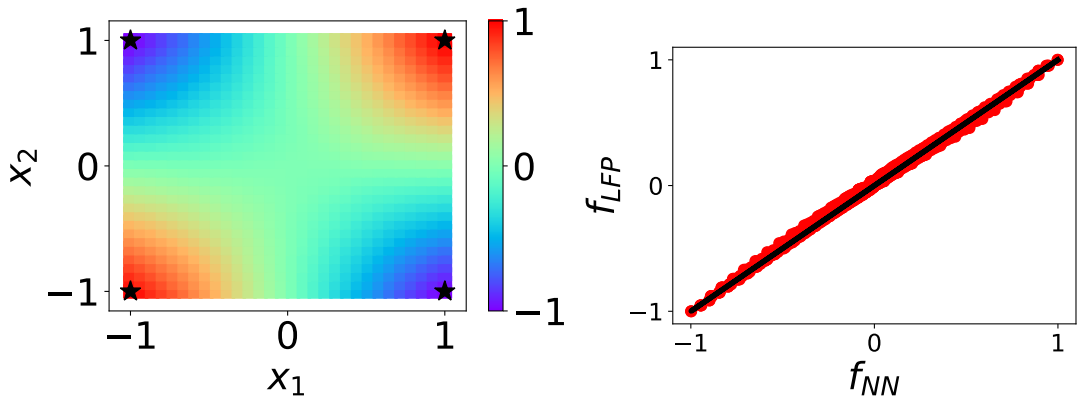


Figure 3. *2-d XOR problem with four training data indicated by black stars learned by a two-layer ReLU NN of* 8000 *hidden neurons.* (a) f_{NN} *illustrated in color scale.* (b) f_{LFP} *(ordinate) versus* f_{NN} *(abscissa) represented by red dots evaluated over whole input domain* $[-1,1]^2$. *The black line indicates the identity function.*

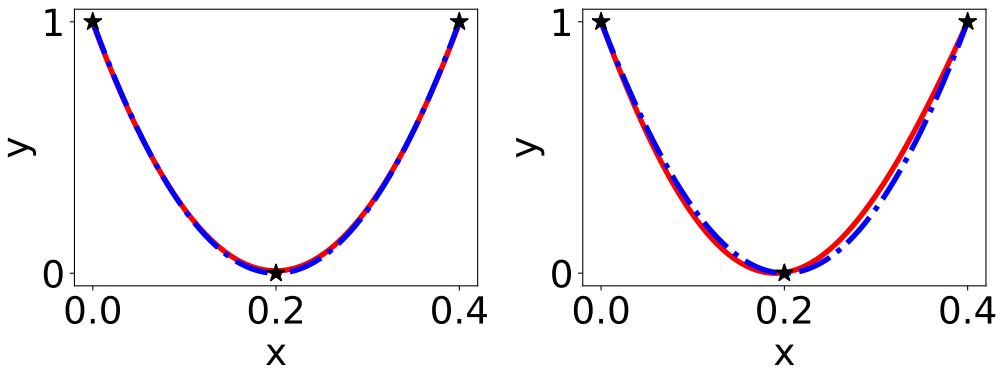


Figure 4. *f_{NN} (red solid) versus f_{LFP} (blue dashed dot) for a 1-d problem. Two-layer tanh NN of* 10000 *hidden neurons is initialized with* (a) $\langle r^2\rangle_r \gg \langle a^2\rangle_a$, *and* (b) $\langle r^2\rangle_r \ll \langle a^2\rangle_a$. *Black stars indicates training data.*

7.3. Generalization error. In this subsection, we use NNs to fit a series of target functions with different frequencies in different trials to see how the generalization error of the NN depends on the frequency of the target function. Note that we train all NNs to a stage where the training loss is sufficiently small and we compare their generalization error only on test samples but not their training speeds in this subsection. The generalization error bound in Theorem 3 is larger as the FP-norm increases, which implies that the generalization error would increase as the frequency of the target function increases. The following experiments of fitting 1-d sinusoidal functions is consistent with this implication.

In each trial, by full-batch gradient descent training, we train a ReLU-NN of width 1-5000-1 to fit 20 uniform samples of $f(x) = \sin(2\pi v x)$ on $[0, 1]$ until the training mean square error loss is smaller than 10^{-6}, where v is the frequency. In each trial, the target function has only a single frequency. The number of training samples is sufficient to recover the frequency of the target function by the Nyquist sampling theorem. We then use 500 uniform samples to test the NN. As the frequency of the target function increases, the FP-norm would increase, thus leading to a looser bound of the generalization error. As shown in Figure 5, the test error increases as the frequency of the target function increases.

8. Discussion. In this work, inspired by the F-Principle, we derive an LFP model for two-layer wide NNs—a model that quantitatively well predicts the output of two-layer ReLU or tanh NNs in an extremely overparameterized regime. We explicitize the implicit bias of the F-Principle by a constrained optimization problem equivalent to the LFP model. This explicitization leads to an a priori estimate of the generalization error bound, which depends on the FP-norm of the target function. Note that our LFP model for other transfer functions can also be derived similarly.

The LFP model advances our qualitative/empirical understandings of the F-Principle to a quantitative level. (i) With the ASI trick [30] offsetting the initial DNN output to zero, the LFP model indicates that the F-Principle also holds for DNNs initialized with large weights. Therefore, "initialized with small parameters" [25, 27] is not a necessary condition for the F-Principle. (ii) Based on the training behavior of the F-Principle, previous works [25, 27]

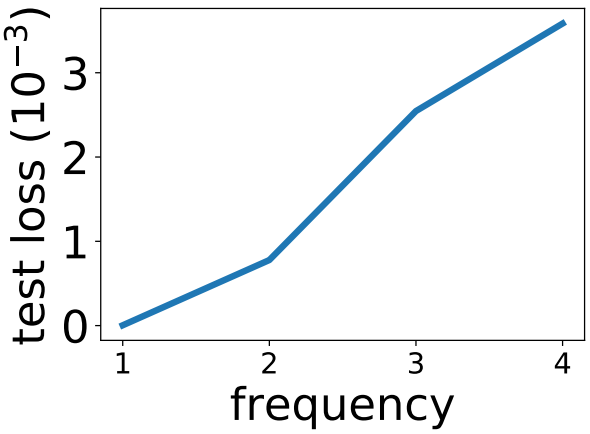


Figure 5. *Test loss plotted as a function of frequency v of the target function* $\sin(2\pi vx)$.

speculated that DNNs prefer to learn the training data by a low frequency function. With an equivalent optimization problem explicitizing the F-Principle, this speculation is demonstrated theoretically by the LFP model.

Our a priori generalization error bound increases as the FP-norm of the target function increases. This explains several important phenomena. First, DNNs fail to generalize well for the parity function [21]. [25] shows that this is due to the inconsistency between the high-frequency-dominant property of the parity function and the low-frequency preference of DNNs. In this work, by our a priori generalization error bound, the dominant high frequency of the parity function quantitatively results in a large FP-norm and, thus, a large generalization error. Second, because randomly labeled data possesses large high-frequency components, which induces a large FP-norm of any function that well matches the training data and test data, we expect a very large generalization error, e.g., no generalization, as observed in experiments. Intuitively, our estimate indicates good generalization of NNs for a well-structured low-frequency-dominant real dataset as well as bad generalization of NNs for randomly labeled data, thus providing insight into the well-known puzzle of generalization of DNNs [29].

The F-Principle, a widely observed implicit bias of DNNs, is also a natural bias for humans. Empirically, when humans see several points of training data, without a specific prior, they tend to interpolate these points by a low-frequency-dominant function. Therefore, the success of DNN may partly result from its adoption of a similar interpolation bias as a human's. In general, there could be multiple types of implicit biases underlying the training dynamics of a DNN. Inspired by the LFP model, discovering and explicitizing these implicit biases could be a key step toward a thorough quantitative understanding of deep learning.

REFERENCES

[1] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, *Fine-grained Analysis of Optimization and Generalization for Overparameterized Two-layer Neural Networks*, preprint, arXiv:1901.08584, 2019.

[2] D. Arpit, S. Jastrzbski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., *A closer look at memorization in deep networks*,

in Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, pp. 233–242.

[3] R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman, *Frequency Bias in Neural Networks for Input of Non-uniform Density*, preprint, arXiv:2003.04560, 2020.

[4] R. Basri, D. Jacobs, Y. Kasten, and S. Kritchman, *The convergence rate of neural networks for learned functions of different frequencies*, in Advances in Neural Information Processing Systems, 2019, pp. 4763–4772.

[5] S. Biland, V. C. Azevedo, B. Kim, and B. Solenthaler, *Frequency-aware Reconstruction of Fluid Simulations with Generative Networks*, preprint, arXiv:1912.08776, 2019.

[6] B. Bordelon, A. Canatar, and C. Pehlevan, *Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks*, preprint, arXiv:2002.02561, 2020.

[7] W. Cai, X. Li, and L. Liu, *A phase shift deep neural network for high frequency approximation and wave problems*, J. Sci. Comput, 42 (2020), pp. A3285–A3312.

[8] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, *Towards Understanding the Spectral Bias of Deep Learning*, arXiv:1912.01198 [cs, stat], 2019.

[9] P. Craven and G. Wahba, *Smoothing noisy data with spline functions*, Numer. Math., 31 (1978), pp. 377–403.

[10] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Syst., 2 (1989), pp. 303–314.

[11] W. E, C. Ma, and L. Wu, *A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics*, Sci. China Math., 63 (2020), pp. 1235–1258.

[12] W. E, C. Ma, and L. Wu, *Machine learning from a continuous viewpoint,* I, Sci. China Math., 63 (2020), pp. 2233–2266.

[13] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis, *Adaptive activation functions accelerate convergence in deep and physics-informed neural networks*, J. Comput. Phys., 404 (2020), 109136, https://doi.org/10.1016/j.jcp.2019.109136.

[14] Z. Liu, W. Cai, and Z.-Q. J. Xu, *Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains,* Commun. Comput. Phys., 28 (2020), pp. 1970–2001.

[15] T. Luo, Z. Ma, Z.-Q. J. Xu, and Y. Zhang, *Theory of the Frequency Principle for General Deep Neural Networks*, preprint, arXiv:1906.09235, 2019.

[16] C. Ma, L. Wu, and W. E, *The slow deterioration of the generalization error of the random feature model*, in Mathematical and Scientific Machine Learning, PMLR 107, 2020, pp. 373–389.

[17] C. Mingard, J. Skalse, G. Valle-Pérez, D. Martínez-Rubio, V. Mikulik, and A. A. Louis, *Neural Networks Are A Priori Biased Towards Boolean Functions with Low Entropy*, preprint, arXiv:1609.08403, 2019.

[18] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, MA, 2017.

[19] P. Nakkiran, G. Kaplun, D. Kalimeris, T. Yang, B. L. Edelman, F. Zhang, and B. Barak, *SGD on neural networks learns functions of increasing complexity*, in Advances in Neural Information Processing Systems, 2019, pp. 3491–3501.

[20] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. *On the spectral bias of neural networks*, in Proceedings of the International Conference on Machine Learning, 2019, pp. 5301–5310.

[21] S. Shalev-Shwartz, O. Shamir, and S. Shammah, *Failures of gradient-based deep learning*, in Proceedings of the International Conference on Machine Learning, 2017, pp. 3067–3075.

[22] G. Valle-Perez, C. Q. Camargo, and A. A. Louis, *Deep learning generalizes because the parameter-function map is biased towards simple functions*, in Proceedings of the International Conference on Learning Representations, 2019.

[23] Z. J. Xu, *Understanding Training and Generalization in Deep Learning by Fourier Analysis*, preprint, arXiv:1808.04295, 2018.

[24] Z.-Q. J. Xu, Y. Zhang, and T. Luo, *Overview Frequency Principle/Spectral Bias in Deep Learning*, preprint, arXiv:2201.07395, 2022.

[25] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, *Frequency principle: Fourier analysis sheds light on deep neural networks*, Commun. Comput. Phys., 28 (2020), pp. 1746-1767.

[26] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, *Training behavior of deep neural network in frequency domain*, in Neural Information Processing, Lecture Notes in Comput. Sci. 11953, Springer, New York, 2019, pp. 264–274, https://doi.org/10.1007/978-3-030-36708-4_22.

[27] G. Yang and H. Salman, *A Fine-grained Spectral Perspective on Neural Networks*, preprint, arXiv:1907.10599, 2019.

[28] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, in Proceedings of the International Conference on Learning Representations, 2017.

[29] Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma, *Explicitizing an Implicit Bias of the Frequency Principle in Two-layer Neural Networks*, preprint, arXiv:1905.10264, 2019.

[30] Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma, *A type of generalization error induced by initialization in deep neural networks*, in Mathematical and Scientific Machine Learning, PMLR, 2020, pp. 144–164.