Explicitizing an Implicit Bias of the Frequency Principle in Two-layer Neural Networks

Yaoyu Zhang
New York University Abu Dhabi and Courant Institute of Mathematical Sciences

Zhi-Qin John Xu*
New York University Abu Dhabi and Courant Institute of Mathematical Sciences

Tao Luo
Department of Mathematics, Purdue University

Zheng Ma
Department of Mathematics, Purdue University

Abstract

It remains a puzzle that why deep neural networks (DNNs), with more parameters than samples, often generalize well. An attempt of understanding this puzzle is to discover implicit biases underlying the training process of DNNs, such as the Frequency Principle (F-Principle), i.e., DNNs often fit target functions from low to high frequencies. Inspired by the F-Principle, we propose an effective model of linear F-Principle (LFP) dynamics which accurately predicts the learning results of two-layer ReLU neural networks (NNs) of large widths. This LFP dynamics is rationalized by a linearized mean field residual dynamics of NNs. Importantly, the long-time limit solution of this LFP dynamics is equivalent to the solution of a constrained optimization problem explicitly minimizing an FP-norm, in which higher frequencies of feasible solutions are more heavily penalized. Using this optimization formulation, an *a priori* estimate of the generalization error bound is provided, revealing that a higher FP-norm of the target function increases the generalization error. Overall, by explicitizing the implicit bias of the F-Principle as an explicit penalty for two-layer NNs, our work makes a step towards a quantitative understanding of the learning and generalization of general DNNs.

1 Introduction

The wide success of deep learning in many fields (LeCun et al. 2015) remains a mystery. For example, a puzzle recently attracts a lot of attention, that is, why Deep Neural Networks (DNNs), with more parameters than samples, often generalize well (Zhang et al. 2016). A major difficulty of resolving this puzzle may be attributed to the lack of an effective model which can accurately predict the final output function of DNNs and yet is simple enough for analysis. Devising such an effective model could propel deep learning into a new era in which quantitative understandings of deep learning replace the qualitative or empirical ones.

Towards this end, we begin with a widely observed phenomenon of DNNs, that is, Frequency Principle (F-Principle) (Xu et al. 2018, Rahaman et al. 2018, Xu et al. 2019):

*Corresponding author: zhiqinxu@nyu.edu

Preprint. Under review.

DNNs initialized with small parameters often fit target functions from low to high frequencies during the training.

Without an explicit mathematical description, it is unclear how this implicit bias of the F-Principle functions quantitatively during the training. Inspired by the F-Principle, we construct a model of linear F-Principle (LFP) dynamics, which explicitly imposes different priorities on different frequencies in the gradient flow dynamics. Experimentally, we show that the LFP model can accurately predict the output of two-layer ReLU neural networks (NNs) of large widths. We then rationalize the LFP model using a linearized mean field residual dynamics of DNNs, which is widely considered in recent theoretical studies of DNNs (Mei et al. 2018, Rotskoff & Vanden-Eijnden 2018, Mei et al. 2019). We prove that the long-time limit solution of this LFP dynamics is equivalent to the solution of a constrained optimization problem minimizing an F-Principle norm (FP-norm), in which higher frequencies of feasible solutions are more heavily penalized. Therefore, by analyzing the explicit regularity underlying the FP-norm, we can obtain a quantitative understanding of the behavior of two-layer NNs.

With a reasonable construction process and an ability of making accurate predictions for two-layer ReLU NNs of large widths, the LFP model qualifies as a primitive candidate of an effective model of DNNs, which is capable of providing quantitative understandings of deep learning. To analyze the generalization error of the LFP model, we first use the FP-norm, the explicit penalty, to induce an FP function space, and estimate its Rademacher complexity. We then provide an *a priori* estimate, i.e., an estimate without the knowledge of the model solution, of the generalization error for the LFP model, which is bounded by the FP-norm of the target function, scales as $1/\sqrt{n}$ as the number of training samples n increases, and is independent of the number of parameters in NNs.

2 Related works

Various approaches have been proposed in an attempt to resolve the generalization puzzle of DNNs. For example, the generalization error has been related to various complexity measures (Bartlett et al. 1999, Bartlett & Mendelson 2002, Bartlett, Foster & Telgarsky 2017, Bartlett, Harvey, Liaw & Mehrabian 2017, Neyshabur et al. 2017, Golowich et al. 2017, Dziugaite & Roy 2017, Neyshabur et al. 2018, E et al. 2018), local properties (sharpness/flatness) of loss functions at minima (Hochreiter & Schmidhuber 1995, Keskar et al. 2016, Dinh et al. 2017, Wu et al. 2017), stability of optimization algorithms (Bousquet & Elisseeff 2002, Xu & Mannor 2012, Hardt et al. 2015), and implicit bias of the training process (Arpit et al. (2017), Rahaman et al. (2018), Xu (2018*a*,*b*), Pérez et al. (2018), Xu et al. (2019, 2018), Neyshabur et al. (2014), Poggio et al. (2018), Soudry et al. (2018)). Recently, analyzing DNNs in an extremely over-parameterized regime renders a promising approach. For example, the training process of two-layer neural networks at the mean-field limit can be described by a partial differential equation (Rotskoff & Vanden-Eijnden 2018, Mei et al. 2018, Sirignano & Spiliopoulos 2018). In addition, the training dynamics of a DNN in an extremely over-parameterized regime is found to be well approximated by the gradient flow of a linearized model of the DNN resembling kernel methods (Jacot et al. 2018, Lee et al. 2019). This result initiates a series of works. For example, Arora et al. (2019), Cao & Gu (2019), E, Ma, Wang & Wu (2019), E, Ma & Wu (2019) utilize the linearized model to study the generalization error bounds of DNN. Note that an *a priori* generalization error bound for two-layer NNs is provided in E et al. (2018), in which an explicit penalty is imposed to the loss function of NNs. In contrast, our *a prior* generalization bound works for NNs without any extra penalty.

3 Notation and experimental setup

3.1 Notation

For a two-layer neural network, its output (also known as the hypothesis function) reads as

$$f(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j=1}^{m} a_j \sigma\left(\boldsymbol{w}_j^{\intercal}\boldsymbol{x} - |\boldsymbol{w}_j| c_j\right), \tag{1}$$

where $\boldsymbol{w}_j, \boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{\theta} = (\boldsymbol{a}^{\intercal}, \boldsymbol{w}_1^{\intercal}, \ldots, \boldsymbol{w}_m^{\intercal}, \boldsymbol{c}^{\intercal})^{\intercal}$, $\boldsymbol{a}, \boldsymbol{c} \in \mathbb{R}^m$ and $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)^{\intercal} \in \mathbb{R}^{m\times d}$, and by default $\sigma(z) = \max(z, 0)$ ($z \in \mathbb{R}$) is the activation function of ReLU. The target function is

denoted by $f(\boldsymbol{x})$. In this work, we consider the mean-squared error (MSE) loss function

$$L = \int_{\mathbb{R}^d} \frac{1}{2} |f(\boldsymbol{x}, \boldsymbol{\theta}) - f(\boldsymbol{x})|^2 \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{2}$$

where $\rho(\boldsymbol{x})$ is a probability density. The following notation will be used in studying the training dynamics: $u(\boldsymbol{x}, t) = f(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f(\boldsymbol{x})$, $u_\rho(\boldsymbol{x}, t) = f_\rho(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f_\rho(\boldsymbol{x})$, where $f_\rho(\boldsymbol{x}, \boldsymbol{\theta}(t)) = f(\boldsymbol{x}, \boldsymbol{\theta}(t))\rho(\boldsymbol{x})$, $f_\rho(\boldsymbol{x}) = f(\boldsymbol{x})\rho(\boldsymbol{x})$. In this work, we focus on $\rho(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, which accounts for the real case of a finite training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with each input $\boldsymbol{x}_i \in \mathbb{R}^d$ and output $y_i \in \mathbb{R}$. Because the target function $f(\boldsymbol{x})$ is not available, without loss of generality, we fixed it to any continuous function satisfying $f(\boldsymbol{x}_i) = y_i$ for $i = 1, \cdots, n$. Then $f_\rho(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n y_i \delta(\boldsymbol{x} - \boldsymbol{x}_i)$. Because $\partial_t f(\boldsymbol{x}) = 0$ for any fixed $f(\boldsymbol{x})$, its choice does not affect the training dynamics of $f(\boldsymbol{x}, \boldsymbol{\theta}(t))$.

For any function g defined on $\mathbb{R}^d$, we use the following convention of the Fourier transform and its inverse:

$$\mathcal{F}[g](\boldsymbol{\xi}) = \int_{\mathbb{R}^d} g(\boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}, \quad g(\boldsymbol{x}) = \int_{\mathbb{R}^d} \mathcal{F}[g](\boldsymbol{\xi}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}^\intercal \boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \in \mathbb{R}^d$ denotes the frequency. For any function g defined on the torus $\Omega := \mathbb{T}^d = [0, 1]^d$, we use the following convention of the Fourier series and its inverse:

$$\mathcal{F}[g](\boldsymbol{\xi}) = \int_{\mathbb{T}^d} g(\boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}, \quad g(\boldsymbol{x}) = \sum_{\boldsymbol{\xi} \in \mathbb{Z}^d} \mathcal{F}[g](\boldsymbol{\xi}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}^\intercal \boldsymbol{\xi}},$$

where $\boldsymbol{\xi} \in \mathbb{Z}^d$ denotes the frequency. For u and u_ρ, their Fourier transforms are written as $\mathcal{F}[u](\boldsymbol{\xi}, t)$ and $\mathcal{F}[u_\rho](\boldsymbol{\xi}, t)$, respectively. $\boldsymbol{\theta}(t) = (\boldsymbol{a}^\intercal(t), \boldsymbol{w}_1^\intercal(t), \ldots, \boldsymbol{w}_m^\intercal(t), \boldsymbol{c}^\intercal(t))^\intercal$ are the parameters at training time t.

3.2 Experimental setup

In our experiments, we use two-layer ReLU NNs of form $f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^m a_j \sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x} - |\boldsymbol{w}_j| c_j)$ for input dimension $d > 1$ and $f(x, \boldsymbol{\theta}) = \sum_{j=1}^m a_j \sigma(w_j(x - c_j))$ for $d = 1$. The NNs are trained with MSE loss and full batch size. The learning rate for Fig. 1 and 2 is 3×10^{-5}, for Fig. 3 is 10^{-4}. The training algorithm for Fig. 1 and 2 is gradient descent, for Fig. 3 is Adam (Kingma & Ba 2014). c_i's are initialized by a uniform distribution on $[-1, 1]$ for Fig. 1 and on $[-4, 4]$ for Fig. 2. For Fig. 2, a_i's and $\boldsymbol{w}_i$'s are initialized by $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 0.49)$. For Fig. 3, We use an NN of 5000 hidden neurons initialized by Xavier normal initialization.

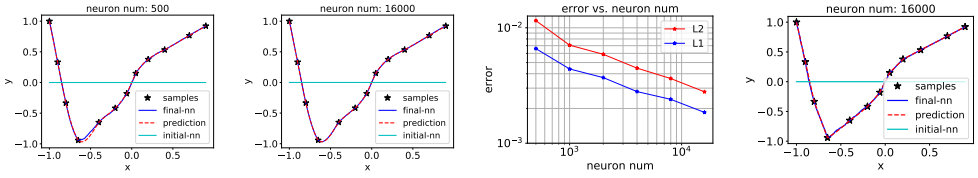
A random non-zero initial output of DNN leads to a specific type of generalization error. To eliminate this error, we use DNNs with an antisymmetrical initialization (ASI) trick (Zhang et al. 2019).

4 An effective model of Linear F-Principle (LFP) dynamics

It is difficult to analyze DNNs theoretically due to its huge number of parameters and highly nonlinear dynamics. In this section, inspired by the F-Principle, we propose a Linear F-Principle (LFP) dynamics to effectively model a two-layer ReLU NN of a large width. Specifically, with the loss of MSE, up to a multiplicative constant in time scale, we model the gradient descent dynamics of the two-layer NN of a sufficiently large width m as

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = -\left(\frac{\frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 + a_j(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 \frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 a_j(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+1}} \right) \mathcal{F}[u_\rho](\boldsymbol{\xi}, t), \tag{3}$$

where, different from $\mathcal{F}[u](\boldsymbol{\xi}, t)$ in the left hand side (LHS), $\mathcal{F}[u_\rho](\boldsymbol{\xi}, t)$ in the right hand side (RHS), which equals $\mathcal{F}(u(\cdot, t)\rho(\cdot)) = \mathcal{F}\left[\sum_{i=1}^n \left(f(\boldsymbol{x}_i, \boldsymbol{\theta}(t)) - y_i\right) \delta(\cdot - \boldsymbol{x}_i)\right]$, incorporates the information from the training dataset. In our numerical experiments, we only consider NNs with ASI trick(Zhang et al. 2019), which guarantees $f(\cdot, \boldsymbol{\theta}_0) = 0$. Note that, for $d = 1$, an NN of the form $f(x, \boldsymbol{\theta}) = \sum_{j=1}^m a_j \sigma(w_j(x - c_j))$ is modeled by the same LFP dynamics (3). For convenience, we refer to the long-time limit solution of the LFP dynamics as the solution of the LFP model. Intuitively, the coefficient as a function of $\boldsymbol{\xi}$ in the RHS characterizes a decaying priority of convergence for $\mathcal{F}[u](\boldsymbol{\xi}, t)$ from low to high frequencies, conforming with the phenomenon of the F-Principle (Xu et al. 2018, 2019).



(a) (b) (c) (d)

Figure 1: LFP model for 1-d training data. For (a, b, c), $1/|\boldsymbol{\xi}|^4$ dominates. For (d), $1/|\boldsymbol{\xi}|^2$ dominates. For (a, b, d), black stars: training samples; blue solid line: 800 uniformly spaced samples; red dashed lines: solutions of the corresponding LFP models; cyan solid curves: zero initial outputs of NNs. (c) $L^1(h_m, h_{\mathrm{LFP}})$ and $L^2(h_m, h_{\mathrm{LFP}})$ (mean of 10 trials) vs. neuron number.

Intuitively, a higher order of decay in the frequency domain, say $1/\|\boldsymbol{\xi}\|^{d+3}$ comparing to $1/\|\boldsymbol{\xi}\|^{d+1}$, leads to a more "smooth" solution. Therefore, adjusting the relative importance of $1/\|\boldsymbol{\xi}\|^{d+3}$ and $1/\|\boldsymbol{\xi}\|^{d+1}$ through their coefficients $\frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 + a_j(0)^2\right)$ and $4\pi^2\frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 a_j(0)^2\right)$, we can obtain solutions of different regularity/smoothness for a given training dataset.

Before we rationalize this model, we first demonstrate the effectiveness of this model through experiments on the synthetic training data of 1-d and 2-d input. Note that the long-time limit solution of dynamics (3) is obtained by solving an equivalent optimization problem numerically (see Section 5 and Appendix 11 for details.).

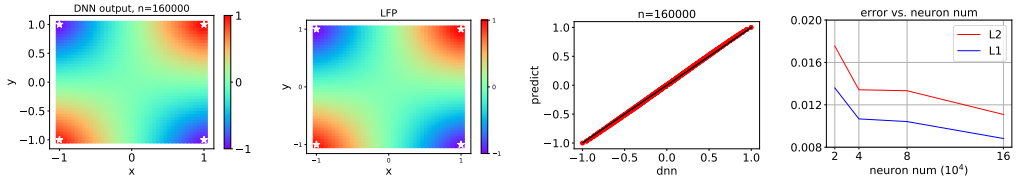
For the case of 1-d input, i.e., $d = 1$, we consider a training dataset of 12 samples as shown in Fig. 1. We first initialize a_i's and w_i's by uniform distributions on $[-0.1, 0.1]$ and $[-0.25, 0.25]$, respectively, such that $1/|\xi|^4$ dominates in Eq. (3). As shown in Fig. 1(a), for the two-layer ReLU NN of 500 hidden neurons, the solution of the corresponding LFP model well approximates the final output of the NN. As we increase the number of hidden neurons to 16000, as shown in Fig. 1(b), the approximation becomes almost perfect. We use L^p norm to quantify the approximation error at testing points $\{x_i\}_{i=1}^{n_{\mathrm{test}}}$, i.e., $L^p(h_m, h_{\mathrm{LFP}}) = \left(\sum_{i=1}^{n_{\mathrm{test}}} |h_m(x_i) - h_{\mathrm{LFP}}(x_i)|^p\right)^{1/p}$, where h_m is the final output of the two-layer ReLU NN of m hidden neurons, h_{LFP} is the solution of the corresponding LFP model. As shown in Fig. 1(c), $L^1(h_m, h_\infty)$ and $L^2(h_m, h_\infty)$ decrease as m increases, indicating that our LFP model accurately models the two-layer NNs of sufficiently large widths. To obtain a less smooth interpolation of the training data, we initialize a_j's and w_j's with uniform distributions on $[-2, 2]$, such that $1/|\xi|^2$ dominates in Eq. (3). As shown in Fig. 1(d), the solution of the LFP model is less smooth compared to that in Fig. 1(a) and (b). In fact, it is close to a piecewise linear function. We note that the LFP model with only the decaying term of $1/|\xi|^2$ in the RHS of Eq. (3) performs the piecewise linear interpolation for $d = 1$. We will elaborate this result in our future works. In this example, the solution of our LFP model almost perfectly overlaps with the final output of the NN.

For the case of 2-d input, i.e., $d = 2$, we consider the training dataset of the famous XOR problem, which cannot be solved by one-layer neural networks. This training dataset consists of four points represented by white stars in Fig. 2(a). As shown in Fig. 2(a-c), our LFP model predicts the final output of NN very accurately over the domain $[-1, 1]^2$. Similar to the 1-d case, as the number of hidden neurons increases, the prediction by the LFP model becomes more accurate. We also present similar experimental results for an asymmetrical training dataset in Fig. 4 in Appendix.

4.1 Rationalization of the LFP dynamics

Our starting point is the following *linearized mean field residual dynamics* (Mei et al. (2018, 2019)):

$$\partial_t u(\boldsymbol{x}, t) = -\int_{\mathbb{R}^d} k_{\boldsymbol{\theta}_0}(\boldsymbol{x}, \boldsymbol{z}) u_\rho(\boldsymbol{z}, t)\, \mathrm{d}\boldsymbol{z}, \tag{4}$$



(a) (b) (c) (d)

Figure 2: LFP model for 2-d training data of the XOR problem. (a) The final output of the NN. (b) The solution of the corresponding LFP model. The training data are marked by white stars. (c) Each dot represents the final output of NN (abscissa) vs. solution of the LFP model (ordinate) evaluated at one of the 1600 evenly spaced test points. The black line indicates the identity function. (d) Decay of $L^1(h_m, h_{\mathrm{LFP}})$ and $L^2(h_m, h_{\mathrm{LFP}})$ (mean of 10 trials) vs. neuron number.

where $\boldsymbol{\theta}_0$ denotes the initial parameters of the NN in the mean field kernel limit. The kernel is defined as

$$k_{\boldsymbol{\theta}_0}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_{(a,\boldsymbol{w},c)\sim\rho_{a,\boldsymbol{w},c}} \left[\nabla_{\boldsymbol{\theta}}\sigma^*(\boldsymbol{x}, \boldsymbol{\theta}_0) \cdot \nabla_{\boldsymbol{\theta}}\sigma^*(\boldsymbol{z}, \boldsymbol{\theta}_0)\right], \tag{5}$$

where $\sigma^*(\cdot, \boldsymbol{\theta}_0) = a_0\sigma\left(\boldsymbol{w}_0^{\intercal}(\cdot) - r_0 c_0\right)$, $r_0 = \|\boldsymbol{w}_0\|$, and σ is the ReLU function. By applying the Fourier transform with respect to $\boldsymbol{x}$ to both sides of Eq. (4), we can approximately derive the following frequency domain dynamics up to a time constant (see Appendix 8 for details), that is

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = -\mathbb{E}_{(a,r)\sim\rho_{a_0,r_0}} \left(\frac{\|\boldsymbol{w}_0\|^2 + a_0^2}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2\|\boldsymbol{w}_0\|^2 a_0^2}{\|\boldsymbol{\xi}\|^{d+1}}\right) \mathcal{F}[u_\rho](\boldsymbol{\xi}, t), \tag{6}$$

where $r = \|\boldsymbol{w}\|$. Then Eq. (3) can be obtained by replacing $\mathbb{E}_{(a,r)\sim\rho_{a_0,r_0}}$ with the mean over m hidden neurons.

5 Explicitizing the implicit bias of the F-Principle

5.1 An equivalent optimization problem to the gradient flow dynamics

In our 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.

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. All proofs are in Appendix 9.

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

$$\langle \mathcal{P}\phi_1, \phi_2\rangle_{H_2} = \langle \phi_1, \mathcal{P}^*\phi_2\rangle_{H_1}, \quad \text{for all} \quad \phi_1 \in H_1, h_2 \in H_2. \tag{7}$$

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

(i) The initial value problem

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \mathcal{P}^*(g - \mathcal{P}\phi), \quad \phi(0) = \phi_{\mathrm{ini}}.$$

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

$$\min_{\phi-\phi_{\mathrm{ini}}\in H_1} \|\phi - \phi_{\mathrm{ini}}\|_{H_1}, \quad \text{s.t.} \quad \mathcal{P}\phi = g.$$

In the following, we will show that it has a unique minimizer which is denoted as $\phi_{\min}$. Now we present the following theorem of the equivalence relation.

Theorem 1. *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*

$$\phi_\infty = \phi_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}(g - \mathcal{P}\phi_{\rm ini}) + \phi_{\rm ini}. \tag{8}$$

The following corollary is obtained directly from Theorem 1.

Corollary 1. *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*

$$\mathcal{P} : \phi \mapsto \left(\int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{\xi} \right)_{i=1}^n = (h(\boldsymbol{x}_i))_{i=1}^n. \tag{9}$$

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

The initial value problem

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

The minimization problem

$$\min_{\phi - \phi_{\rm ini} \in H_\Gamma} \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2} |\phi(\boldsymbol{\xi}) - \phi_{\rm ini}(\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi}, \quad \text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n.$$

Note that in Appendix 9, we provide another version of Corollary 1 for the discretized frequency, which is considered in Section 6.

5.2 Explicitizing the implicit bias for two-layer NNs

By Corollary 1, we derive the following constrained optimization problem explicitly minimizing an FP-norm (see Section 6.1), whose solution is equivalent to that of the LFP model (3), that is,

$$\min_{h - h_{\rm ini} \in F_\gamma} \int_{\mathbb{R}^d} \left(\frac{\frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 + a_j(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 \frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 a_i(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+1}} \right)^{-1} |\mathcal{F}[h](\boldsymbol{\xi}) - \mathcal{F}[h]_{\rm ini}(\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi}, \tag{10}$$

subject to constraints $h(\boldsymbol{x}_i) = y_i$ for $i = 1, \cdots, n$. The F_γ is defined in the next section. Note that the solutions of the LFP models in Figs. (1, 2) are obtained by solving another form of this optimization problem (see Appendix 11). This explicit penalty indicates that the learning of DNN is biased towards functions with more power at low frequencies (more precisely, functions of smaller FP-norm), which is speculated in Xu et al. (2018), Rahaman et al. (2018), Xu et al. (2019). Next, we extend this special example to the case of a general weight function $(\gamma(\boldsymbol{\xi}))^{-2}$ in the frequency domain.

6 FP-norm and an *a priori* generalization error bound

The equivalent explicit optimization problem (10) provides a way to analyze the generalization of sufficiently wide two-layer NNs. We begin with the definition of an FP-norm, which naturally induces a 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* estimates 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 FP-norm and FP-space

We denote $\mathbb{Z}^{d*} := \mathbb{Z}^d \backslash \{\boldsymbol{0}\}$. Given a frequency weight function $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$ or $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ satisfying

$$\|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} < +\infty \quad \text{or} \quad \|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} < +\infty, \tag{11}$$

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

$$\|h\|_\gamma := \|\mathcal{F}[h]\|_{H_\Gamma} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h](\boldsymbol{k})|^2 \right)^{\frac{1}{2}}. \tag{12}$$

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

$$F_\gamma(\Omega) = \{h \in L^2(\Omega) : \|h\|_\gamma < \infty\}. \tag{13}$$

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.

6.2 *a priori* generalization error bound

The following lemma shows that the FP-norm closely relates to the Rademacher complexity, which is defined as

$$R(\mathcal{H}) = \frac{1}{n} \mathbb{E}_{\boldsymbol{\tau}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^n \tau_i h(\boldsymbol{x}_i) \right]. \tag{14}$$

for the function space $\mathcal{H}$ and data set $\{\boldsymbol{x}_i\}_{i=1}^n$.

Lemma 1. *(i) For $\mathcal{H}_Q = \{h : \|h\|_\gamma \le Q\}$ with $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, we have*

$$R(\mathcal{H}_Q) \le \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{15}$$

(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

$$R(\mathcal{H}'_Q) \le \frac{c_0}{\sqrt{n}} + \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{16}$$

By Lemma 4 in Appendix 10, the FP-norm of the solution for the optimization problem in Corollary 1 is bounded by $Q = \|f - h_{\mathrm{ini}}\|_\gamma$. Then, we obtain the following estimate of the generalization error bound.

Theorem 2. *Suppose that the real-valued target function $f \in F_\gamma(\Omega)$, the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ satisfies $y_i = f(\boldsymbol{x}_i)$, $i = 1, \cdots, n$, and h_n is the solution of the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in F_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma, \quad \textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n. \tag{17}$$

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 samples, the population risk has the bound

$$L(h_n) \le \|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2} \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right). \tag{18}$$

(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 samples, the population risk has the bound

$$L(h_n) \le (\|f - h_{\mathrm{ini}}\|_\infty + 2\|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2}) \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right). \tag{19}$$

Remark 1. *By the assumption in the theorem, the target function f belongs to $F_\gamma(\Omega)$ 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 neural network due to universal approximation theory Cybenko (1989).*

Our a priori generalization error bound in Theorem 2 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 (Shalev-Shwartz et al. 2017), whose power concentrates at high frequencies. In the following, We use experiments to illustrate that, as predicted by our a priori generalization error bound, larger FP-norm of the target function indicates a larger generalization error.

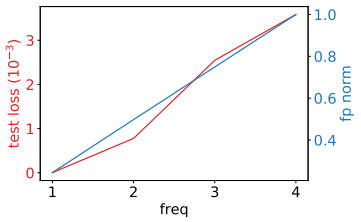


Figure 3: normalized FP-norm and test loss are plotted as a function of frequency v of the target function $\sin(2\pi vx)$.

6.3 Experiment

In this section, we train a ReLU-NN of width 1-5000-1 to fit 20 uniform samples of $f(x) = \sin(2\pi vx)$ on $[0, 1]$ until the training MSE loss is smaller than 10^{-6}, where v is the frequency. We then use 500 uniform samples to test the NN. The FP-norm of the target function is computed by

$$\|h\|_\gamma = \left(\sum_{\boldsymbol{\xi} \in \mathbb{Z}^{d*}} \left(\frac{\frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 + a_j(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 \frac{1}{m}\sum_{j=1}^m \left(\|\boldsymbol{w}_j(0)\|^2 a_j(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+1}} \right)^{-1} |\mathcal{F}[h](\boldsymbol{\xi})|^2 \right)^{1/2}, \tag{20}$$

where $\mathcal{F}[h](\boldsymbol{\xi})$ is computed by the discrete Fourier transform of $h(\boldsymbol{x})$. As shown in Fig. 3, a larger FP-norm of the target function is related to a larger test error.

7 Discussion

In this work, inspired by the F-Principle, we propose an effective LFP model for NNs — a model quantitatively well predicts the output of wide two-layer ReLU NNs and is theoretically rationalized by their training dynamics in an extremely over-parameterized 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 based on the ReLU transfer function can be naturally extended to other transfer functions following a similar construction process.

As a candidate of an effective model of DNNs, the LFP model advances our qualitative/empirical understandings of the F-Principle to a quantitative level. i) With ASI trick (Zhang et al. 2019) 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" (Xu et al. 2018, 2019) is not a necessary condition for the F-Principle. ii) Based on the qualitative behavior of F-Principle, previous works (Xu et al. 2018, 2019, Rahaman et al. 2018) speculate that "DNNs prefer to learn the training data by a low frequency function". With an equivalent optimization problem explicitizing the F-Principle, the LFP model quantifies this speculation.

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 (Shalev-Shwartz et al. 2017). Xu et al. (2019) 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, thus, a large generalization error. Second, because randomly labeled data possesses large high frequency components, which induces a large FP-norm of any function 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 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 (Zhang et al. 2016).

The F-Principle, a widely observed implicit bias of DNNs, is also a natural bias for human. Empirically, when a human see several points of training data, without a specific prior, one tends 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 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 towards a thorough quantitative understanding of deep learning.

References

Arora, S., Du, S. S., Hu, W., Li, Z. & Wang, R. (2019), 'Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks', *arXiv preprint arXiv:1901.08584* .

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y. et al. (2017), 'A closer look at memorization in deep networks', *arXiv preprint arXiv:1706.05394* .

Bartlett, P. L., Foster, D. J. & Telgarsky, M. J. (2017), Spectrally-normalized margin bounds for neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 6240–6249.

Bartlett, P. L., Harvey, N., Liaw, C. & Mehrabian, A. (2017), 'Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks', *arXiv preprint arXiv:1703.02930* .

Bartlett, P. L., Maiorov, V. & Meir, R. (1999), Almost linear vc dimension bounds for piecewise polynomial networks, *in* 'Advances in Neural Information Processing Systems', pp. 190–196.

Bartlett, P. L. & Mendelson, S. (2002), 'Rademacher and gaussian complexities: Risk bounds and structural results', *Journal of Machine Learning Research* **3**(Nov), 463–482.

Bousquet, O. & Elisseeff, A. (2002), 'Stability and generalization', *Journal of machine learning research* **2**(Mar), 499–526.

Cao, Y. & Gu, Q. (2019), 'A generalization theory of gradient descent for learning over-parameterized deep relu networks', *arXiv preprint arXiv:1902.01384* .

Cybenko, G. (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of control, signals and systems* **2**(4), 303–314.

Dinh, L., Pascanu, R., Bengio, S. & Bengio, Y. (2017), 'Sharp minima can generalize for deep nets', *arXiv preprint arXiv:1703.04933* .

Dziugaite, G. K. & Roy, D. M. (2017), 'Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data', *arXiv preprint arXiv:1703.11008* .

E, W., Ma, C., Wang, Q. & Wu, L. (2019), 'Analysis of the gradient descent algorithm for a deep neural network model with skip-connections', *arXiv preprint arXiv:1904.05263* .

E, W., Ma, C. & Wu, L. (2018), 'A priori estimates of the generalization error for two-layer neural networks', *arXiv preprint arXiv:1810.06397* .

E, W., Ma, C. & Wu, L. (2019), 'A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics', *arXiv preprint arXiv:1904.04326* .

Golowich, N., Rakhlin, A. & Shamir, O. (2017), 'Size-independent sample complexity of neural networks', *arXiv preprint arXiv:1712.06541* .

Hardt, M., Recht, B. & Singer, Y. (2015), 'Train faster, generalize better: Stability of stochastic gradient descent', *arXiv preprint arXiv:1509.01240* .

Hochreiter, S. & Schmidhuber, J. (1995), Simplifying neural nets by discovering flat minima, *in* 'Advances in neural information processing systems', pp. 529–536.

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

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. (2016), 'On large-batch training for deep learning: Generalization gap and sharp minima', *arXiv preprint arXiv:1609.04836* .

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. & Pennington, J. (2019), 'Wide neural networks of any depth evolve as linear models under gradient descent', *arXiv preprint arXiv:1902.06720* .

Mei, S., Misiakiewicz, T. & Montanari, A. (2019), 'Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit', *arXiv preprint arXiv:1902.06015* .

Mei, S., Montanari, A. & Nguyen, P.-M. (2018), 'A mean field view of the landscape of two-layer neural networks', *Proceedings of the National Academy of Sciences* **115**(33), E7665–E7671.

Neyshabur, B., Bhojanapalli, S., McAllester, D. & Srebro, N. (2017), Exploring generalization in deep learning, *in* 'Advances in Neural Information Processing Systems', pp. 5947–5956.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. & Srebro, N. (2018), 'Towards understanding the role of over-parametrization in generalization of neural networks', *arXiv preprint arXiv:1805.12076* .

Neyshabur, B., Tomioka, R. & Srebro, N. (2014), 'In search of the real inductive bias: On the role of implicit regularization in deep learning', *arXiv preprint arXiv:1412.6614* .

Pérez, G. V., Louis, A. A. & Camargo, C. Q. (2018), 'Deep learning generalizes because the parameter-function map is biased towards simple functions', *arXiv preprint arXiv:1805.08522* .

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J. & Mhaskar, H. (2018), Theory of deep learning iii: the non-overfitting puzzle, Technical report, Technical report, CBMM memo 073.

Rahaman, N., Arpit, D., Baratin, A., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y. & Courville, A. (2018), 'On the spectral bias of deep neural networks', *arXiv preprint arXiv:1806.08734* .

Rotskoff, G. & Vanden-Eijnden, E. (2018), Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks, *in* 'Advances in neural information processing systems', pp. 7146–7155.

Shalev-Shwartz, S. & Ben-David, S. (2014), *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Shalev-Shwartz, S., Shamir, O. & Shammah, S. (2017), 'Failures of gradient-based deep learning', *arXiv preprint arXiv:1703.07950* .

Sirignano, J. & Spiliopoulos, K. (2018), 'Mean field analysis of neural networks: A central limit theorem', *arXiv preprint arXiv:1808.09372* .

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. & Srebro, N. (2018), 'The implicit bias of gradient descent on separable data', *Journal of Machine Learning Research* **19**(70).

Wu, L., Zhu, Z. & E, W. (2017), 'Towards understanding generalization of deep learning: Perspective of loss landscapes', *arXiv preprint arXiv:1706.10239* .

Xu, H. & Mannor, S. (2012), 'Robustness and generalization', *Machine learning* **86**(3), 391–423.

Xu, Z. J. (2018*a*), 'Understanding training and generalization in deep learning by fourier analysis', *arXiv preprint arXiv:1808.04295* .

Xu, Z.-Q. J. (2018*b*), 'Frequency principle in deep learning with general loss functions and its potential application', *arXiv preprint arXiv:1811.10146* .

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y. & Ma, Z. (2019), 'Frequency principle: Fourier analysis sheds light on deep neural networks', *arXiv preprint arXiv:1901.06523* .

Xu, Z.-Q. J., Zhang, Y. & Xiao, Y. (2018), 'Training behavior of deep neural network in frequency domain', *arXiv preprint arXiv:1807.01251* .

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2016), 'Understanding deep learning requires rethinking generalization', *arXiv preprint arXiv:1611.03530* .

Zhang, Y., Xu, Z.-Q. J., Luo, T. & Ma, Z. (2019), 'A type of generalization error induced by initialization in deep neural networks', *arXiv:1905.07777 [cs, stat]* . arXiv: 1905.07777.
URL: *http://arxiv.org/abs/1905.07777*

Appendix

8 Rationalization of the LFP dynamics

In this section, we derive the dynamics of each frequency component of the loss function when a two-layer ReLU-NN is used to fit a d-dimensional function. Under mild assumption, we can clearly see that a lower frequency component has a faster convergence speed.

The starting point is the following *linearized mean field residual dynamics* Mei et al. (2018, 2019):

$$\partial_t u(\boldsymbol{x}, t) = -\int_{\mathbb{R}^d} k_{\boldsymbol{\theta}_0}(\boldsymbol{x}, \boldsymbol{z}) u_\rho(\boldsymbol{z}, t)\, \mathrm{d}\boldsymbol{z}, \tag{21}$$

here $\boldsymbol{\theta}_0$ denotes the initial parameters of the training dynamics, considering the linear kernel regime. The kernel is defined as

$$k_{\boldsymbol{\theta}_0}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_{(a_0, \boldsymbol{w}_0, c_0)\sim\rho_{a,\boldsymbol{w},c}} \left[\nabla_{\boldsymbol{\theta}} \sigma^*(\boldsymbol{x}, \boldsymbol{\theta}_0) \cdot \nabla_{\boldsymbol{\theta}} \sigma^*(\boldsymbol{z}, \boldsymbol{\theta}_0)\right], \tag{22}$$

with $\sigma^*(\cdot, \boldsymbol{\theta}_0) = a_0\sigma\left(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0\right)$, $r_0 = \|\boldsymbol{w}_0\|$, and σ is the ReLU function. For simplicity, we assume the parameters are isotropic and $\rho_c(c_0) = C_{c_0}$, $\rho_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}}_0) = C_{\hat{\boldsymbol{w}}_0}$ to be specified later that is,

$$\rho_{a,\boldsymbol{w},c}(a_0, \boldsymbol{w}_0, c_0)\, \mathrm{d}a_0\, \mathrm{d}\boldsymbol{w}_0\, \mathrm{d}c_0 = C_{\hat{\boldsymbol{w}}_0} C_{c_0} \rho_a(a_0)\rho_r(r_0)\, \mathrm{d}a_0\, \mathrm{d}r_0\, \mathrm{d}\hat{\boldsymbol{w}}_0\, \mathrm{d}c_0 := \rho_{a,\boldsymbol{w},c}(\mathrm{d}\boldsymbol{\theta}_0), \tag{23}$$

where $\hat{\boldsymbol{w}}_0$ is the unit vector of $\boldsymbol{w}_0$. Then gradient of σ^* with respect to the parameters is

$$\nabla_{\boldsymbol{\theta}} \sigma^*(\cdot, \boldsymbol{\theta}_0) = \begin{pmatrix} \sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \\ a_0(\cdot - c_0 e_{\boldsymbol{w}_0})\sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \\ -a_0 r_0 \sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \end{pmatrix}. \tag{24}$$

For any given vector $\boldsymbol{w}_0$, we can decompose the gradient with respect to $\boldsymbol{w}_0$, that is, second row in the above gradient vector, into two directions: parallel and perpendicular to $\boldsymbol{w}_0$, i.e.,

$$\nabla_{\boldsymbol{w}_0} \sigma^*(\cdot, \boldsymbol{\theta}_0) = \frac{\boldsymbol{w}_0 \cdot \nabla_{\boldsymbol{w}_0} \sigma^*(\cdot, \boldsymbol{\theta}_0)}{r_0^2} \boldsymbol{w}_0 + \left[\nabla_{\boldsymbol{w}_0} \sigma^*(\cdot, \boldsymbol{\theta}_0) - \frac{\boldsymbol{w}_0 \cdot \nabla_{\boldsymbol{w}_0} \sigma^*(\cdot, \boldsymbol{\theta}_0)}{r_0^2} \boldsymbol{w}_0\right] \tag{25}$$

$$= a_0(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)\sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)\frac{\boldsymbol{w}_0}{r_0^2} + a_0[\cdot]_{\perp \boldsymbol{w}_0}\sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \tag{26}$$

$$= a_0\sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)\frac{\boldsymbol{w}_0}{r_0^2} + a_0[\cdot]_{\perp \boldsymbol{w}_0}\sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0), \tag{27}$$

where $[\cdot]_{\perp \boldsymbol{w}_0} = [\cdot] - \frac{\boldsymbol{w}_0^\intercal(\cdot)}{r_0^2}\boldsymbol{w}_0$ and $[\cdot]\sigma'(\cdot) = \sigma(\cdot)$ due to the property of the ReLU function. Thus, the gradients (24) can be written into two parts

$$\nabla_{\boldsymbol{\theta}} \sigma^*(\cdot, \boldsymbol{\theta}_0) = \begin{pmatrix} \sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \\ \frac{a_0}{r_0^2}\sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)\boldsymbol{w}_0 \\ -a_0 r_0 \sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \end{pmatrix} + \begin{pmatrix} 0 \\ a_0\sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)[\cdot]_{\perp \boldsymbol{w}_0} \\ 0 \end{pmatrix} \tag{28}$$

$$:= \boldsymbol{A}(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) + \boldsymbol{B}. \tag{29}$$

Then the kernel (22) can be split into two parts,

$$k_{\boldsymbol{\theta}_0}(\boldsymbol{x}, \boldsymbol{z}) = \int_{\mathbb{R}^{d+2}} (\boldsymbol{A}(\boldsymbol{w}_0^\intercal \boldsymbol{x} - r_0 c_0))^\intercal\, \boldsymbol{A}(\boldsymbol{w}_0^\intercal \boldsymbol{z} - r_0 c_0)\rho_{a,\boldsymbol{w},c}(\mathrm{d}\boldsymbol{\theta}_0) + \int_{\mathbb{R}^{d+2}} \boldsymbol{B}^\intercal \boldsymbol{B}\rho_{a,\boldsymbol{w},c}(\mathrm{d}\boldsymbol{\theta}_0) \tag{30}$$

$$=: k_{\boldsymbol{A}} + k_{\boldsymbol{B}}, \tag{31}$$

In the following computation, we will drop the term $k_{\boldsymbol{B}}$. Since this term corresponds to the direction perpendicular to $\boldsymbol{w}_0$, it is not very easy to check by numerical experiments and we only consider the dynamics of the parallel part.

Taking Fourier transform with respect to $\boldsymbol{x}$,

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = -\int_{\mathbb{R}^d} \hat{k}_{\boldsymbol{\theta}_0}(\boldsymbol{\xi}, \boldsymbol{z}) u_\rho(\boldsymbol{z}, t)\, \mathrm{d}\boldsymbol{z} \tag{32}$$

$$= -\int_{\mathbb{R}^{2d+2}} (\mathcal{F}[\boldsymbol{A}(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)](\boldsymbol{\xi}))^\intercal\, \boldsymbol{A}(\boldsymbol{w}_0^\intercal \boldsymbol{z} - r_0 c_0) u_\rho(\boldsymbol{z}, t)\rho_{a,\boldsymbol{w},c}(\mathrm{d}\boldsymbol{\theta}_0)\, \mathrm{d}\boldsymbol{z}. \tag{33}$$

Notice for any function g defined on $\mathbb{R}$, the following results hold:

$$\mathcal{F}[g(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)](\boldsymbol{\xi}) = \int_{\mathbb{R}^d} g(\boldsymbol{w}_0^\intercal \boldsymbol{x} - r_0 c_0) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{x}} \,\mathrm{d}\boldsymbol{x} \tag{34}$$

$$= \int_{\mathbb{R}^d} g(\boldsymbol{w}_0^\intercal \boldsymbol{x} - r_0 c_0) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal (\boldsymbol{x} - c_0 \hat{\boldsymbol{w}}_0)} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0 c_0} \,\mathrm{d}\boldsymbol{x} \tag{35}$$

$$= \mathcal{F}[g(\boldsymbol{w}_0^\intercal(\cdot))](\boldsymbol{\xi}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0 c_0}, \tag{36}$$

and

$$\mathcal{F}[g(\boldsymbol{w}_0^\intercal(\cdot))](\boldsymbol{\xi}) = \int_{\mathbb{R}^d} g(\boldsymbol{w}_0^\intercal \boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{x}} \,\mathrm{d}\boldsymbol{x} \tag{37}$$

$$= \int_{\mathbb{R}^d} g(\boldsymbol{w}_0^\intercal \boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{x}_{\| \boldsymbol{w}_0}} \mathrm{e}^{-2\pi \boldsymbol{\xi}^\intercal \boldsymbol{x}_{\perp \boldsymbol{w}_0}} \,\mathrm{d}\boldsymbol{x}_{\| \boldsymbol{w}_0} \,\mathrm{d}\boldsymbol{x}_{\perp \boldsymbol{w}_0} \tag{38}$$

$$= \left(\int_{\mathbb{R}} g(\boldsymbol{w}_0^\intercal \boldsymbol{x}_{\| \boldsymbol{w}_0}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^T \boldsymbol{x}_{\| \boldsymbol{w}_0}} \,\mathrm{d}\boldsymbol{x}_{\| \boldsymbol{w}_0} \right) \left(\int_{\mathbb{R}^{d-1}} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^T \boldsymbol{x}_{\perp \boldsymbol{w}_0}} \,\mathrm{d}\boldsymbol{x}_{\perp \boldsymbol{w}_0} \right) \tag{39}$$

$$= \left(\int_{\mathbb{R}} g(r_0 y) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0 y} \,\mathrm{d}y \right) \delta(\boldsymbol{\xi}_{\perp \boldsymbol{w}_0}) \tag{40}$$

$$= \frac{1}{r_0} \mathcal{F}[g] \left(\frac{\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0}{r_0} \right) \delta(\boldsymbol{\xi}_{\perp \boldsymbol{w}_0}), \tag{41}$$

where $\boldsymbol{x}_{\| \boldsymbol{w}_0} = \frac{\boldsymbol{w}_0^\intercal \boldsymbol{x}}{r_0^2} \boldsymbol{w}_0$, $\boldsymbol{x}_{\perp \boldsymbol{w}_0} = \boldsymbol{x} - \boldsymbol{x}_{\| \boldsymbol{w}_0}$, $\boldsymbol{\xi}_{\perp \boldsymbol{w}_0} = \boldsymbol{\xi} - \frac{\boldsymbol{\xi}^\intercal \boldsymbol{w}_0}{r_0^2} \boldsymbol{w}_0$ and $y = \hat{\boldsymbol{w}}_0^\intercal \boldsymbol{x}$. The above delta function is defined as, for any function $g(\hat{\boldsymbol{w}}_0)$

$$\int_{\mathbb{S}^{d-1}} \delta(\boldsymbol{\xi}_{\perp \boldsymbol{w}_0}) g(\hat{\boldsymbol{w}}_0) \rho_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}}_0) \,\mathrm{d}\hat{\boldsymbol{w}}_0 = C_{\hat{\boldsymbol{w}}_0} \|\boldsymbol{\xi}\|^{-(d-1)} g(\hat{\boldsymbol{\xi}}), \tag{42}$$

where $C_{\hat{\boldsymbol{w}}_0} = \frac{\Gamma(d/2)}{2\pi^{d/2}}$. Combining the above results, one obtains

$$\mathcal{F}[\boldsymbol{A}(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)](\boldsymbol{\xi}) = \mathcal{F} \left[\begin{pmatrix} \sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0)] \\ a_0 \sigma(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \frac{\boldsymbol{w}_0}{r_0^2} \\ -a_0 r_0 \sigma'(\boldsymbol{w}_0^\intercal(\cdot) - r_0 c_0) \end{pmatrix} \right] (\boldsymbol{\xi}) \tag{43}$$

$$= -\frac{1}{4\pi^2} \begin{pmatrix} 1 \\ \frac{a_0 \boldsymbol{w}_0}{r_0^2} \\ 2\pi \mathrm{i} a_0 (\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0) \end{pmatrix} \frac{r_0}{(\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0)^2} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0 c_0} \delta(\boldsymbol{\xi}_{\perp \boldsymbol{w}_0}). \tag{44}$$

So

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = \frac{1}{4\pi^2} \int_{\mathbb{R}^{2d+2}} \delta(\boldsymbol{\xi}_{\perp \boldsymbol{w}_0}) \frac{r_0}{(\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0)^2} \Bigg[\left(1 + \frac{a_0^2}{r_0^2} \right) \sigma(\boldsymbol{w}_0^\intercal \boldsymbol{z} - r_0 c_0)$$

$$- 2\pi \mathrm{i} r_0 a_0^2 (\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0) \sigma'(\boldsymbol{w}_0^\intercal \boldsymbol{z} - r_0 c_0) \Bigg] \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}_0 c_0} u_\rho(\boldsymbol{z}, t) \rho_{a, \boldsymbol{w}, c}(\mathrm{d}\boldsymbol{\theta}_0) \,\mathrm{d}\boldsymbol{z}, \tag{45}$$

then we first integrate the variable $\hat{\boldsymbol{w}}_0$ using (42) to get

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi},t) = \frac{C_{\hat{\boldsymbol{w}}_0} C_{c_0}}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}} \int_{\mathbb{R}^{d+2}\times\mathbb{R}^+} \Bigg[\left(1+\frac{a_0^2}{r_0^2}\right) r_0 \sigma(r_0 \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z} - r_0 c_0)$$

$$- 2\pi \mathrm{i} a_0^2 r_0^2 \|\boldsymbol{\xi}\| \sigma'(r_0 \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z} - r_0 c_0) \Bigg] \mathrm{e}^{-2\pi \mathrm{i}\|\boldsymbol{\xi}\| c_0} u_\rho(\boldsymbol{z},t) \rho_a(a_0) \rho_{\boldsymbol{w}}(r_0) r_0^{d-1} \,\mathrm{d}a_0 \,\mathrm{d}r_0 \,\mathrm{d}c_0 \,\mathrm{d}\boldsymbol{z} \tag{46}$$

$$= \frac{C_{\hat{\boldsymbol{w}}_0} C_{c_0}}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}} \int_{\mathbb{R}^{d+2}\times\mathbb{R}^+} \Bigg[\left(1+\frac{a_0^2}{r_0^2}\right) r_0 \sigma(r_0 \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z} - r_0 c_0) \mathrm{e}^{-2\pi \mathrm{i}\|\boldsymbol{\xi}\|(c_0 - \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z})}$$

$$- 2\pi \mathrm{i} a_0^2 r_0^2 \|\boldsymbol{\xi}\| \sigma'(r_0 \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z} - r_0 c_0) \mathrm{e}^{-2\pi \mathrm{i}\|\boldsymbol{\xi}\|(c_0 - \hat{\boldsymbol{\xi}}^\intercal \boldsymbol{z})} \Bigg] u_\rho(\boldsymbol{z},t) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{\xi}^\intercal \boldsymbol{z}} \rho_a(a_0) \rho_r(r_0) r_0^{d-1} \,\mathrm{d}a_0 \,\mathrm{d}r_0 \,\mathrm{d}c_0 \,\mathrm{d}\boldsymbol{z} \tag{47}$$

$$= \frac{C_{\hat{\boldsymbol{w}}_0} C_{c_0}}{4\pi^2\|\boldsymbol{\xi}\|^{d+1}} \int_{\mathbb{R}\times\mathbb{R}^+} \Big[(r_0^2 + a_0^2) \overline{\mathcal{F}[\sigma](\|\boldsymbol{\xi}\|)}$$

$$- 2\pi \mathrm{i} a_0^2 r_0^2 \|\boldsymbol{\xi}\| \overline{\mathcal{F}[\sigma'(r_0 \cdot)](\|\boldsymbol{\xi}\|)} \Big] \mathcal{F}[u_\rho](\boldsymbol{\xi},t) \rho_a(a_0) \rho_r(r_0) r_0^{d-1} \,\mathrm{d}a_0 \,\mathrm{d}r_0. \tag{48}$$

Notice that

$$\mathcal{F}[\sigma](\|\boldsymbol{\xi}\|) = -\frac{1}{4\pi^2\|\boldsymbol{\xi}\|^2}, \tag{49}$$

$$\mathcal{F}[\sigma'(r_0 \cdot)](\|\boldsymbol{\xi}\|) = \frac{1}{2\pi \mathrm{i}\|\boldsymbol{\xi}\|}, \tag{50}$$

thus

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi},t) = -\frac{C_{\hat{\boldsymbol{w}}_0} C_{c_0}}{16\pi^4} \mathbb{E}_{(a,r)\sim\rho_{a_0,r_0}} \left(\frac{r^2+a^2}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 r^2 a^2}{\|\boldsymbol{\xi}\|^{d+1}} \right) \mathcal{F}[u_\rho](\boldsymbol{\xi},t). \tag{51}$$

Finally, by choosing a suitable time scale the above dynamics can be written as

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi},t) = -\mathbb{E}_{(a,r)\sim\rho_{a_0,r_0}} \left(\frac{r^2+a^2}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 r^2 a^2}{\|\boldsymbol{\xi}\|^{d+1}} \right) \mathcal{F}[u_\rho](\boldsymbol{\xi},t). \tag{52}$$

9 Proofs of the equivalence theorems

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

$$\langle \mathcal{P}\phi_1, \phi_2 \rangle_{H_2} = \langle \phi_1, \mathcal{P}^* \phi_2 \rangle_{H_1}, \quad \text{for all} \quad \phi_1 \in H_1, \phi_2 \in H_2. \tag{53}$$

Lemma 2. *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 non-negative. 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.*

Proof. We consider the eigenvalue problem $\mathcal{P}^*\mathcal{P}\phi_1 = \lambda\phi_1$. Taking inner product with ϕ_1, we have $\langle \phi_1, \mathcal{P}^*\mathcal{P}\phi_1 \rangle_{H_1} = \lambda \|\phi_1\|_{H_1}^2$. Note that the left hand side is $\|\mathcal{P}\phi_1\|_{H_2}^2$ which is non-negative. Thus $\lambda \geq 0$. Similarly, the eigenvalues of $\mathcal{P}\mathcal{P}^*$ are also non-negative.

Now if $\mathcal{P}^*\mathcal{P}$ has a positive eigenvalue $\lambda > 0$, then $\mathcal{P}^*\mathcal{P}\phi_1 = \lambda\phi_1$ with non-zero vector $\phi_1 \in H_1$. It follows that $\mathcal{P}\mathcal{P}^*(\mathcal{P}\phi_1) = \lambda(\mathcal{P}\phi_1)$. It is sufficient to prove that $\mathcal{P}\phi_1$ is non-zero. Indeed, if $\mathcal{P}\phi_1 = 0$, then $\mathcal{P}^*\mathcal{P}\phi_1 = 0$ and $\lambda = 0$ which contradicts with our assumption. Therefore, any positive eigenvalue of $\mathcal{P}^*\mathcal{P}$ is an eigenvalue of $\mathcal{P}\mathcal{P}^*$. Similarly, any positive eigenvalue of $\mathcal{P}\mathcal{P}^*$ is an eigenvalue of $\mathcal{P}^*\mathcal{P}$.

Next, suppose that $\mathcal{P}\mathcal{P}^*$ is surjective. We show that $\mathcal{P}\mathcal{P}^*\phi_2 = 0$ has only the trivial solution $\phi_2 = 0$. In fact, $\mathcal{P}\mathcal{P}^*\phi_2 = 0$ implies that $\|\mathcal{P}^*\phi_2\|_{H_1}^2 = \langle \phi_2, \mathcal{P}\mathcal{P}^*\phi_2 \rangle_{H_2} = 0$, i.e., $\mathcal{P}^*\phi_2 = 0$. Thanks to the surjectivity of $\mathcal{P}\mathcal{P}^*$, there exists a vector $\phi_3 \in H_2$ such that $\phi_2 = \mathcal{P}\mathcal{P}^*\phi_3$. Let $\phi_1 = \mathcal{P}^*\phi_3 \in H_1$. Hence $\phi_2 = \mathcal{P}\phi_1$ and $\mathcal{P}^*\mathcal{P}h_1 = 0$. Taking inner product with ϕ_1, we have $\|\mathcal{P}\phi_1\|_{H_2}^2 = \langle \phi_1, \mathcal{P}^*\mathcal{P}\phi_1 \rangle_{H_1} = 0$, i.e., $\phi_2 = \mathcal{P}\phi_1 = 0$. Therefore $\mathcal{P}\mathcal{P}^*$ is injective. This with the surjectivity assumption of $\mathcal{P}\mathcal{P}^*$ leads to that $\mathcal{P}\mathcal{P}^*$ is invertible. □

Remark 2. *For the finite dimensional case $H_2 = \mathbb{R}^n$, conditions for the operator $\mathcal{P}$ in Lemma 2 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

$$\begin{aligned} &\min_{\phi-\phi_{\mathrm{ini}}\in H_1} \|\phi - \phi_{\mathrm{ini}}\|_{H_1}, \\ &\text{s.t.} \quad \mathcal{P}\phi = g. \end{aligned}$$

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

Now we show the following equivalent theorem.

Theorem 3. *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*

$$\phi_\infty = h_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}(g - \mathcal{P}\phi_{\mathrm{ini}}) + \phi_{\mathrm{ini}}. \tag{54}$$

Proof. Let $\tilde{\phi} = \phi - \phi_{\mathrm{ini}}$ and $\tilde{g} = g - \mathcal{P}\phi_{\mathrm{ini}}$. Then it is sufficient to show the following problems (i') and (ii') are equivalent.

(i') The initial value problem

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

(ii') The minimization problem

$$\begin{aligned} &\min_{\tilde{\phi}} \|\tilde{\phi}\|_{H_1}^2, \\ &\text{s.t.} \quad \mathcal{P}\tilde{\phi} = \tilde{g}. \end{aligned}$$

We claim that $\tilde{\phi}_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g}$. Thanks to Lemma 2, $\mathcal{P}\mathcal{P}^*$ is invertible, and thus $\phi_{\min}$ is well-defined and satisfies that $\mathcal{P}\tilde{\phi} = \tilde{g}$. It remains to show that this solution is unique. In fact, for any $\tilde{\phi}$ satisfying $\mathcal{P}\tilde{\phi} = \tilde{g}$, we have

$$\begin{aligned} \langle \tilde{\phi} - \tilde{\phi}_{\min}, \tilde{\phi}_{\min} \rangle_{H_1} &= \langle \tilde{\phi} - \tilde{\phi}_{\min}, \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} \rangle_{H_1} \\ &= \langle \mathcal{P}(\tilde{\phi} - \tilde{\phi}_{\min}), (\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} \rangle_{H_2} \\ &= \langle \mathcal{P}\tilde{\phi}, (\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} \rangle_{H_2} - \langle \mathcal{P}\tilde{\phi}_{\min}, (\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} \rangle_{H_2} \\ &= 0. \end{aligned}$$

Therefore,

$$\|\tilde{\phi}\|_{H_1}^2 = \|\tilde{\phi}_{\min}\|_{H_1}^2 + \|\tilde{\phi} - \tilde{\phi}_{\min}\|_{H_1}^2 \geq \|\tilde{\phi}_{\min}\|_{H_1}^2.$$

The equality holds if and only if $\tilde{\phi} = \tilde{\phi}_{\min}$.

For problem (i'), from the theory of ordinary differential equations on Hilbert spaces, we have that its solution can be written as

$$\tilde{\phi}(t) = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} + \sum_{i\in I} c_i v_i \exp(-\lambda_i t),$$

where λ_i, $i \in \mathcal{I}$ are positive eigenvalues of $\mathcal{P}\mathcal{P}^*$, $\mathcal{I}$ is an index set with at most countable cardinality, and v_i, $i \in \mathcal{I}$ are eigenvectors in H_1. Thus $\tilde{\phi}_\infty = \tilde{\phi}_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g}$.

Finally, by back substitution, we have

$$\phi_\infty = \phi_{\min} = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}\tilde{g} + \phi_0 = \mathcal{P}^*(\mathcal{P}\mathcal{P}^*)^{-1}(g - \mathcal{P}\phi_{\rm ini}) + \phi_{\rm ini}.$$

□

The following corollaries are obtained directly from Theorem 3.

Corollary 2. *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

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

(A2) The minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}-\boldsymbol{\theta}_{\rm ini}\in\mathbb{R}^m} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\rm ini}\|_2, \\ &\textit{s.t.} \quad \boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{Y}. \end{aligned}$$

The next corollary is a weighted version of Theorem 3.

Corollary 3. *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

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

(B2) The minimization problem

$$\begin{aligned} &\min_{\phi-\phi_0\in H_\Gamma} \|\phi - \phi_{\rm ini}\|_{H_\Gamma}, \\ &\textit{s.t.} \quad \mathcal{P}\phi = g. \end{aligned}$$

Proof. The operator $\Gamma : H_1 \to H_\Gamma$ is bijective. Hence $\Gamma^{-1} : H_\Gamma \to H_1$ is well-defined and H_Γ with norm $\|\cdot\|_{H_\Gamma}$ is a Hilbert space. The equivalence result holds by applying Theorem 3 with proper replacements. More precisely, we replace ϕ by $\Gamma^{-1}\phi$ and $\mathcal{P}$ by $\mathcal{P}\Gamma$. □

Corollary 4. *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*

$$\mathcal{P} : \phi \mapsto \left(\int_{\mathbb{R}^d} \phi(\boldsymbol{\xi})\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{x}_1^\intercal\boldsymbol{\xi}}\,\mathrm{d}\boldsymbol{\xi}, \ldots, \int_{\mathbb{R}^d} \phi(\boldsymbol{\xi})\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{x}_n^\intercal\boldsymbol{\xi}}\,\mathrm{d}\boldsymbol{\xi}\right)^\intercal = (h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_n))^\intercal. \tag{55}$$

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

(C1) The initial value problem

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

(C2) The minimization problem

$$\begin{aligned} &\min_{\phi-\phi_{\rm ini}\in H_\Gamma} \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2}|\phi(\boldsymbol{\xi}) - \phi_{\rm ini}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}, \\ &\textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n. \end{aligned}$$

Proof. Let $H_1 = L^2(\mathbb{R}^d)$, $H_2 = \mathbb{R}^n$, $g = \boldsymbol{Y}$. By definition, Γ is injective. Then by Corollary 3, we have that $\Gamma^{-1} : H_\Gamma \to L^2(\mathbb{R}^d)$ exists and H_Γ is a Hilbert space with norm $\|\phi\|_{H_\Gamma} := \|\Gamma^{-1}\phi\|_{L^2(\mathbb{R}^d)}$. Moreover, $\|\phi - \phi_{\mathrm{ini}}\|_{H_\Gamma}^2 = \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2} |\phi(\boldsymbol{\xi}) - \phi_{\mathrm{ini}}(\boldsymbol{\xi})|^2 \,\mathrm{d}\boldsymbol{\xi}$. We note that $[\mathcal{P}^* Y](\boldsymbol{\xi}) = \sum_{i=1}^n y_i \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}}$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$. Thus

$$
\begin{aligned}
[\mathcal{P}^*\mathcal{P}\phi](\boldsymbol{\xi}) &= \left[\mathcal{P}^* \left(\int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}') \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}'} \,\mathrm{d}\boldsymbol{\xi}' \right)_{i=1}^n \right](\boldsymbol{\xi}) \\
&= \sum_{i=1}^n \int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}') \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}'} \,\mathrm{d}\boldsymbol{\xi}' \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{\xi}} \\
&= \sum_{i=1}^n \int_{\mathbb{R}^d} \phi(\boldsymbol{\xi}') \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\boldsymbol{\xi} - \boldsymbol{\xi}')} \,\mathrm{d}\boldsymbol{\xi}' \\
&= \sum_{i=1}^n \left[\phi * \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\cdot)} \right](\boldsymbol{\xi}).
\end{aligned}
$$

The equivalence result then follows from Corollary 3. □

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

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

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

Corollary 5. *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*

$$
P : \phi \mapsto \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} \phi(\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_1^\intercal \boldsymbol{k}}, \ldots, \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \phi(\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_n^\intercal \boldsymbol{k}} \right)^\intercal. \tag{56}
$$

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

(D1) The initial value problem

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

(D2) The minimization problem

$$
\begin{aligned}
&\min_{\phi - \phi_{\mathrm{ini}} \in H_\Gamma} \sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\phi(\boldsymbol{k}) - \phi_{\mathrm{ini}}(\boldsymbol{k})|^2, \\
&\textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n.
\end{aligned}
$$

Proof. Let $H_1 = \ell^2(\mathbb{Z}^d)$, $H_2 = \mathbb{R}^n$, and $g = \boldsymbol{Y}$. By definition, Γ is injective. Then by Corollary 3, we have that $\Gamma^{-1} : H_\Gamma \to \ell^2(\mathbb{Z}^d)$ exists and H_Γ is a Hilbert space with norm $\|\phi\|_{H_\Gamma} := \|\Gamma^{-1}\phi\|_{\ell^2(\mathbb{Z}^d)}$. Moreover, $\|\phi - \phi_{\mathrm{ini}}\|_{H_\Gamma}^2 = \sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\phi(\boldsymbol{k}) - \phi_{\mathrm{ini}}(\boldsymbol{k})|^2$. We

note that $[P^*\boldsymbol{Y}](\boldsymbol{k}) = \sum_{i=1}^n y_i \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{k}}$ for all $\boldsymbol{k} \in \mathbb{Z}^d$. Thus

$$\begin{aligned}
[P^*P\phi](\boldsymbol{k}) &= \left[P^* \left(\sum_{\boldsymbol{k}' \in \mathbb{Z}^d} \phi(\boldsymbol{k}') \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{k}'} \right)_{i=1}^n \right] (\boldsymbol{k}) \\
&= \sum_{i=1}^n \sum_{\boldsymbol{k}' \in \mathbb{Z}^d} \phi(\boldsymbol{k}') \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{k}'} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal \boldsymbol{k}} \\
&= \sum_{i=1}^n \sum_{\boldsymbol{k}' \in \mathbb{Z}^d} \phi(\boldsymbol{k}') \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\boldsymbol{k} - \boldsymbol{k}')} \\
&= \sum_{i=1}^n \left[\phi * \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}_i^\intercal (\cdot)} \right] (\boldsymbol{k}).
\end{aligned}$$

The equivalence result then follows from Corollary 3. □

10 Proof of the *a priori* generalization error bound

10.1 Problem Setup

We focus on 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

$$\ell(h, \boldsymbol{x}, y) = |h(\boldsymbol{x}) - y|^2, \tag{57}$$

with population risk

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \ell(h, \boldsymbol{x}, f(\boldsymbol{x})) \tag{58}$$

and empirical risk

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \ell(h, \boldsymbol{x}_i, y_i). \tag{59}$$

10.2 FP-space

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

$$\|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} < +\infty \quad \text{or} \quad \|\gamma\|_{\ell^2} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} < +\infty, \tag{60}$$

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

$$\|h\|_\gamma := \|\mathcal{F}[h]\|_{H_\Gamma} = \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h](\boldsymbol{k})|^2 \right)^{\frac{1}{2}}. \tag{61}$$

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

Then we define the FP-space

$$\mathcal{F}_\gamma(\Omega) = \{h \in L^2(\Omega) : \|h\|_\gamma < \infty\}. \tag{62}$$

By our definition, the FP-space is always a subspace of $L^2(\Omega)$. In addition, if $\gamma : \boldsymbol{k} \mapsto |\boldsymbol{k}|^{-m}$ 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 $H^m(\Omega)$.

10.3 A prior generalization error bound

Lemma 3. *(i) For $\mathcal{H}_Q = \{h : \|h\|_\gamma \le Q\}$ with $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, we have*

$$R(\mathcal{H}_Q) \le \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{63}$$

(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

$$R(\mathcal{H}'_Q) \le \frac{c_0}{\sqrt{n}} + \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{64}$$

Proof. We first prove (ii) since it is more involved. By the definition of the Rademacher complexity

$$R(\mathcal{H}'_Q) = \frac{1}{n} \mathbb{E}_{\boldsymbol{\tau}} \left[\sup_{h \in \mathcal{H}'_Q} \sum_{i=1}^n \tau_i h(\boldsymbol{x}_i) \right]. \tag{65}$$

Let $\tau(\boldsymbol{x}) = \sum_{i=1}^n \tau_i \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, where τ_i's are i.i.d. random variables with $\mathbb{P}(\tau_i = 1) = \mathbb{P}(\tau_i = -1) = \frac{1}{2}$. We have $\mathcal{F}[\tau](\boldsymbol{k}) = \int_\Omega \sum_{i=1}^n \tau_i \delta(\boldsymbol{x} - \boldsymbol{x}_i) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{k}^\intercal \boldsymbol{x}} \, \mathrm{d}\boldsymbol{x} = \sum_{i=1}^n \tau_i \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{k}^\intercal \boldsymbol{x}_i}$. Note that

$$\sup_{h \in \mathcal{H}'_Q} \sum_{i=1}^n \tau_i h(\boldsymbol{x}_i) = \sup_{h \in \mathcal{H}'_Q} \sum_{i=1}^n \tau_i \bar{h}(\boldsymbol{x}_i) = \sup_{h \in \mathcal{H}'_Q} \sum_{i=1}^n \tau_i \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \overline{\mathcal{F}[h](\boldsymbol{k})} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{k}^\intercal \boldsymbol{x}_i} \tag{66}$$

$$= \sup_{h \in \mathcal{H}'_Q} \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \mathcal{F}[\tau](\boldsymbol{k}) \overline{\mathcal{F}[h](\boldsymbol{k})}. \tag{67}$$

By the Cauchy–Schwarz inequality,

$$\sup_{h \in \mathcal{H}'_Q} \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \mathcal{F}[\tau](\boldsymbol{k}) \overline{\mathcal{F}[h](\boldsymbol{k})}$$

$$\le \sup_{h \in \mathcal{H}_Q} \left[\mathcal{F}[\tau](\mathbf{0}) \overline{\mathcal{F}[h](\mathbf{0})} + \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 |\mathcal{F}[\tau](\boldsymbol{k})|^2 \right)^{1/2} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^{-2} |\overline{\mathcal{F}[h](\boldsymbol{k})}|^2 \right)^{1/2} \right] \tag{68}$$

$$\le c_0 |\mathcal{F}[\tau](\mathbf{0})| + Q \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 |\mathcal{F}[\tau](\boldsymbol{k})|^2 \right)^{1/2}. \tag{69}$$

Since $\mathbb{E}_{\boldsymbol{\tau}} |\mathcal{F}[\tau](\mathbf{0})| \le (\mathbb{E}_{\boldsymbol{\tau}} |\mathcal{F}[\tau](\mathbf{0})|^2)^{1/2} = \sqrt{n}$, $\mathbb{E}_{\boldsymbol{\tau}} |\mathcal{F}[\tau](\boldsymbol{k})|^2 = \mathbb{E}_{\boldsymbol{\tau}} \sum_{i,j=1}^n \tau_i \tau_j \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{k}^\intercal (\boldsymbol{x}_i - \boldsymbol{x}_j)} = n$, we obtain

$$\mathbb{E}_{\boldsymbol{\tau}} \left[\sup_{h \in \mathcal{H}'_Q} \sum_{i=1}^n \tau_i h(\boldsymbol{x}_i) \right] \le c_0 \sqrt{n} + Q \mathbb{E}_{\boldsymbol{\tau}} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 |\mathcal{F}[\tau](\boldsymbol{k})|^2 \right)^{1/2} \tag{70}$$

$$\le c_0 \sqrt{n} + Q \left(\mathbb{E}_{\boldsymbol{\tau}} \sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 |\mathcal{F}[\tau](\boldsymbol{k})|^2 \right)^{1/2} \tag{71}$$

$$= c_0 \sqrt{n} + Q \sqrt{n} \|\gamma\|_{\ell^2}. \tag{72}$$

This leads to

$$R(\mathcal{H}'_Q) \le \frac{c_0}{\sqrt{n}} + \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{73}$$

For (ii), the proof is similar to (i). We have

$$\mathbb{E}_{\boldsymbol{\tau}} \left[\sup_{h \in \mathcal{H}_Q} \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \mathcal{F}[\tau](\boldsymbol{k}) \overline{\mathcal{F}[h](\boldsymbol{k})} \right] \le Q \mathbb{E}_{\boldsymbol{\tau}} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^2 |\mathcal{F}[\tau](\boldsymbol{k})|^2 \right)^{1/2} \le Q \sqrt{n} \|\gamma\|_{\ell^2}. \tag{74}$$

Therefore

$$R(\mathcal{H}_Q) \leq \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{75}$$

□

Lemma 4. *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 $y_i = f(\boldsymbol{x}_i)$, $i = 1, \cdots, n$. If $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, then there exists a unique solution h_n to the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma, \quad \textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n. \tag{76}$$

Moreover, we have

$$\|h_n - h_{\mathrm{ini}}\|_\gamma \leq \|f - h_{\mathrm{ini}}\|_\gamma. \tag{77}$$

Proof. By the definition of the FP-norm, we have $\|h_n - h_{\mathrm{ini}}\|_\gamma = \|\mathcal{F}[h]_n - \mathcal{F}[h]_{\mathrm{ini}}\|_{H_\Gamma}$. According to Corollary 5, the minimizer of problem (76) exists, i.e., h_n exists. Since the target function $f(x)$ satisfies the constraints $f(x_i) = y_i$, $i = 1, \cdots, n$, we have $\|h_n - h_{\mathrm{ini}}\|_\gamma \leq \|f - h_{\mathrm{ini}}\|_\gamma$. □

Lemma 5. *Suppose that the real-valued target function $f \in \mathcal{F}_\gamma(\Omega)$ and the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ satisfies $y_i = f(\boldsymbol{x}_i)$, $i = 1, \cdots, n$. 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*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma, \quad \textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n. \tag{78}$$

Moreover, we have

$$|\mathcal{F}[h_n - h_{\mathrm{ini}}](\mathbf{0})| \leq \|f - h_{\mathrm{ini}}\|_\infty + \|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2}. \tag{79}$$

Proof. Let $f' = f - h_{\mathrm{ini}}$. Since $h_n(\boldsymbol{x}_i) - f(\boldsymbol{x}_i) = 0$ for $i = 1, \cdots, n$, we have $h_n(\boldsymbol{x}_i) - f'(\boldsymbol{x}_i) - h_{\mathrm{ini}}(\boldsymbol{x}_i) = 0$. Therefore

$$\begin{aligned}
|\mathcal{F}[h_n - h_{\mathrm{ini}}](\mathbf{0})| &= \left| f'(\boldsymbol{x}_i) - \sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} \mathcal{F}[h_n - h_{\mathrm{ini}}](\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{k}^\intercal \boldsymbol{x}_i} \right| && (80)\\
&\leq \|f'\|_\infty + \sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} |\mathcal{F}[h_n - h_{\mathrm{ini}}](\boldsymbol{k})| && (81)\\
&\leq \|f'\|_\infty + \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h_n - h_{\mathrm{ini}}](\boldsymbol{k})|^2 \right)^{\frac{1}{2}} && (82)\\
&\leq \|f'\|_\infty + \|h_n - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2} && (83)\\
&\leq \|f'\|_\infty + \|f'\|_\gamma \|\gamma\|_{\ell^2}. && (84)
\end{aligned}$$

We remark that the last step is due to the same reason as Lemma 4. □

Theorem 4. *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 $y_i = f(\boldsymbol{x}_i)$, $i = 1, \cdots, n$, and h_n is the solution of the regularized model*

$$\min_{h - h_{\mathrm{ini}} \in \mathcal{F}_\gamma(\Omega)} \|h - h_{\mathrm{ini}}\|_\gamma, \quad \textit{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n. \tag{85}$$

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

$$L_{\mathcal{D}}(h_n) \leq \|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2} \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right). \tag{86}$$

(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

$$L_{\mathcal{D}}(h_n) \leq (\|f - h_{\mathrm{ini}}\|_\infty + 2\|f - h_{\mathrm{ini}}\|_\gamma \|\gamma\|_{\ell^2}) \left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right). \tag{87}$$

Proof. Let $f' = f - h_{\mathrm{ini}}$ and $Q = \|f'\|_\gamma$.

(i) Given $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, we set $\mathcal{H}_Q = \{h : \|h - h_{\mathrm{ini}}\|_\gamma \le Q\}$. According to Lemma 4, the solution of problem (85) $h_n \in \mathcal{H}_Q$. By the relation between generalization gap and Rademacher complexity (Bartlett & Mendelson (2002), Shalev-Shwartz & Ben-David (2014)),

$$|L_{\mathcal{D}}(h_n) - L_S(h_n)| \le 2R(\mathcal{H}_Q) + 2 \sup_{h,h' \in \mathcal{H}_Q} \|h - h'\|_\infty \sqrt{\frac{2\log(4/\delta)}{n}}. \tag{88}$$

One of the component can be bounded as follows

$$\sup_{h,h' \in \mathcal{H}_Q} \|h - h'\|_\infty \le \sup_{h \in \mathcal{H}_Q} 2\|h - h_{\mathrm{ini}}\|_\infty \tag{89}$$

$$\le \sup_{h \in \mathcal{H}_Q} 2 \max_{\boldsymbol{x}} \left| \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \mathcal{F}[h - h_{\mathrm{ini}}](\boldsymbol{k}) \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{k}^\intercal \boldsymbol{x}} \right| \tag{90}$$

$$\le \sup_{h \in \mathcal{H}_Q} 2 \sum_{\boldsymbol{k} \in \mathbb{Z}^d} |\mathcal{F}[h - h_{\mathrm{ini}}](\boldsymbol{k})| \tag{91}$$

$$\le 2 \sup_{h \in \mathcal{H}_Q} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^d} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h - h_{\mathrm{ini}}](\boldsymbol{k})|^2 \right)^{\frac{1}{2}} \tag{92}$$

$$\le 2Q\|\gamma\|_{\ell^2}. \tag{93}$$

By Lemma 3,

$$R(\mathcal{H}_Q) \le \frac{1}{\sqrt{n}} Q \|\gamma\|_{\ell^2}. \tag{94}$$

By optimization problem (85), $L_S(h_n) \le L_S(f') = 0$. Therefore we obtain

$$L_{\mathcal{D}}(h) \le \frac{2}{\sqrt{n}} \|f'\|_\gamma \|\gamma\|_{\ell^2} + 4\|f'\|_\gamma \|\gamma\|_{\ell^2} \sqrt{\frac{2\log(4/\delta)}{n}}. \tag{95}$$

(ii) Given $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ with $\gamma(\mathbf{0})^{-1} := 0$, set $c_0 = \|f'\|_\infty + \|f'\|_\gamma \|\gamma\|_{\ell^2}$. By Lemma 3, 4, and 5, define $\mathcal{H}'_Q = \{h : \|h - h_{\mathrm{ini}}\|_\gamma \le Q, |\mathcal{F}[h - h_{\mathrm{ini}}](\mathbf{0})| \le c_0\}$, we obtain

$$R(\mathcal{H}'_Q) \le \frac{1}{\sqrt{n}} \|f'\|_\infty + \frac{2}{\sqrt{n}} \|f'\|_\gamma \|\gamma\|_{\ell^2}. \tag{96}$$

Also

$$\sup_{h,h' \in \mathcal{H}'_Q} \|h - h'\|_\infty \le \sup_{h \in \mathcal{H}'_Q} 2 \sum_{\boldsymbol{k} \in \mathbb{Z}^d} |\mathcal{F}[h - h_{\mathrm{ini}}](\boldsymbol{k})| \tag{97}$$

$$\le 2 \sup_{h \in \mathcal{H}'_Q} \left[|\mathcal{F}[h - h_{\mathrm{ini}}](\mathbf{0})| + \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^2 \right)^{\frac{1}{2}} \left(\sum_{\boldsymbol{k} \in \mathbb{Z}^{d*}} (\gamma(\boldsymbol{k}))^{-2} |\mathcal{F}[h - h_{\mathrm{ini}}](\boldsymbol{k})|^2 \right)^{\frac{1}{2}} \right] \tag{98}$$

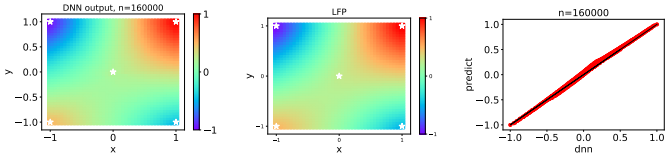
$$\le 2\|f'\|_\infty + 4\|f'\|_\gamma \|\gamma\|_{\ell^2}. \tag{99}$$

Then

$$L_{\mathcal{D}}(h_n) \le \frac{2}{\sqrt{n}} \|f'\|_\infty + \frac{4}{\sqrt{n}} \|f'\|_\gamma \|\gamma\|_{\ell^2} + (4\|f'\|_\infty + 8\|f'\|_\gamma \|\gamma\|_{\ell^2}) \sqrt{\frac{2\log(4/\delta)}{n}}. \tag{100}$$

□

Remark 3. *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 neural network due to universal approximation theory Cybenko (1989).*



(a) (b) (c)

Figure 4: Same as Fig. 2 (a,b,c) except that the training data consists of 5 points and is asymmetrical.

11 Numerical solution of the LFP model

Numerically, we solve the following ridge regression problem (Mei et al. (2019)) to approximate the solution of the optimization problem (10)

$$\min_{h} \sum_{i=1}^{n} (h(\boldsymbol{x}_i) - y_i)^2 + \lambda \sum_{\boldsymbol{\xi} \in \mathbb{L}^d} \left(\frac{\frac{1}{m} \sum_{i=1}^{m} \left(|\boldsymbol{w}_i(0)|^2 + a_i(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+3}} + \frac{4\pi^2 \frac{1}{m} \sum_{i=1}^{m} \left(|\boldsymbol{w}_i(0)|^2 a_i(0)^2 \right)}{\|\boldsymbol{\xi}\|^{d+1}} \right)^{-1} |\mathcal{F}[h](\boldsymbol{\xi})|^2, \tag{101}$$

where the truncated lattice $\mathbb{L}^d = \frac{1}{L'} \left(\mathbb{Z}^d \cap [-K+1, K-1]^d \right)$. ε is fixed to 10^{-6}. The numerical error contributed by λ is very small in a proper range. For the case of $d = 1$ as shown in Fig. 1, we choose $L' = 20$, $K = 2000$ for the computation of the solution of problem (101). For the case of $d = 2$ as shown in Figs. 2 and 4, we choose $L' = 24$, $K = 120$ for the computation of solution. For higher dimensional cases, because the size of set $\mathbb{L}^d$ grows exponentially with d, we will encounter curse of dimensionality for solving Eq. (101). Therefore, we do not test the LFP model for $d > 2$ in this work.