An Upper Limit of Decaying Rate with Respect to Frequency in Linear Frequency Principle Model

Tao Luo[1,2*]   LUOTAO41@SJTU.EDU.CN

Zheng Ma[1,2]   ZHENGMA@SJTU.EDU.CN

Zhiwei Wang[1†]   VICTORYWZW@SJTU.EDU.CN

Zhi-Qin John Xu[1]   XUZHIQIN@SJTU.EDU.CN

Yaoyu Zhang[1,3]   ZHYY.SJTU@SJTU.EDU.CN

[1]*School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China*

[2] *CMA-Shanghai*

[3] *Shanghai Center for Brain Science and Brain-Inspired Technology*

Editors: Bin Dong, Qianxiao Li, Lei Wang, Zhi-Qin John Xu

Abstract

Deep neural network (DNN) usually learns the target function from low to high frequency, which is called frequency principle or spectral bias. This frequency principle sheds light on a high-frequency curse of DNNs — difficult to learn high-frequency information. Inspired by the frequency principle, a series of works are devoted to develop algorithms for overcoming the high-frequency curse. A natural question arises: what is the upper limit of the decaying rate w.r.t. frequency when one trains a DNN? In this work, we abstract a paradigm for modeling and analysis of algorithm suitable for supervised learning problems from the sufficiently wide two-layer neural network, i.e., linear frequency principle (LFP) model. Our theory confirms that there is a critical decaying rate w.r.t. frequency in LFP model. It is precisely because of the existence of this limit that a sufficiently wide DNN interpolates the training data by a function with a certain regularity. However, if the decay rate of an algrithm in our paradigm is above such upper limit, then this algorithm interpolates the training data by a trivial function, i.e., a function is only non-zero at training data points. This work rigorously proves that the high-frequency curse is an intrinsic difficulty of the LFP model, which provides similar insight to DNN.

Keywords: Deep neural network, frequency principle, upper limit bound

1. Introduction

The study of generalization in deep learning attracts much attention in recent years due to the contradiction to the traditional wisdom (Breiman, 1995; Zhang et al., 2017), that is, over-parameterized DNNs often generalize well in real dataset. Recently, a series of works have demonstrated a frequency principle (F-Principle), an implicit bias in Fourier domain, that is, a DNN tends to learn a target function from low to high frequencies during the training (Xu et al., 2019, 2020; Rahaman et al., 2019). Xu (2018); Xu et al. (2020) propose that the low-frequency bias is due to that a function with a certain regularity decays w.r.t. frequency in the Fourier domain with a certain rate.

---

[*] Corresponding author

[†] PhD student of School of Mathematical Sciences at Shanghai Jiao Tong University.

This mechanism is further confirmed by a series of theoretical works (Luo et al., 2019; Zhang et al., 2019; Luo et al., 2020; Cao et al., 2019; Yang and Salman, 2019; Ronen et al., 2019; Bordelon et al., 2020; E et al., 2020). The frequency principle implies a rational that DNNs generalize well for real datasets, which are often low-frequency dominant (Xu et al., 2020). Meanwhile, such low-frequency bias also suggests a high-frequency curse, i.e., DNNs are difficult to learn high-frequency information. To overcome the high-frequency curse, various approaches are proposed (Xu et al., 2020; Jagtap et al., 2020; Biland et al., 2019; Cai et al., 2020; Peng et al., 2020; Cai and Xu, 2019; Liu et al., 2020; Li et al., 2020; Wang et al., 2020; Tancik et al., 2020; Mildenhall et al., 2020; Agarwal et al., 2020; Campo et al., 2020; Jiang et al., 2020; Xi et al., 2020).

A natural question is that what is the upper limit of the decaying rate w.r.t. frequency in DNN. The decaying rate means the decaying speed of the Fourier coefficient of the difference of the neural network function and the target function with respect to the frequency. Such an upper limit characterizes the boundary of the frequency bias, providing a better understanding of the implicit bias of DNNs in Fourier domain. In addition, it also provides a guidance for algorithm design of DNNs which could be more efficient in capturing high-frequency information. In this work, we use Linear Frequency Principle (LFP) (Luo et al., 2020) dynamics to model the dynamical behavior of a sufficiently wide two-layer neural network in the linear regime Jacot et al. (2018), and prove that there is a critical decaying rate w.r.t. frequency in LFP model. It is precisely because of the existence of this limit that the LFP model interpolates the training data by a function with a certain regularity. However, if the decay rate of an algrithm in our paradigm is above such upper limit, then this algorithm interpolates the training data by a trivial function, i.e., a function is only non-zero at training data points.

Theoretical works have estimated the decaying rate w.r.t. frequency follows a power law for DNNs with a certain regularity activation function in the gradient descent training (Luo et al., 2019; Zhang et al., 2019; Luo et al., 2020; Cao et al., 2019; Ronen et al., 2019; Bordelon et al., 2020). The long-time limit solution of such gradient descent training is proved to be equivalent to solving a Fourier-domain variational problem (Zhang et al., 2019; Luo et al., 2020). Inspired by above works about the F-Principle, in this paper, we propose a general Fourier-domain variational formulation for supervised learning problem, including DNNs in linear regime, and study its well-posedness. In continuum modelling, it is often difficult to impose the constraint of given values on isolated data points in a function space without sufficient regularity, e.g., a L^p space. We circumvent this difficulty by regarding the Fourier-domain variation as the primal problem and the constraint of isolated data points is imposed through a linear operator. Under a necessary and sufficient condition within our unified framework, we establish the well-posedness of the Fourier-domain variational problem. We show that the well-posedness depends on a critical exponent, which equals to the data dimension. This is a stark difference compared with a traditional partial differential equation (PDE) problem. For example, in a boundary value problem of any PDE in a d-dimensional domain, the boundary data should be prescribed on the $(d-1)$-dimensional boundary of the domain, where the dimension d plays an important role. However, in a well-posed supervised learning problem, the constraint is always on isolated points, which are 0-dimensional independent of d, while the model has to satisfy a well-posedness condition depending on the dimension. In practice, common DNNs in linear regime is a convenient way to implement our formulation. Therefore, the convergence rate of high-frequency has a upper limit. An algorithm with too fast high-frequency learning would lead to a learned function only non-zero at training data points. Such understanding of the upper limit of decaying rate indicates a better way to overcome the high-frequency curse is to design a proper

2

pre-condition approach to shift high-frequency information to low-frequency one, which coincides with several previous developed algorithms for fast learning high-frequency information (Cai et al., 2020; Cai and Xu, 2019; Liu et al., 2020; Li et al., 2020; Wang et al., 2020; Tancik et al., 2020; Mildenhall et al., 2020; Agarwal et al., 2020).

The rest of the paper is organized as follows. Section 2 shows some related work. In section 4, we propose a Fourier-domain variational formulation for supervised learning problems. The necessary and sufficient condition for the well-posedness of our model is presented in section 5. Section 6 is devoted to the numerical demonstration in which we solve the Fourier-domain variational problem using band-limited functions. Finally, we present a short conclusion and discussion in section 7.

2. Related Works

It has been an important approach to study machine learning from the perspective of implicit bias (Neyshabur et al., 2014), such as the implicit bias of training algorithms (Gunasekar et al., 2018a; Soudry et al., 2018), dropout (Mianjy et al., 2018), linear network (Gunasekar et al., 2018b) and DNNs under different initializations (Luo et al., 2021).

The low-frequency implicit bias is named as frequency principle (F-Principle) (Xu et al., 2019, 2020) or spectral bias (Rahaman et al., 2019) and can be robustly observed no matter how over-parameterized NNs are. Xu (2018); Xu et al. (2020) propose a key mechanism of the F-Principle that the regularity of the activation function converts into the decay rate of a loss function in the frequency domain. Theoretical studies subsequently show that the F-Principle holds in general setting with infinite samples (Luo et al., 2019) and in the regime of wide NNs (Neural Tangent Kernel (NTK) regime (Jacot et al., 2018)) with finite samples (Zhang et al., 2019; Luo et al., 2020) or samples distributed uniformly on sphere (Cao et al., 2019; Yang and Salman, 2019; Ronen et al., 2019; Bordelon et al., 2020). E et al. (2020) show that the integral equation would naturally leads to the F-Principle. In addition to characterizing the training speed of DNNs, the F-Principle also implicates that DNNs prefer low-frequency function and generalize well for low-frequency functions (Xu et al., 2020; Zhang et al., 2019; Luo et al., 2020).

3. Notations

In the following, we consider the regression problem of fitting a target function $f^* \in C_c(\mathbb{R}^d)$. Clearly, $f^* \in L^2(\mathbb{R}^d)$. Specifically, we use a sufficiently wide two-layer DNN, $h(\boldsymbol{x}, \boldsymbol{\theta}(t))$ with a parameter set $\boldsymbol{\theta}(t)$ which varies with the training time t, to fit the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of n sample points, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i = f^*(\boldsymbol{x}_i)$ for each i. For the convenience of notation, we denote $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^\mathsf{T}$.

It has been shown in Jacot et al. (2018); Lee et al. (2019) that, if the number of neurons in each hidden layer is sufficiently large, then $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \ll 1$ for any $t \geq 0$. In such cases, $h(\boldsymbol{x}, \boldsymbol{\theta}(t))$ can be approximated up to the first order in the Taylor expansion at $\theta(0)$, which means the the following function

$$h_{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{\theta}) = h(\boldsymbol{x}, \boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{x}, \boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{1}$$

is a very good approximation of DNN output $h(\boldsymbol{x}, \boldsymbol{\theta}(t))$ with $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. Note that, we have the following requirement for h which is easily satisfied for common DNNs: for the gradient in equation (1), we require that for any $\boldsymbol{\theta} \in \mathbb{R}^m$, there exists a weak derivative of $h(\cdot, \boldsymbol{\theta}_0)$ with respect to $\boldsymbol{\theta}$ satisfying $\nabla_{\boldsymbol{\theta}} h(\cdot, \boldsymbol{\theta}_0) \in L^2(\mathbb{R}^d)$.

A two-layer neural network is

$$h(\boldsymbol{x}, \boldsymbol{\theta}(t)) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} + b_j), \tag{2}$$

where σ is the activation function, $\boldsymbol{w}_j$ is the input weight, a_j is the output weight, b_j is the bias term.

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

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

4. Fourier-domain Variational Problem for Supervised Learning

To study the decaying rate limit w.r.t. frequency in DNN training, we propose a Fourier-domain variational problem for supervised learning, in which frequency bias can be imposed by weight term $\langle \boldsymbol{\xi} \rangle^\alpha$ in equation (4). To show the motivation and the rationality of the variational problem, we first introduce a linear frequency principle.

4.1. Motivation: Linear Frequency Principle

In the large width limit, it is reasonable (Jacot et al., 2018; Lee et al., 2019) to assume a linear condition, i.e., $h(\boldsymbol{x}, \boldsymbol{\theta}) = h_{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{\theta})$. Based on the linear condition, Zhang et al. (2019); Luo et al. (2020) derived a Linear F-Principle (LFP) dynamics to effectively study the training dynamics of a two-layer NN with the mean square loss in the large width limit. Up to a multiplicative constant in the time scale, the gradient descent dynamics of a sufficiently wide two-layer NN is approximated by

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = -(\gamma(\boldsymbol{\xi}))^2 \mathcal{F}[u_\rho](\boldsymbol{\xi}), \tag{3}$$

where $u(\boldsymbol{x}, t) = h(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f^*(\boldsymbol{x})$, $u_\rho(\boldsymbol{x}, t) = u(\boldsymbol{x}, t)\rho(\boldsymbol{x})$, $\rho(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, accounting for the real case of a finite training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and $\gamma(\boldsymbol{\xi})$ depends on the initialization and frequency. For ReLU activation function,

$$(\gamma(\boldsymbol{\xi}))^2 = \mathbb{E}_{a(0), r(0)} \left[\frac{r(0)^3}{16\pi^4 \|\boldsymbol{\xi}\|^{d+3}} + \frac{a(0)^2 r(0)}{4\pi^2 \|\boldsymbol{\xi}\|^{d+1}} \right],$$

where $r(0) = |\boldsymbol{w}(0)|$ and the two-layer NN parameters at initial $a(0)$ and $\boldsymbol{w}(0)$ are random variables with certain given distribution. The reason why there are $1/\|\boldsymbol{\xi}\|^{d+1}$ and $1/\|\boldsymbol{\xi}\|^{d+3}$ is explained with example $d = 1$ as follows. In the gradient flow, the gradient of $h(x, \boldsymbol{\theta})$ w.r.t. a_j leaves a ReLU function, which decays in $1/|\xi|^2$, and the gradient of $h(x, \boldsymbol{\theta})$ w.r.t. w_j leaves a Heaviside function, which decays in $1/|\xi|$. Due to the chain rule, when we considering the evolution of the function, there will be square of gradients, thus, there are $1/\|\boldsymbol{\xi}\|^2$ and $1/\|\boldsymbol{\xi}\|^4$.

The dynamics in equation (3) shows the low frequency components converges faster than high frequency components. The RHS is only affected by the training data due to the sample distribution, that is, this dynamics stops evolution only when the error on training data is zero. However, the LHS contains the Fourier transform of the whole fitting function. Therefore, the fitting function beyond the training data can be inferred. The number of functions that satisfy the constraints on training

data is infinite. It can be proved that the faster convergence of low-frequency components leads to that the model prefers to select a function with large low-frequency coefficients. Zhang et al. (2019); Luo et al. (2020) rigorously deduced the solution of the LFP model (3) is equivalent to that of the following optimization problem in a proper hypothesis space F_γ,

$$\min_{h-h_{\text{ini}}\in F_\gamma} \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2} |\mathcal{F}[h](\boldsymbol{\xi}) - \mathcal{F}[h_{\text{ini}}](\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi},$$

subject to constraints $h(\boldsymbol{x}_i) = y_i$ for $i = 1, \ldots, n$. Here, $h_{\text{ini}} = h(\boldsymbol{x}, \boldsymbol{\theta}(0))$ is the original output of DNN. The weight $(\gamma(\boldsymbol{\xi}))^{-2}$ grows as the frequency $\boldsymbol{\xi}$ increases, which means that a large penalty is imposed on the high frequency part of $h(\boldsymbol{x}) - h_{\text{ini}}(\boldsymbol{x})$. Based on such analysis, $h(x) - h_{\text{ini}}(x)$ is a low-frequency function (which is a function whose Fourier transform has large value only near the origin). But if $h_{\text{ini}}(x)$ is not 0, $h(x)$ may be not a low-frequency function. In most cases, the target function in the real world is a low-frequency function. Thus, the low-frequency function is usually more likely to have good generalization in real data set. Therefore, 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., 2020), which guarantees $h_{\text{ini}}(\boldsymbol{x}) = 0$. Then the final output $h(\boldsymbol{x})$ is dominated by low frequency, and the DNN model possesses a good generalization.

4.2. Fourier-domain Variational Formulation

Inspired by the variational formulation of LFP model, we propose a new continuum model for the supervised learning, which includes DNNs with gradient flow learning. In the minimization problem proposed above, for large $\|\boldsymbol{\xi}\|$, the main component of the weight term is $\|\boldsymbol{\xi}\|^{d+1}$. Therefore, after ignoring the constant and low-order term, we can use $\langle\boldsymbol{\xi}\rangle^{d+1}$ to replace this weight item, where $\langle\boldsymbol{\xi}\rangle = (1 + \|\boldsymbol{\xi}\|^2)^{\frac{1}{2}}$ is the "Japanese bracket" of $\boldsymbol{\xi}$. We use this form to connect our variational problem with Sobolev space, which is convenient for us to carry out theoretical analysis.

Further, we can replace the index $d + 1$ by a more general constant α to be determined later. Thus we obtain the following variational problem:

$$\min_{h\in\mathcal{H}} Q_\alpha[h] = \int_{\mathbb{R}^d} \langle\boldsymbol{\xi}\rangle^\alpha |\mathcal{F}[h](\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi}, \tag{4}$$

$$\text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n, \tag{5}$$

where $\mathcal{H} = \{h(x)|\int_{\mathbb{R}^d} \langle\boldsymbol{\xi}\rangle^\alpha |\mathcal{F}[h](\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi} < \infty\}$. According to the equivalent theorem in Luo et al. (2020), $-\alpha$ is the decaying rate w.r.t. frequency in the gradient flow dynamics in (4). In this work, we study how the property of the solution in the variational problem depends on α.

The solution of the variational problem is expected to have a sufficient regularity, e.g., it belongs to L^2 space. But in the spatial domain, the evaluation on n known data points is meaningless in the sence of L^2 function since isolated points cannot reflect the overall regularity of functions in L^2 space. Therefore, we consider the problem in the frequency domain and define a linear operator $\mathcal{P}_{\boldsymbol{X}} : L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \to \mathbb{R}^n$ for the given sample set $\boldsymbol{X}$ to transform the original constraints into the ones in the Fourier domain: $\mathcal{P}_{\boldsymbol{X}}\phi^* = \boldsymbol{Y}$. More precisely, we define for $\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$

$$\mathcal{P}_{\boldsymbol{X}}\phi := \left(\int_{\mathbb{R}^d} \phi(\boldsymbol{\xi})e^{2\pi i\boldsymbol{\xi}\cdot\boldsymbol{x}_1} \, \mathrm{d}\boldsymbol{\xi}, \cdots, \int_{\mathbb{R}^d} \phi(\boldsymbol{\xi})e^{2\pi i\boldsymbol{\xi}\cdot\boldsymbol{x}_n} \, \mathrm{d}\boldsymbol{\xi} \right)^{\mathsf{T}}. \tag{6}$$

The admissible function class reads as

$$\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}} = \{\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \mid \mathcal{P}_{\boldsymbol{X}}\phi = \boldsymbol{Y}\}.$$

Notice that $\|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}} = \left(\int_{\mathbb{R}^d}\langle\boldsymbol{\xi}\rangle^{\alpha}|\phi(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}\right)^{\frac{1}{2}}$ is a Sobolev norm, which characterizes the regularity of the final output function $h(\boldsymbol{x}) = \mathcal{F}^{-1}[\phi](\boldsymbol{x})$. The larger the exponent α is, the better the regularity becomes. For example, when $d = 1$ and $\alpha = 2$, by Parseval's theorem,

$$\|u\|_{H^1}^2 = \int_{\mathbb{R}}(1 + |\xi|^2)|\mathcal{F}[u](\xi)|^2\,\mathrm{d}\xi = \int_{\mathbb{R}} u^2 + \frac{1}{4\pi^2}|\nabla u|^2\,\mathrm{d}x.$$

Accordingly, the Fourier-domain variational problem reads as a standard variational problem in spatial domain. This is true for any quadratic Fourier-domain variational problem, but of course our Fourier-domain variational formulation is not necessarily being quadratic. The details for general cases (non-quadratic ones) are left to future work. For the quadratic setting with exponent α, i.e., Problem (4), it is roughly equivalent to the following spatial-domain variational problem:

$$\min \int_{\mathbb{R}^d}(u^2 + |\nabla^{\frac{\alpha}{2}}u|^2)\,\mathrm{d}x.$$

This is clear for integer $\alpha/2$, while fractional derivatives are required for non-integer $\alpha/2$.

Back to our problem, after the above transformation, our goal is transformed into studying the following Fourier-domain variational problem,

Problem 1 *Find a minimizer ϕ^* in $\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ such that*

$$\phi^* \in \arg\min_{\phi\in\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}\|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}}^2. \tag{7}$$

We should point out that this variational problem is a paradigm for supervised learning problems, and a two-layer infinitely wide neural network corresponds to the special case of taking $\alpha = d + 1$. And not all α can be obtained by the neural network with commonly-used activation functions. However, by analogy, we believe that the infinitely wide multi-layer neural network in NTK regime also belongs to the paradigm we proposed and corresponds the case $\alpha > d$.

We remark that the operator $\mathcal{P}_{\boldsymbol{X}}$ is the inverse Fourier transform with evaluations on sample points $\boldsymbol{X}$. Actually, the linear operator $\mathcal{P}_{\boldsymbol{X}}$ projects a function defined on $\mathbb{R}^d$ to a function defined on 0-dimensional manifold $\boldsymbol{X}$. Just like the (linear) trace operator T in a Sobolev space projects a function defined on d-dimensional manifold into a function defined on $(d-1)$-dimensional boundary manifold. Note that the only function space over the 0-dimensional manifold $\boldsymbol{X}$ is the n-dimensional vector space $\mathbb{R}^n$, where n is the number of data points, while any Sobolev (or Besov) space over d-dimensional manifold ($d \geq 1$) is an infinite dimensional vector space.

5. The Critical Decaying Rate

In this section, we consider a critical exponent for α, which leads to the existence/non-existence dichotomy to Problem 1. We first prove that there is no solution to the Problem 1 in subcritical case $\alpha < d$, and for $\alpha > d$ the optimal function is a continuous and nontrivial solution (See proof in Appendix.). Therefore, we conclude that to obtain a non-trivial interpolation among training data for supervised learning, such as DNN fitting, the decaying rate of high-frequency information can not be too fast, i.e., there exists a upper limit of the decaying rate w.r.t. frequency.

5.1. Subcritical Case: $\alpha < d$

In order to prove the nonexistence of the solution to the Problem 1 in $\alpha < d$ case, at first we need to find a class of functions that make the norm tend to zero. Let $\psi_\sigma(\boldsymbol{\xi}) = (2\pi)^{\frac{d}{2}}\sigma^d e^{-2\pi^2\sigma^2\|\boldsymbol{\xi}\|^2}$, then by direct calculation, we have $\mathcal{F}^{-1}[\psi_\sigma](\boldsymbol{x}) = e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}}$. For $\alpha < d$ the following proposition shows that the norm $\|\mathcal{F}^{-1}[\psi_\sigma]\|^2_{H^{\frac{\alpha}{2}}}$ can be sufficiently small as $\sigma \to 0$.

Proposition 1 (critical exponent) *For any input dimension d, we have*

$$\lim_{\sigma \to 0}\|\mathcal{F}^{-1}[\psi_\sigma]\|^2_{H^{\frac{\alpha}{2}}} = \begin{cases} 0, & \alpha < d, \\ C_d, & \alpha = d, \\ \infty, & \alpha > d. \end{cases} \tag{8}$$

Here the constant $C_d = \frac{1}{2}(d-1)!(2\pi)^{-d}\frac{2\pi^{d/2}}{\Gamma(d/2)}$ only depends on the dimension d.

Remark 1 *The function $\mathcal{F}^{-1}[\psi]$ can be any function in the Schwartz space, not necessarily Gaussian. Proposition 1 still holds with (possibly) different C_d.*

For every small σ, we can use n rapidly decreasing functions $\mathcal{F}^{-1}[\psi_\sigma](\boldsymbol{x} - \boldsymbol{x}_i)$ to construct the solution $\mathcal{F}^{-1}[\phi_\sigma](\boldsymbol{x})$ of the supervised learning problem. However, according to Proposition 1, when the parameter σ tends to 0, the limit is the zero function in the sense of $L^2(\mathbb{R}^d)$. Therefore we have the following theorem:

Theorem 1 (non-existence) *Suppose that $\boldsymbol{Y} \neq \boldsymbol{0}$. For $\alpha < d$, there is no function $\phi^* \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ satisfying*

$$\phi^* \in \arg\min_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}\|\mathcal{F}^{-1}[\phi]\|^2_{H^{\frac{\alpha}{2}}}.$$

In other words, there is no solution to the Problem 1.

5.2. Supercritical Case: $\alpha > d$

We then provide a theorem to establish the existence of the minimizer for Problem 1 in the case of $\alpha > d$.

Theorem 2 (existence) *For $\alpha > d$, there exists $\phi^* \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ satisfying*

$$\phi^* \in \arg\min_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}\|\mathcal{F}^{-1}[\phi]\|^2_{H^{\frac{\alpha}{2}}}.$$

In other words, there exists a solution to the Problem 1.

Remark 2 *Note that, according to the Sobolev embedding theorem (Adams and Fournier, 2003; Evans, 1999), the minimizer in Theorem 2 has smoothness index no less than $[\frac{\alpha-d}{2}]$.*

6. Numerical Results

In this section, we illustrate our results by solving Fourier-domain variational problems numerically. We use uniform mesh in frequency domain with mesh size $\Delta\xi$ and band limit $M\Delta\xi$. In this discrete setting, the considered space becomes $\mathbb{R}^{(2M)^d}$. We emphasize that the numerical solution with this setup always exists even for the subcritical case which corresponds to the non-existence theorem. However, as we will show later, the numerical solution is trivial in nature when $\alpha < d$.

6.1. Special Case: One Data Point in One Dimension

To simplify the problem, we start with a single point $X = 0 \in \mathbb{Z}$ with the label $Y = 2$. Denote $\phi_j = \phi(\xi_j)$ for $j \in \mathbb{Z}$. We also assume that the function ϕ is an even function. Then according to the definition of $\mathcal{P}_X$, we have the following problem:

Example 1 (Problem 1 with a particular discretization)

$$\min_{\phi \in \mathbb{R}^M} \sum_{j=1}^{M} (1 + j^2 \Delta \xi^2)^{\frac{\alpha}{2}} |\phi_j|^2, \tag{9}$$

$$\text{s.t.} \quad \sum_{j=1}^{M} \phi_j \Delta \xi = 1, \tag{10}$$

If we denote $\phi = (\phi_1, \phi_2, \ldots, \phi_M)^\mathsf{T}$, $b = \frac{1}{\Delta \xi}$, $A = (1, 1, \ldots, 1) \in \mathbb{R}^M$ and $\Gamma \in \mathbb{R}^M$ where $\Gamma_{ii} = \sqrt{\lambda}(1 + i^2 \Delta \xi^2)^{\frac{\alpha}{4}}$, $i = 1, 2, \cdots, M$.

In fact this is a standard Tikhonov regularization (Tikhonov and Arsenin, 1977), also known as ridge regression problem with the Lagrange multiplier λ. The corresponding ridge regression problem is,

$$\min_{\phi} \|A\phi - b\|_2^2 + \|\Gamma\phi\|_2^2, \tag{11}$$

This problem admits an explicit and unique solution (Tikhonov and Arsenin, 1977),

$$\phi = (A^\mathsf{T} A + \Gamma^\mathsf{T} \Gamma)^{-1} A^\mathsf{T} b. \tag{12}$$

By using the inverse discrete Fourier transform, we obtain the final output function in x space:

$$h(x) = \frac{2}{(Z^2 + \lambda)} \sum_{j=1}^{M} (1 + j^2 \Delta \xi^2)^{-\frac{\alpha}{2}} \cos(2\pi j x). \tag{13}$$

Fig. 1 shows that for this special case with a large M, $h(x)$ is not an trivial function in $\alpha > d$ case and degenerates to a trivial function in $\alpha < d$ case.

6.2. General Case: n Points in d Dimension

Assume that we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and each data point has d components which means $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^\mathsf{T}$, and denote the corresponding label as $(y_1, y_2, \ldots, y_n)^\mathsf{T}$. For the sake of simplicity, we denote the vector $(j_1, j_2, \cdots, j_d)^\mathsf{T}$ by $\boldsymbol{J}_{j_1 \ldots j_d}$. Then our problem becomes

Example 2 (Problem 1 with general discretization)

$$\min_{\phi \in \mathbb{R}^{(2M)^d}} \sum_{j_1, \ldots, j_d = -M}^{M} (1 + \|\boldsymbol{J}_{j_1 \ldots j_d}\|^2 \Delta \xi^2)^{\frac{\alpha}{2}} |\phi_{j_1 \ldots j_d}|^2, \tag{14}$$

$$\text{s.t.} \quad \sum_{j_1, \ldots, j_d = -M}^{M} \phi_{j_1 \ldots j_d} e^{2\pi i \Delta \xi \boldsymbol{J}_{j_1 \ldots j_d}^\mathsf{T} \boldsymbol{x}_k} = y_k, \quad k = 1, 2, \ldots, d \tag{15}$$
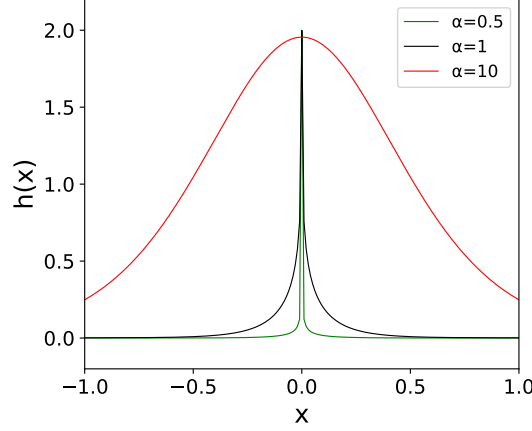
Figure 1: Fitting the function $h(x)$ shown in equation (13) with different exponent α's. Here we take $M = 10^6$, $\Delta\xi = 0.01$, $\lambda = 1$ and different α and observe that $h(x)$ is not an trivial function in $\alpha > d$ case and degenerates to a trivial function in $\alpha < d$ case.
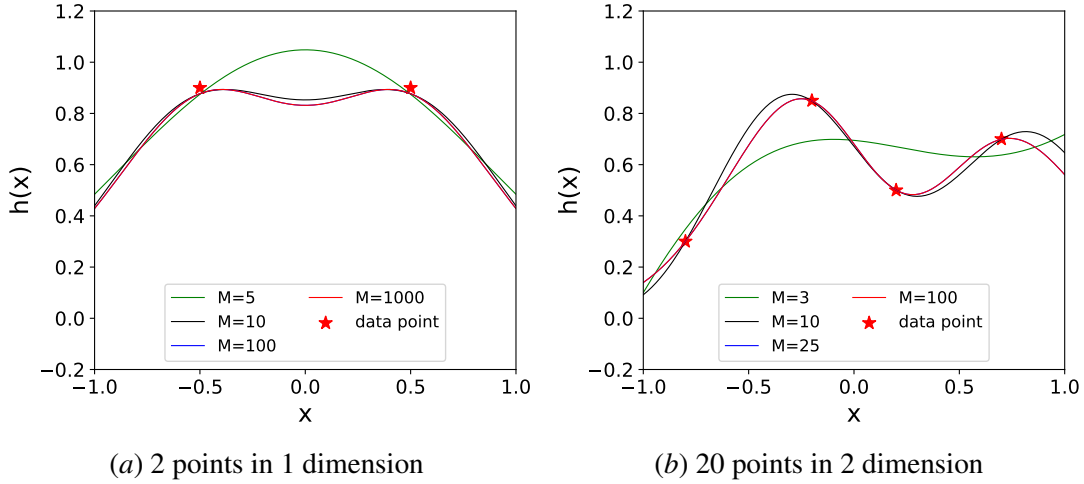

(*a*) 2 points in 1 dimension

(*b*) 20 points in 2 dimension

Figure 2: Fitting data points in different dimensions with different band limit M. We use a proper α ($\alpha > d$) and observe that even for a large M, the function $h(x)$ does not degenerate to a trivial function. Note that the blue curve and the red one overlap with each. Here the trivial function represents a function whose value decays rapidly to zero except for the given training points.

The calculation of this example can be completed by the method analogous to the one used in Example 6.1.

In Fig.2, we set $\alpha = 10$ in both cases to ensure $\alpha > d$ and change the band limit M. We observe that as M increases, the fitting curve converges to a non-trivial curve. In Fig.3, we set $M = 1000$ in 1-dimensional case and $M = 100$ in 2-dimensional case. By changing exponent α, we can see in all cases, the fitting curves are non-trivial when $\alpha > d$, but degenerate when $\alpha < d$.

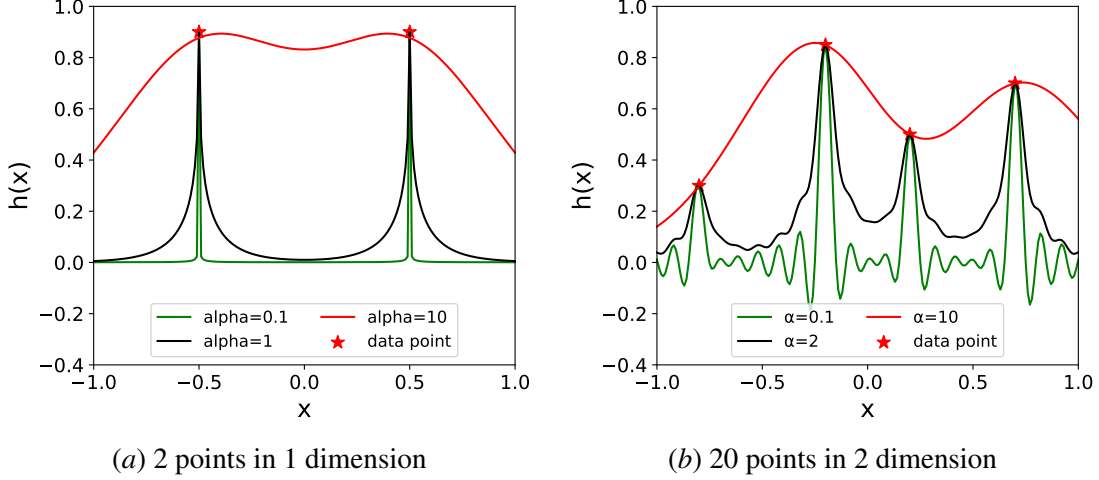*(a)* 2 points in 1 dimension   *(b)* 20 points in 2 dimension

Figure 3: Fitting data points in different dimensions with different exponent α's. We observe that with a proper M, the function $h(x)$ is not a trivial function for $\alpha > d$ case and degenerates to a trivial function for $\alpha < d$ case.

7. Conclusion

To understand the limit of the frequency bias in DNNs, we propose a Fourier-domain variational formulation based on the LFP model and establish the sufficient and necessary conditions for the well-posedness of the Fourier-domain variational problem, followed by numerical demonstration. Our work suggests that there is a upper limit of the decaying rate w.r.t. frequency, i.e., high frequency cannot converge too fast, in order to obtain a nontrivial solution in LFP model, thus, pointing out the intrinsic high-frequency curse. For two-layer infinite-width neural networks, existing works have shown their solutions are equivalent to the solutions of particular Fourier-domain variational problems (Luo et al., 2020). However, for general non-linear DNNs, this equivalence is only qualitative.

In addition, our Fourier-domain variational formulation provides a novel viewpoint for modelling machine learning problem, that is, imposing more constraints, e.g., higher regularity, on the model rather than the data (always isolated points in practice) can give us the well-posedness as dimension of the problem increases. This is different from the modelling in physics and traditional point cloud problems, in which the model is independent of dimension in general. Our work suggests a potential approach of algorithm design by considering a dimension-dependent model for data modelling.

Acknowledgments

References

Robert A. Adams and John J.F. Fournier. *Sobolev spaces*. Elsevier Science, 2003.

Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *arXiv preprint arXiv:2004.13912*, 2020.

Simon Biland, Vinicius C Azevedo, Byungsoo Kim, and Barbara Solenthaler. Frequency-aware reconstruction of fluid simulations with generative networks. *arXiv preprint arXiv:1912.08776*, 2019.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Leo Breiman. Reflections after refereeing papers for nips. *The Mathematics of Generalization*, XX: 11–15, 1995.

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

Wei Cai, Xiaoguang Li, and Lizuo Liu. A phase shift deep neural network for high frequency approximation and wave problems. *SIAM Journal on Scientific Computing*, 42(5):A3285–A3312, 2020.

Miguel Campo, Zhengxing Chen, Luke Kung, Kittipat Virochsiri, and Jianyu Wang. Band-limited soft actor critic model. *arXiv preprint arXiv:2006.11431*, 2020.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint, I. *Science China Mathematics*, pages 1–34, 2020.

Lawrence C. Evans. Partial differential equations. *Mathematical Gazette*, 83(496):185, 1999.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018b.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

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

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for generative models. *arXiv preprint arXiv:2012.12821*, 2020.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8572–8583, 2019.

Xi-An Li, Zhi-Qin John Xu, and Lei Zhang. A multi-scale DNN algorithm for nonlinear elliptic equations with multiple scales. *Communications in Computational Physics*, 28(5):1886–1906, 2020.

Ziqi Liu, Wei Cai, and Zhi-Qin John Xu. Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains. *Communications in Computational Physics*, 28(5):1970–2001, 2020.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *arXiv preprint arXiv:1906.09235*, 2019.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. On the exact computation of linear frequency principle dynamics and its generalization. *arXiv preprint arXiv:2010.08153*, 2020.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. In *International Conference on Machine Learning*, pages 3540–3548. PMLR, 2018.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Wei Peng, Weien Zhou, Jun Zhang, and Wen Yao. Accelerating physics-informed neural network training with prior dictionaries. *arXiv preprint arXiv:2004.08151*, 2020.

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

Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, pages 4763–4772, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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

Andrej Nikolaevich Tikhonov and Vasiliy Yakovlevich Arsenin. Solutions of ill-posed problems. *Mathematics of Computation*, 32(144):491–491, 1977.

Feng Wang, Alberto Eljarrat, Johannes Müller, Trond R Henninen, Rolf Erni, and Christoph T Koch. Multi-resolution convolutional neural networks for inverse problems. *Scientific reports*, 10(1):1–11, 2020.

Yue Xi, Wenjing Jia, Jiangbin Zheng, Xiaochen Fan, Yefan Xie, Jinchang Ren, and Xiangjian He. Drl-gan: Dual-stream representation learning gan for low-resolution image classification in uav applications. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 2020.

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. *International Conference on Neural Information Processing*, pages 264–274, 2019.

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

Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018.

Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. Explicitizing an implicit bias of the frequency principle in two-layer neural networks. *arXiv:1905.10264*, May 2019.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. In Jianfeng Lu and Rachel Ward, editors, *Proceedings*

of The First Mathematical and Scientific Machine Learning Conference, volume 107, pages 144–164, 2020.

Appendix A. Lemma 1

Lemma 1 *Let the function $\psi_\sigma(\boldsymbol{\xi}) = (2\pi)^{\frac{d}{2}}\sigma^d \mathrm{e}^{-2\pi^2\sigma^2\|\boldsymbol{\xi}\|^2}$, $\boldsymbol{\xi} \in \mathbb{R}^d$. We have*

$$\lim_{\sigma\to 0}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|^\alpha|\psi_\sigma(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} = \begin{cases} 0, & \alpha < d, \\ C_d, & \alpha = d, \\ \infty, & \alpha > d. \end{cases} \tag{16}$$

Here the constant $C_d = \frac{1}{2}(d-1)!(2\pi)^{-d}\frac{2\pi^{d/2}}{\Gamma(d/2)}$ only depends on the dimension d.

Proof In fact,

$$\lim_{\sigma\to 0}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|^\alpha|\psi_\sigma(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} = \lim_{\sigma\to 0}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|^\alpha(2\pi)^d\sigma^{2d}\mathrm{e}^{-4\pi^2\sigma^2\|\boldsymbol{\xi}\|^2}\,\mathrm{d}\boldsymbol{\xi}$$

$$= \lim_{\sigma\to 0}(2\pi)^d\sigma^{d-\alpha}\int_{\mathbb{R}^d}\|\sigma\boldsymbol{\xi}\|^\alpha \mathrm{e}^{-4\pi^2\|\sigma\boldsymbol{\xi}\|^2}\,\mathrm{d}(\sigma\boldsymbol{\xi})$$

$$= \lim_{\sigma\to 0}(2\pi)^d\sigma^{d-\alpha}\int_0^\infty r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r \cdot \omega_d,$$

where $\omega_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$ is the surface area of a unit $(d-1)$-sphere.

Notice that

$$\int_0^\infty r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r = \int_0^1 r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r + \int_1^\infty r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r$$

$$\leq \int_0^\infty \mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r + \int_0^\infty r^{[\alpha]+d}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r$$

$$= \frac{1}{8\pi^{\frac{3}{2}}} + \int_0^\infty r^{[\alpha]+d}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r$$

and

$$\int_0^\infty r^{[\alpha]+d}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r = \begin{cases} \frac{1}{2}\left(\frac{[\alpha]+d-1}{2}\right)!(2\pi)^{-([\alpha]+d+1)}, & [\alpha]+d \text{ is odd}, \\ \frac{\sqrt{\pi}}{2}(2\pi)^{-([\alpha]+d+1)}(\frac{1}{2})^{\frac{[\alpha]+d}{2}}([\alpha]+d-1)!!, & [\alpha]+d \text{ is even}. \end{cases}$$

Therefore, in both cases, the integral $\int_0^\infty r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r$ is finite. Then we have

$$\lim_{\sigma\to 0}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|^\alpha|\psi_\sigma(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} = \lim_{\sigma\to 0}(2\pi)^d\sigma^{d-\alpha}\int_0^\infty r^{\alpha+d-1}\mathrm{e}^{-4\pi^2 r^2}\,\mathrm{d}r \cdot \omega_d$$

$$= \begin{cases} 0, & \alpha < d, \\ \infty, & \alpha > d. \end{cases}$$

When $\alpha = d$, it follows that

$$\int_0^\infty r^{\alpha+d-1} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r = \frac{1}{2}(2\pi)^{-2d}(d-1)!.$$

Therefore

$$\lim_{\sigma \to 0} \int_{\mathbb{R}^d} \|\boldsymbol{\xi}\|^\alpha |\psi_\sigma(\xi)|^2 \, \mathrm{d}\xi = \frac{1}{2}(d-1)!(2\pi)^{-d} \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)},$$

which completes the proof. ∎

Appendix B. Proof of Proposition 1

Proof Similar to the proof of Lemma 1, we have

$$\lim_{\sigma \to 0} \|\mathcal{F}^{-1}[\psi_\sigma]\|_{H^{\frac{\alpha}{2}}}^2 = \lim_{\sigma \to 0} (2\pi)^d \sigma^{d-\alpha} \int_{\mathbb{R}^d} (\sigma^2 + \|\sigma\boldsymbol{\xi}\|^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2\|\sigma\boldsymbol{\xi}\|^2} \, \mathrm{d}(\sigma\boldsymbol{\xi})$$

$$= \lim_{\sigma \to 0} (2\pi)^d \sigma^{d-\alpha} \int_0^\infty r^{d-1} (\sigma^2 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d.$$

For $\sigma < 1$, the following integrals are bounded from below and above, respectively:

$$\int_0^\infty r^{d-1}(\sigma^2 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r \geq \int_0^\infty r^{\alpha+d-1} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r = C_1 > 0,$$

and

$$\int_0^\infty r^{d-1}(\sigma^2 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r \leq \int_0^1 r^{d-1}(1 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r + \int_1^\infty r^{d-1}((2r)^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r$$

$$\leq \int_0^1 r^{d-1}(1 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r + 2^\alpha \int_0^\infty r^{\alpha+d-1} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r$$

$$= C_2 < \infty,$$

where $C_1 = \int_0^\infty r^{\alpha+d-1} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r$ and $C_2 = \int_0^1 r^{d-1}(1+r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r + 2^\alpha \int_0^\infty r^{\alpha+d-1} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r$.
Therefore, we obtain the results for the subcritical ($\alpha < d$) and supercritical ($\alpha > d$) cases

$$\lim_{\sigma \to 0} \|\mathcal{F}^{-1}[\psi_\sigma]\|_{H^{\frac{\alpha}{2}}}^2 = \lim_{\sigma \to 0} (2\pi)^d \sigma^{d-\alpha} \int_0^\infty r^{d-1}(\sigma^2 + r^2)^{\frac{\alpha}{2}} \mathrm{e}^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d$$

$$= \begin{cases} 0, & \alpha < d, \\ \infty, & \alpha > d. \end{cases}$$

For the critical case $\alpha = d$, we have

$$
\lim_{\sigma \to 0} \| \mathcal{F}^{-1}[\psi_\sigma] \|^2_{H^{\frac{\alpha}{2}}}
$$

$$
= \lim_{\sigma \to 0} (2\pi)^d \int_0^\infty r^{d-1} (\sigma^2 + r^2)^{\frac{\alpha}{2}} e^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d
$$

$$
= \lim_{\sigma \to 0} (2\pi)^d \int_0^\infty r^{2d-1} e^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d + \lim_{\sigma \to 0} \left[\frac{\alpha}{2} (2\pi)^d \sigma^2 \int_0^\infty r^{2d-3} e^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d + o(\sigma^2) \right]
$$

$$
= \lim_{\sigma \to 0} (2\pi)^d \int_0^\infty r^{2d-1} e^{-4\pi^2 r^2} \, \mathrm{d}r \cdot \omega_d
$$

$$
= \frac{1}{2} (d-1)! (2\pi)^{-d} \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}.
$$

Therefore the proposition holds. ∎

Appendix C. Proof of Theorem 1

Proof Given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^\mathsf{T}$, let $\boldsymbol{A} = \left(\exp(-\frac{\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2}{2\sigma^2}) \right)_{n \times n}$ be an $n \times n$ matrix. For sufficiently small σ, the matrix $\boldsymbol{A}$ is diagonally dominant, and hence invertible. So the linear system $\boldsymbol{A} \boldsymbol{g}^{(\sigma)} = \boldsymbol{Y}$ has a solution $\boldsymbol{g}^{(\sigma)} = \left(g_1^{(\sigma)}, g_2^{(\sigma)}, \cdots, g_n^{(\sigma)} \right)^\mathsf{T}$. Let

$$
\phi_\sigma(\boldsymbol{\xi}) = \sum_i g_i^{(\sigma)} e^{-2\pi i \boldsymbol{\xi}^\mathsf{T} \boldsymbol{x}_i} \psi_\sigma(\boldsymbol{\xi}),
$$

where $\psi_\sigma(\boldsymbol{\xi}) = (2\pi)^{\frac{d}{2}} \sigma^d e^{-2\pi^2 \sigma^2 \|\boldsymbol{\xi}\|^2}$ satisfying $\mathcal{F}^{-1}[\psi_\sigma](\boldsymbol{x}) = e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}}$. Thus

$$
\mathcal{F}^{-1}[\phi_\sigma](\boldsymbol{x}) = \sum_i g_i^{(\sigma)} \mathcal{F}^{-1}[\psi_\sigma](\boldsymbol{x} - \boldsymbol{x}_i) = \sum_i g_i^{(\sigma)} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}}.
$$

In particular, for all $i = 1, 2, \cdots, n$

$$
\mathcal{F}^{-1}[\phi_\sigma](\boldsymbol{x}_i) = \sum_j g_j^{(\sigma)} e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}} = (\boldsymbol{A} \boldsymbol{g}^{(\sigma)})_i = y_i.
$$

Therefore, $\phi_\sigma \in \mathcal{A}_{\boldsymbol{X}, \boldsymbol{Y}}$ for sufficiently small $\sigma > 0$.

According to the above discussion, we can construct a sequence $\{\phi_{\frac{1}{m}}\}_{m=M}^\infty \subset \mathcal{A}_{\boldsymbol{X}, \boldsymbol{Y}}$, where M is a sufficiently large positive integer to make the matrix $\boldsymbol{A}$ invertible. As Proposition 1 shows,

$$
\lim_{m \to +\infty} \| \mathcal{F}^{-1}[\phi_{\frac{1}{m}}] \|^2_{H^{\frac{\alpha}{2}}} = 0.
$$

Now, suppose that there exists a solution to the Problem 1, denoted as $\phi^* \in \mathcal{A}_{\boldsymbol{X}, \boldsymbol{Y}}$. By definition,

$$
\| \mathcal{F}^{-1}[\phi^*] \|^2_{H^{\frac{\alpha}{2}}} \le \min_{\phi \in \mathcal{A}_{\boldsymbol{X}, \boldsymbol{Y}}} \| \mathcal{F}^{-1}[\phi] \|^2_{H^{\frac{\alpha}{2}}} \le \lim_{m \to +\infty} \| \mathcal{F}^{-1}[\phi_{\frac{1}{m}}] \|^2_{H^{\frac{\alpha}{2}}} = 0.
$$

Therefore, $\phi^*(\boldsymbol{\xi}) \equiv 0$ and $\mathcal{P}_{\boldsymbol{X}} \phi^* = \boldsymbol{0}$, which contradicts to the restrictive condition $\mathcal{P}_{\boldsymbol{X}} \phi^* = \boldsymbol{Y}$ for the situation that $\boldsymbol{Y} \ne \boldsymbol{0}$. The proof is completed. ∎

Appendix D. Proof of Theorem 2

Proof 1. We introduce a distance for functions $\phi, \psi \in L^2(\mathbb{R}^d)$:

$$\mathrm{dist}(\phi, \psi) = \|\mathcal{F}^{-1}[\phi] - \mathcal{F}^{-1}[\psi]\|_{H^{\frac{\alpha}{2}}}.$$

Under the topology induced by this distance, the closure of the admissible function class $\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ reads as

$$\overline{\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}} := \overline{\{\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \mid \mathcal{P}_{\boldsymbol{X}}\phi = \boldsymbol{Y}\}}^{\,\mathrm{dist}(\cdot,\cdot)}.$$

2. We will consider an auxiliary minimization problem: to find ϕ^* such that

$$\phi^* \in \arg\min_{\phi \in \overline{\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}}. \tag{17}$$

Let $m := \inf_{\phi \in \overline{\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}}$. According to the proof of Proposition 1 and Theorem 1, for a small enough $\sigma > 0$, the inverse Fourier transform of function

$$\phi_\sigma(\boldsymbol{\xi}) = \sum_i g_i^{(\sigma)} \mathrm{e}^{-2\pi\mathrm{i}\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x}_i} \psi_\sigma(\boldsymbol{\xi})$$

has finite Sobolev norm $\|\mathcal{F}^{-1}[\phi_\sigma]\|_{H^{\frac{\alpha}{2}}} < \infty$, where $\psi_\sigma(\boldsymbol{\xi})$ satisfies $\mathcal{F}^{-1}[\psi_\sigma](\boldsymbol{x}) = \mathrm{e}^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}}$, $\boldsymbol{A} = \left(\exp(-\frac{\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2}{2\sigma^2})\right)_{n \times n}$ and $\boldsymbol{g}^{(\sigma)} = \left(g_1^{(\sigma)}, g_2^{(\sigma)}, \cdots, g_n^{(\sigma)}\right)^{\mathsf{T}} = \boldsymbol{A}^{-1}\boldsymbol{Y}$. Thus $m < +\infty$.

3. Choose a minimizing sequence $\{\bar{\phi}_k\}_{k=1}^{\infty} \subset \overline{\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}$ such that

$$\lim_{k\to\infty} \|\mathcal{F}^{-1}[\bar{\phi}_k]\|_{H^{\frac{\alpha}{2}}} = m.$$

By definition of the closure, there exists a function $\phi_k \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ for each k such that

$$\|\mathcal{F}^{-1}[\bar{\phi}_k] - \mathcal{F}^{-1}[\phi_k]\|_{H^{\frac{\alpha}{2}}} \leq \frac{1}{k}.$$

Therefore $\{\phi_k\}_{k=1}^{\infty} \subset \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$ is also a minimizing sequence, i.e.,

$$\lim_{k\to\infty} \|\mathcal{F}^{-1}[\phi_k]\|_{H^{\frac{\alpha}{2}}} = m.$$

Then $\{\mathcal{F}^{-1}[\phi_k]\}_{k=1}^{\infty}$ is bounded in the Sobolev space $H^{\frac{\alpha}{2}}(\mathbb{R}^d)$. Hence there exist a weakly convergent subsequence $\{\mathcal{F}^{-1}[\phi_{n_k}]\}_{k=1}^{\infty}$ and a function $\mathcal{F}^{-1}[\phi^*] \in H^{\frac{\alpha}{2}}(\mathbb{R}^d)$ such that

$$\mathcal{F}^{-1}[\phi_{n_k}] \rightharpoonup \mathcal{F}^{-1}[\phi^*] \quad \text{in } H^{\frac{\alpha}{2}}(\mathbb{R}^d) \text{ as } k \to \infty.$$

Note that

$$m = \inf_{\phi \in \overline{\mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}} \leq \|\mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} \leq \liminf_{\phi_{n_k}} \|\mathcal{F}^{-1}[\phi_{n_k}]\|_{H^{\frac{\alpha}{2}}} = m,$$

where we have used the lower semi-continuity of the Sobolev norm of $H^{\frac{\alpha}{2}}(\mathbb{R}^d)$ in the third inequality. Hence $\|\mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} = m$.

4. We further establish the strong convergence that $\mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*] \to 0$ in $H^{\frac{\alpha}{2}}(\mathbb{R}^d)$ as $k \to \infty$.

In fact, since $\mathcal{F}^{-1}[\phi_{n_k}] \rightharpoonup \mathcal{F}^{-1}[\phi^*]$ in $H^{\frac{\alpha}{2}}(\mathbb{R}^d)$ as $k \to \infty$ and $\lim_{k\to\infty} \|\mathcal{F}^{-1}[\phi_{n_k}]\|_{H^{\frac{\alpha}{2}}} = m = \|\mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}}$, we have

$$\lim_{k\to\infty} \|\mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}}^2 = \lim_{k\to\infty} \langle \mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*] \rangle$$

$$= \lim_{k\to\infty} \langle \mathcal{F}^{-1}[\phi_{n_k}], \mathcal{F}^{-1}[\phi_{n_k}] \rangle + \langle \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi^*] \rangle - \langle \mathcal{F}^{-1}[\phi_{n_k}], \mathcal{F}^{-1}[\phi^*] \rangle - \langle \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi_{n_k}] \rangle$$

$$= m^2 + m^2 - \lim_{k\to\infty} \left(\langle \mathcal{F}^{-1}[\phi_{n_k}], \mathcal{F}^{-1}[\phi^*] \rangle + \langle \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi_{n_k}] \rangle \right)$$

$$= m^2 + m^2 - \langle \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi^*] \rangle - \langle \mathcal{F}^{-1}[\phi^*], \mathcal{F}^{-1}[\phi^*] \rangle = 0.$$

Here $\langle \cdot, \cdot \rangle$ is the inner product of the Hilbert space $H^{\frac{\alpha}{2}}$.

5. We have $\phi^* \in L^1(\mathbb{R}^d)$ because

$$\int_{\mathbb{R}^d} |\phi^*(\boldsymbol{\xi})| \, \mathrm{d}\boldsymbol{\xi} = \int_{\mathbb{R}^d} \frac{\langle \boldsymbol{\xi} \rangle^{\frac{\alpha}{2}} |\phi^*(\boldsymbol{\xi})|}{\langle \boldsymbol{\xi} \rangle^{\frac{\alpha}{2}}} \, \mathrm{d}\boldsymbol{\xi} \leq \|\mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} \left(\int_{\mathbb{R}^d} \frac{1}{\langle \boldsymbol{\xi} \rangle^{\alpha}} \, \mathrm{d}\boldsymbol{\xi} \right)^{\frac{1}{2}} = Cm < +\infty,$$

where $C := \left(\int_{\mathbb{R}^d} \frac{1}{\langle \boldsymbol{\xi} \rangle^{\alpha}} \, \mathrm{d}\boldsymbol{\xi} \right)^{\frac{1}{2}} < +\infty$. Hence $\phi^* \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and $\mathcal{P}_{\boldsymbol{X}}\phi^*$ is well-defined.

6. Recall that $\mathcal{P}_{\boldsymbol{X}}\phi_{n_k} = \boldsymbol{Y}$. We have

$$|\boldsymbol{Y} - \mathcal{P}_{\boldsymbol{X}}\phi^*| = \lim_{k\to+\infty} |\mathcal{P}_{\boldsymbol{X}}\phi_{n_k} - \mathcal{P}_{\boldsymbol{X}}\phi^*|$$

$$= \lim_{k\to+\infty} \left| \int_{\mathbb{R}^d} (\phi_{n_k} - \phi^*) e^{2\pi i \boldsymbol{x}\boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{\xi} \right|$$

$$= \lim_{k\to+\infty} \left| \int_{\mathbb{R}^d} \frac{\langle \boldsymbol{\xi} \rangle^{\frac{\alpha}{2}} (\phi_{n_k} - \phi^*)}{\langle \boldsymbol{\xi} \rangle^{\frac{\alpha}{2}}} e^{2\pi i \boldsymbol{x}\boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{\xi} \right|$$

$$\leq \lim_{k\to+\infty} \|\mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} \left(\int_{\mathbb{R}^d} \frac{|e^{2\pi i \boldsymbol{x}\boldsymbol{\xi}}|^2}{\langle \boldsymbol{\xi} \rangle^{\alpha}} \, \mathrm{d}\boldsymbol{\xi} \right)^{\frac{1}{2}}$$

$$= C \lim_{k\to+\infty} \|\mathcal{F}^{-1}[\phi_{n_k}] - \mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} = 0.$$

Hence $\mathcal{P}_{\boldsymbol{X}}\phi^* = \boldsymbol{Y}$ and $\phi^* \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}$.

7. Note that

$$m = \inf_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}} \leq \inf_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}} \leq \|\mathcal{F}^{-1}[\phi^*]\|_{H^{\frac{\alpha}{2}}} = m.$$

This implies that $\inf_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}} = m$ and $\phi^* \in \arg\min_{\phi \in \mathcal{A}_{\boldsymbol{X},\boldsymbol{Y}}} \|\mathcal{F}^{-1}[\phi]\|_{H^{\frac{\alpha}{2}}}$, which completes the proof. ∎

Appendix E. Details of numerical experiments

E.1. Special Case: One Data Point in One Dimension

To simplify the problem, we start with a single point $X = 0 \in \mathbb{Z}$ with the label $Y = 2$. Denote $\phi_j = \phi(\xi_j)$ for $j \in \mathbb{Z}$. We also assume that the function ϕ is an even function. Then according to the definition of $\mathcal{P}_{\boldsymbol{X}}$, we have the following problem:

Example 3 (Problem 1 with a particular discretization)

$$\min_{\phi \in \mathbb{R}^M} \sum_{j=1}^{M} (1 + j^2 \Delta \xi^2)^{\frac{\alpha}{2}} \left| \phi_j \right|^2, \tag{18}$$

$$\text{s.t.} \quad \sum_{j=1}^{M} \phi_j \Delta \xi = 1, \tag{19}$$

where we further assume $\phi_0 = \phi(0) = 0$. If we denote $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_M)^{\mathsf{T}}$, $b = \frac{1}{\Delta \xi}$, $\boldsymbol{A} = (1, 1, \ldots, 1) \in \mathbb{R}^M$ and

$$\boldsymbol{\Gamma} = \sqrt{\lambda} \begin{pmatrix} (1 + 1^2 \Delta \xi^2)^{\frac{\alpha}{4}} & & & \\ & (1 + 2^2 \Delta \xi^2)^{\frac{\alpha}{4}} & & \\ & & \ddots & \\ & & & (1 + M^2 \Delta \xi^2)^{\frac{\alpha}{4}} \end{pmatrix}.$$

In fact this is a standard Tikhonov regularization (Tikhonov and Arsenin, 1977) also known as ridge regression problem with the Lagrange multiplier λ. The corresponding ridge regression problem is,

$$\min_{\phi} \|\boldsymbol{A}\phi - b\|_2^2 + \|\boldsymbol{\Gamma}\phi\|_2^2, \tag{20}$$

where we put λ in the optimization term $\|\boldsymbol{\Gamma}\phi\|_2^2$, instead of the constraint term $\|\boldsymbol{A}\phi - b\|_2^2$. This problem admits an explicit and unique solution (Tikhonov and Arsenin, 1977),

$$\phi = (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} + \boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{A}^{\mathsf{T}}b. \tag{21}$$

Here we need to point out that the above method is also applicable to the case that the matrix $\boldsymbol{\Gamma}$ is not diagonal.

Back to our problem, in order to obtain the explicit expression for the optimal ϕ we need the following relation between the solution of the ridge regression and the singular-value decomposition (SVD).

By denoting $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{I}$ and

$$\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Gamma}^{-1} = \frac{1}{\sqrt{\lambda}} \left((1 + 1^2 \Delta \xi^2)^{\frac{\alpha}{4}}, (1 + 2^2 \Delta \xi^2)^{\frac{\alpha}{4}}, \ldots, (1 + M^2 \Delta \xi^2)^{\frac{\alpha}{4}} \right),$$

where $\boldsymbol{I}$ is the diagonal matrix, the optimal solution (21) can be written as

$$\phi = (\boldsymbol{\Gamma}^{\mathsf{T}})^{-1} \left(\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}} + \boldsymbol{I} \right)^{-1} \boldsymbol{\Gamma}^{-1}\boldsymbol{A}^{\mathsf{T}}b = (\boldsymbol{\Gamma}^{\mathsf{T}})^{-1} \left(\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}} + \boldsymbol{I} \right)^{-1} \tilde{\boldsymbol{A}}^{\mathsf{T}}b = (\boldsymbol{\Gamma}^{\mathsf{T}})^{-1}\tilde{\phi},$$

where $\tilde{\phi} = \left(\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}} + \boldsymbol{I} \right)^{-1} \tilde{\boldsymbol{A}}^{\mathsf{T}}b$ is the solution of ridge regression with $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{\Gamma}}$. In order to obtain the explicit expression for $\tilde{\phi}$ we need the following relation between the solution of the ridge regression and the singular-value decomposition (SVD).

Lemma 2 *If $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{I}$, then this least-squares solution can be solved using SVD. Given the singular value decomposition*

$$\tilde{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}},$$

with singular values σ_i, the Tikhonov regularized solution can be expressed aspects

$$\tilde{\phi} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}b,$$

where $\boldsymbol{D}$ has diagonal values

$$D_{ii} = \frac{\sigma_i}{\sigma_i^2 + 1},$$

and is zero elsewhere.

Proof In fact, $\tilde{\phi} = (\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}} + \tilde{\boldsymbol{\Gamma}}^{\mathsf{T}}\tilde{\boldsymbol{\Gamma}})^{-1}\tilde{\boldsymbol{A}}^{\mathsf{T}}b = \boldsymbol{V}(\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + 1\boldsymbol{I})^{-1}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}b$
$= \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}b$, which completes the proof. ∎

Since $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\mathsf{T}} = \frac{1}{\lambda}\sum_{j=1}^{M}(1 + j^2\Delta\xi^2)^{-\frac{\alpha}{2}}$, we have $\tilde{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}$ with

$$U = 1, \quad \Sigma = \frac{1}{\sqrt{\lambda}}\left(\sum_{j=1}^{M}(1 + j^2\Delta\xi^2)^{-\frac{\alpha}{2}}\right)^{\frac{1}{2}} := Z/\sqrt{\lambda},$$

$$\boldsymbol{V} = \left((1 + 1^2\Delta\xi^2)^{-\frac{\alpha}{2}}/Z, (1 + 2^2\Delta\xi^2)^{-\frac{\alpha}{2}}/Z, \ldots, (1 + M^2\Delta\xi^2)^{-\frac{\alpha}{2}}/Z\right)^{\mathsf{T}}.$$

Then we get the diagonal value

$$D = \frac{Z/\sqrt{\lambda}}{Z^2/\lambda + 1}.$$

Therefore, by Lemma 2

$$\tilde{\phi} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}b = \frac{1/\sqrt{\lambda}}{Z^2/\lambda + 1}\left((1 + 1^2\Delta\xi^2)^{-\frac{\alpha}{2}}, (1 + 2^2\Delta\xi^2)^{-\frac{\alpha}{2}}, \ldots, (1 + M^2\Delta\xi^2)^{-\frac{\alpha}{2}}\right)^{\mathsf{T}}b.$$

Finally, for the original optimal solution

$$\phi = (\boldsymbol{\Gamma}^{\mathsf{T}})^{-1}\tilde{\phi} = \frac{1}{(Z^2 + \lambda)\Delta\xi}\left((1 + 1^2\Delta\xi^2)^{-\frac{\alpha}{2}}, (1 + 2^2\Delta\xi^2)^{-\frac{\alpha}{2}}, \ldots, (1 + M^2\Delta\xi^2)^{-\frac{\alpha}{2}}\right)^{\mathsf{T}},$$

which means

$$\phi_j = \frac{(1 + j^2\Delta\xi^2)^{-\frac{\alpha}{2}}}{(Z^2 + \lambda)\Delta\xi}.$$

To derive the function in x space, say $h(x)$, then

$$h(x) = \frac{1}{(Z^2 + \lambda)}\sum_{j=-M}^{M}(1 + j^2\Delta\xi^2)^{-\frac{\alpha}{2}}e^{2\pi ijx}$$

$$= \frac{2}{(Z^2 + \lambda)}\sum_{j=1}^{M}(1 + j^2\Delta\xi^2)^{-\frac{\alpha}{2}}\cos(2\pi jx). \tag{22}$$

Fig. 1 shows that for this special case with a large M, $h(x)$ is not an trivial function in $\alpha > d$ case and degenerates to a trivial function in $\alpha < d$ case.

E.2. General Case: n Points in d Dimension

Assume that we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and each data point has d components:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^\mathsf{T}$$

and denote the corresponding label as $(y_1, y_2, \ldots, y_n)^\mathsf{T}$. For the sake of simplicity, we denote the vector $(j_1, j_2, \cdots, j_d)^\mathsf{T}$ by $\boldsymbol{J}_{j_1 \ldots j_d}$. Then our problem becomes

Example 4 (Problem 1 with general discretization)

$$\min_{\phi \in \mathbb{R}^{(2M)^d}} \sum_{j_1, \ldots, j_d = -M}^{M} (1 + \|\boldsymbol{J}_{j_1 \ldots j_d}\|^2 \Delta \xi^2)^{\frac{\alpha}{2}} |\phi_{j_1 \ldots j_d}|^2, \tag{23}$$

$$\text{s.t.} \quad \sum_{j_1, \ldots, j_d = -M}^{M} \phi_{j_1 \ldots j_d} e^{2\pi \mathrm{i} \Delta \xi \boldsymbol{J}_{j_1 \ldots j_d}^\mathsf{T} \boldsymbol{x}_k} = y_k, \quad k = 1, 2, \ldots, d \tag{24}$$

The calculation of this example can be completed by the method analogous to the one used in subsection E.1. Let

$$\boldsymbol{A}_j = \left(e^{2\pi \mathrm{i} \Delta \xi \boldsymbol{J}_{-M-M\ldots-M}^\mathsf{T} \boldsymbol{x}_j}, \ldots, e^{2\pi \mathrm{i} \Delta \xi \boldsymbol{J}_{j_1 j_2 \ldots j_d}^\mathsf{T} \boldsymbol{x}_j}, \ldots, e^{2\pi \mathrm{i} \Delta \xi \boldsymbol{J}_{MM\ldots M}^\mathsf{T} \boldsymbol{x}_j} \right)^\mathsf{T}, \quad j = 1, 2, \ldots, n, \tag{25}$$

$$\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)^\mathsf{T} \in \mathbb{R}^{n \times (2M)^d}, \quad \boldsymbol{b} = (y_1, y_2, \ldots, y_n)^\mathsf{T} \in \mathbb{R}^{n \times 1}, \tag{26}$$

$$\boldsymbol{\Gamma} = \lambda \begin{pmatrix} \ddots & & \\ & (1 + \|\boldsymbol{J}_{j_1 j_2 \ldots j_d}\|^2 \Delta \xi^2)^{\frac{\alpha}{4}} & \\ & & \ddots \end{pmatrix} \in \mathbb{R}^{(2M)^d \times (2M)^d}. \tag{27}$$

We just need to solve the following equation:

$$\boldsymbol{\phi} = (\boldsymbol{A}^\mathsf{T} \boldsymbol{A} + \boldsymbol{\Gamma}^\mathsf{T} \boldsymbol{\Gamma})^{-1} \boldsymbol{A}^\mathsf{T} b. \tag{28}$$

Then we can obtain the output function $h(x)$ by using inverse Fourier transform:

$$h(\boldsymbol{x}) = \sum_{j_1, \ldots, j_d = -M}^{M} \phi_{j_1 \ldots j_d} e^{2\pi \mathrm{i} \Delta \xi \boldsymbol{J}_{j_1 \ldots j_d} \cdot \boldsymbol{x}} \tag{29}$$

Since the size of the matrix is too large, it is difficult to solve ϕ by an explicit calculation. Thus we choose special n, d and M and show that $h(x)$ is not a trivial solution (non-zero function).

In our experiment, we set the hyper-parameter $M, \alpha, \lambda, \Delta \xi$ in advance. We set $\lambda = 0, 5, \Delta \xi = 0.1$ in 1-dimensional case and $\lambda = 0.2, \Delta \xi = 0.1$ in 2-dimensional case. We select two data points $\{(-0.5, 0.9), (0.5, 0.9)\}$ as the given points in 1-dimensional case and four points as given points in 2-dimensional case whose second coordinates are 0.5 so that it is convenient to observe the phenomenon. At first, we use formula (25), (26) and (27) to calculate matrix $\boldsymbol{A}, \boldsymbol{\Gamma}$ and vector $\boldsymbol{b}$. Then from the equation (28) we can deduce vector ϕ. The final output function $h(\boldsymbol{x})$ is obtained by inverse discrete Fourier transform (29).

In Fig.2, we set $\alpha = 10$ in both cases to ensure $\alpha > d$ and change the band limit M. We observe that as M increases, the fitting curve converges to a non-trivial curve. In Fig.3, we set $M = 1000$ in 1-dimensional case and $M = 100$ in 2-dimensional case. By changing exponent α, we can see in all cases, the fitting curves are non-trivial when $\alpha > d$, but degenerate when $\alpha < d$.