On the Generalization Power of Overfitted Two-Layer Neural Tangent Kernel Models

Peizhong Ju [1] **Xiaojun Lin** [1] **Ness B. Shroff** [2]

Abstract

In this paper, we study the generalization performance of min ℓ_2-norm overfitting solutions for the neural tangent kernel (NTK) model of a two-layer neural network with ReLU activation that has no bias term. We show that, depending on the ground-truth function, the test error of overfitted NTK models exhibits characteristics that are different from the "double-descent" of other overparameterized linear models with simple Fourier or Gaussian features. Specifically, for a class of learnable functions, we provide a new upper bound of the generalization error that approaches a small limiting value, even when the number of neurons p approaches infinity. This limiting value further decreases with the number of training samples n. For functions outside of this class, we provide a lower bound on the generalization error that does not diminish to zero even when n and p are both large.

1. Introduction

Recently, there is significant interest in understanding why overparameterized deep neural networks (DNNs) can still generalize well (Zhang et al., 2017; Advani et al., 2020), which seems to defy the classical understanding of *bias-variance tradeoff* in statistical learning (Bishop, 2006; Hastie et al., 2009; Stein, 1956; James & Stein, 1992; LeCun et al., 1991; Tikhonov, 1943). Towards this direction, a recent line of study has focused on overparameterized linear models (Belkin et al., 2018b; 2019; Bartlett et al., 2020; Hastie et al., 2019; Muthukumar et al., 2019; Ju et al., 2020; Mei & Montanari, 2019). For linear models with simple features (e.g., Gaussian features and Fourier features) (Belkin et al., 2018b; 2019; Bartlett et al., 2020; Hastie et al., 2019; Muthukumar et al., 2019; Ju et al., 2020), an interesting "double-descent" phenomenon has been observed. Thus, there is a region where the number of model parameters (or linear features) is larger than the number of samples (and thus overfitting occurs), but the generalization error actually decreases with the number of features. However, linear models with these simple features are still quite different from nonlinear neural networks. Thus, although such results provide some hint why overparameterization and overfitting may be harmless, it is still unclear whether similar conclusions apply to neural networks.

In this paper, we are interested in linear models based on the neural tangent kernel (NTK) (Jacot et al., 2018), which can be viewed as a useful intermediate step towards modeling nonlinear neural networks. Essentially, NTK can be seen as a linear approximation of neural networks when the weights of the neurons do not change much. Indeed, (Li & Liang, 2018; Du et al., 2018) have shown that, for a wide and fully-connected two-layer neural network, both the neuron weights and their activation patterns do not change much after gradient descent (GD) training with a sufficiently small step size. As a result, such a shallow and wide neural network is approximately linear in the weights when there are a sufficient number of neurons, which suggests the utility of the NTK model.

Despite its linearity, however, characterizing the double descent of such a NTK model remains elusive. The work in (Mei & Montanari, 2019) also studies the double-descent of a linear version of two-layer neural network. It uses the so-called "random-feature" model, where the bottom-layer weights are random and fixed, and only the top-layer weights are trained. (In comparison, the NTK model for such a two-layer neural network corresponds to training only the bottom-layer weights.) However, the setting there requires the number of neurons, the number of samples, and the data dimension to all grow proportionally to infinity. In contrast, we are interested in the setting where the number of samples is given, and the number of neurons is allowed to be much larger than the number of samples. As a consequence of the different setting, in (Mei & Montanari, 2019) eventually only *linear* ground-truth functions can be learned. (Similar settings are also studied in (d'Ascoli et al.,

[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA. [2]Department of ECE and CSE, The Ohio State University, Columbus, Ohio, USA. Correspondence to: Xiaojun Lin <linx@purdue.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

2020).) In contrast, we will show that far more complex functions can be learned in our setting. In a related work, (Ghorbani et al., 2019) shows that both the random-feature model and the NTK model can approximate highly *non-linear* ground-truth functions with a sufficient number of neurons. However, (Ghorbani et al., 2019) mainly studies the *expressiveness* of the models, and therefore does not explain why overfitting solutions can still generalize well. To the best of our knowledge, our work is the first to characterize the double-descent of overfitting solutions based on the NTK model.

Specifically, in this paper we study the generalization error of the min ℓ_2-norm overfitting solution for a linear model based on the NTK of a two-layer neural network with ReLU activation that has no bias. Only the bottom-layer weights are trained. We are interested in min ℓ_2-norm overfitting solutions because gradient descent (GD) can be shown to converge to such solutions while driving the training error to zero (Zhang et al., 2017) (see also Section 2). Given a class of ground truth functions (see details in Section 3), which we refer to as "learnable functions," our main result (Theorem 1) provides an upper bound on the generalization error of the min ℓ_2-norm overfitting solution for the two-layer NTK model with n samples and p neurons (for any finite p larger than a polynomial function of n). This upper bound confirms that the generalization error of the overfitting solution indeed exhibits descent in the overparameterized regime when p increases. Further, our upper bound can also account for the noise in the training samples.

Our results reveal several important insights. First, we find that the (double) descent of the overfitted two-layer NTK model is drastically different from that of linear models with simple Gaussian or Fourier features (Belkin et al., 2018b; 2019; Bartlett et al., 2020; Hastie et al., 2019; Muthukumar et al., 2019). Specifically, for linear models with simple features, when the number of features p increases, the generalization error will eventually grow again and approach the so-called "null risk" (Hastie et al., 2019), which is the error of a trivial model that predicts zero. In contrast, for the class of learnable functions described earlier, the generalization error of the overfitted NTK model will continue to descend as p grows to infinity, and will approach a limiting value that depends on the number of samples n. Further, when there is no noise, this limiting value will decrease to zero as the number of samples n increases. This difference is shown in Fig. 1(a). As p increases, the test mean-square-error (MSE) of min-ℓ_1 and min-ℓ_2 overfitting solutions for Fourier features (blue and red curves) eventually grow back to the null risk (the black dashed line), even though they exhibit a descent at smaller p. In contrast, the error of the overfitted NTK model continues to descend to a much lower level.

The second important insight is that the aforementioned behavior critically depends on the ground-truth function belonging to the class of "learnable functions." Further, this class of learnable functions depend on the specific network architecture. For our NTK model (with RELU activation that has no bias), we precisely characterize this class of learnable functions. Specifically, for ground-truth functions that are outside the class of learnable functions, we show a lower bound on the generalization error that does not diminish to zero for any n and p (see Proposition 2 and Section 4). This difference is shown in Fig. 1(b), where we use an almost identical setting as Fig. 1(a), except a different ground-truth function. We can see in Fig. 1(b) that the test-error of the overfitted NTK model is always above the null risk and looks very different from that in Fig. 1(a). We note that whether certain functions are learnable or not critically depends on the specific structure of the NTK model, such as the choice of the activation unit. Recently, (Satpathi & Srikant, 2021) shows that all polynomials can be learned by 2-layer NTK model with ReLU activation that has a bias term, provided that the number of neurons p is sufficiently large. (See further discussions in Remark 2. However, (Satpathi & Srikant, 2021) does not characterize the descent of generalization errors as p increases.) This difference in the class of learnable functions between the two settings (ReLU with or without bias) also turns out to be consistent with the difference in the expressiveness of the neural networks. That is, shallow networks with biased-ReLU are known to be universal function approximators (Ji et al., 2019), while those without bias can only approximate the sum of linear functions and even functions (Ghorbani et al., 2019).

A closely related result to ours is the work in (Arora et al., 2019), which characterizes the generalization performance of wide two-layer neural networks whose bottom-layer weights are trained by gradient descent (GD) to overfit the training samples. In particular, our class of learnable functions almost coincides with that of (Arora et al., 2019). This is not surprising because, when the number of neurons is large, NTK becomes a close approximation of such two-layer neural networks. In that sense, the results in (Arora et al., 2019) are even more faithful in following the GD dynamics of the original two-layer network. However, the advantage of the NTK model is that it is easier to analyze. In particular, the results in this paper can quantify how the generalization error descends with p. In contrast, the results in (Arora et al., 2019) provide only a generalization bound that is independent of p (provided that p is sufficiently large), but do not quantify the descent behavior as p increases. Our numerical results in Fig. 1(a) suggest that, over a wide range of p, the descent behavior of the NTK model (the green curve) matches well with that of two-layer neural networks trained by gradient descent (the cyan curve). Thus, we believe that our results also provide guidance for the latter

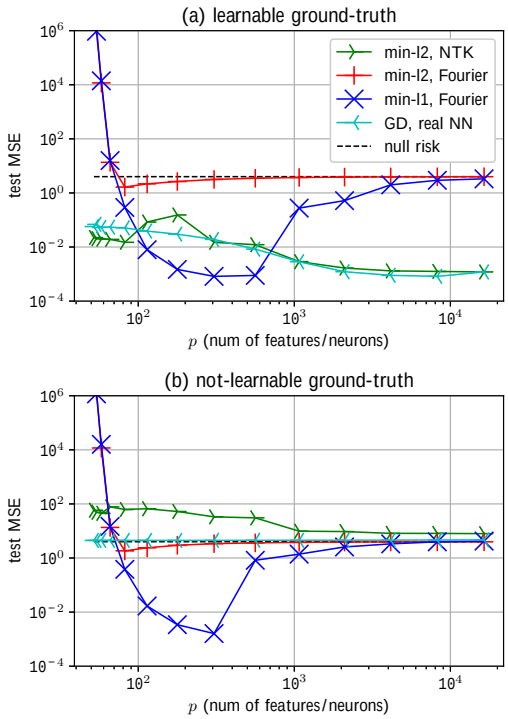

Figure 1. The test mean-square-error(MSE) vs. the number of features/neurons p for **(a)** learnable function and **(b)** not-learnable function when $n = 50$, $d = 2$, $k\epsilon k_2^2 = 0.01$. The corresponding ground-truth are **(a)** $f(\theta) = \sum_{k \in \{0,1,2,4\}} (\sin(k\theta) + \cos(k\theta))$, and **(b)** $f(\theta) = \sum_{k \in \{3,5,7,9\}} (\sin(k\theta) + \cos(k\theta))$. (Note that in 2-dimension every input $\boldsymbol{x}$ on a unit circle can be represented by an angle $\theta \in [-\pi, \pi]$. See the end of Section 4.) Every curve is the average of 9 random simulation runs. For GD on the real neural network (NN), we use the step size $1/\sqrt{p}$ and the number of training epochs is fixed at 2000.

model. The work in (Fiat et al., 2019) studies a different neural network architecture with gated ReLU, whose NTK model turns out to be the same as ours. However, similar to (Arora et al., 2019), the result in (Fiat et al., 2019) does not capture the speed of descent with respect to p either. Second, (Arora et al., 2019) only provides upper bounds on the generalization error. There is no corresponding lower bound to explain whether ground-truth functions outside a certain class are *not* learnable. Our result in Proposition 2 provides such a lower bound, and therefore more completely characterizes the class of learnable functions. (See further comparison in Remark 1 of Section 3 and Remark 3 of Section 5.) Another related work (Allen-Zhu et al., 2019) also characterizes the class of learnable functions for two-layer and three-layer networks. However, (Allen-Zhu et al., 2019) studies a training method that takes a new sample in every iteration, and thus does not overfit all training data. Finally, our paper studies generalization of NTK models for the regression setting, which is different from the classification

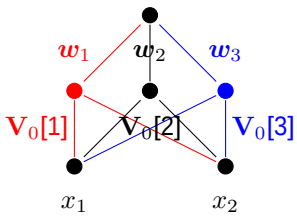

Figure 2. A two-layer neural network where $d = 2$, $p = 3$.

setting that assumes a separability condition, e.g., in (Ji & Telgarsky, 2019).

2. Problem Setup

We assume the following data model $y = f(\boldsymbol{x}) + \epsilon$, with the input $\boldsymbol{x} \in \mathbb{R}^d$, the output $y \in \mathbb{R}$, the noise $\epsilon \in \mathbb{R}$, and $f : \mathbb{R}^d \mapsto \mathbb{R}$ denotes the ground-truth function. Let $(\mathbf{X}_i, y_i)$, $i = 1, 2, \ldots, n$ denote n training samples. We collect them as $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_n] \in \mathbb{R}^{d \times n}$, $\boldsymbol{y} = [y_1\ y_2\ \ldots\ y_n]^T \in \mathbb{R}^n$, $\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2\ \ldots\ \epsilon_n]^T \in \mathbb{R}^n$, and $\mathbf{F}(\mathbf{X}) = [f(\mathbf{X}_1)\ f(\mathbf{X}_2)\ \ldots\ f(\mathbf{X}_n)]^T \in \mathbb{R}^n$. Then, the training samples can be written as $\boldsymbol{y} = \mathbf{F}(\mathbf{X}) + \boldsymbol{\epsilon}$. After training (to be described below), we denote the trained model by the function $\hat{f}$. Then, for any new test data $\boldsymbol{x}$, we will calculate the test error by $|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|$, and the mean squared error (MSE) by $\mathbb{E}_{\boldsymbol{x}}[\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})]^2$.

For training, consider a fully-connected two-layer neural network with p neurons. Let $\boldsymbol{w}_j \in \mathbb{R}$ and $\mathbf{V}_0[j] \in \mathbb{R}^d$ denote the top-layer and bottom-layer weights, respectively, of the j-th neuron, $j = 1, 2, \ldots, p$ (see Fig. 2). We collect them into $\boldsymbol{w} = [\boldsymbol{w}_1\ \boldsymbol{w}_2\ \ldots\ \boldsymbol{w}_p]^T \in \mathbb{R}^p$, and $\mathbf{V}_0 = [\mathbf{V}_0[1]^T\ \mathbf{V}_0[2]^T\ \ldots\ \mathbf{V}_0[p]^T]^T \in \mathbb{R}^{dp}$ (a column vector with dp elements). Note that with this notation, for any row or column vector $\boldsymbol{v}$ with dp elements, $\boldsymbol{v}[j]$ denotes a (row/column) vector that consists of the $(jd+1)$-th to $(jd+d)$-th elements of $\boldsymbol{v}$. We choose ReLU as the activation function for all neurons and there is no bias term in the ReLU activation function.

Now we are ready to introduce the NTK model (Jacot et al., 2018). We fix the top-layer weights $\boldsymbol{w}$, and let the initial bottom-layer weights $\mathbf{V}_0$ be randomly chosen. We then train only the bottom-layer weights. Let $\mathbf{V}_0 + \overline{\mathbf{V}}$ denote the bottom-layer weights after training. Thus, the change of the output after training is

$$\sum_{j=1}^{n} \boldsymbol{w}_j \mathbf{1}_{\{\boldsymbol{x}^T(\mathbf{V}_0[j] + \overline{\Delta\mathbf{V}}[j]) > 0\}} (\mathbf{V}_0[j] + \overline{\mathbf{V}}[j])^T \boldsymbol{x} - \sum_{j=1}^{n} \boldsymbol{w}_j \mathbf{1}_{\{\boldsymbol{x}^T \mathbf{V}_0[j] > 0\}} \mathbf{V}_0[j]^T \boldsymbol{x}.$$

In the NTK model, one assumes that $\overline{\mathbf{V}}$ is very small. As

a result, $\mathbf{1}_{f\boldsymbol{x}^T(\mathbf{V}_0[j]+\overline{\Delta\mathbf{V}}[j])>0g} = \mathbf{1}_{f\boldsymbol{x}^T\mathbf{V}_0[j]>0g}$ for most $\boldsymbol{x}$. Thus, the change of the output can be approximated by

$$\sum_{j=1}^{n} \boldsymbol{w}_j \mathbf{1}_{f\boldsymbol{x}^T\mathbf{V}_0[j]>0g} \quad \overline{\mathbf{V}}[j]^T\boldsymbol{x} = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V},$$

where $\mathbf{V} \; 2 \; \mathsf{R}^{dp}$ is given by $\mathbf{V}[j] := \boldsymbol{w}_j \overline{\mathbf{V}}[j]$, $j = 1, 2, \quad , p$, and $\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \; 2 \; \mathsf{R}^{1 \quad (dp)}$ is given by

$$\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}[j] := \mathbf{1}_{f\boldsymbol{x}^T\mathbf{V}_0[j]>0g} \quad \boldsymbol{x}^T, \; j = 1, 2, \quad , p. \tag{1}$$

In the NTK model, we assume that the output of the trained model is exactly given by Eq. (1), i.e.,

$$\hat{f}_{\Delta\mathbf{V},\mathbf{V}_0}(\boldsymbol{x}) := \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V}. \tag{2}$$

In other words, the NTK model can be viewed as a linear approximation of the two-layer network when the change of the bottom-layer weights is small.

Define $\mathbf{H} \; 2 \; \mathsf{R}^{n \quad (dp)}$ such that its i-th row is $\mathbf{H}_i := \boldsymbol{h}_{\mathbf{V}_0,\mathbf{X}_i}$. Throughout the paper, we will focus on the following min-ℓ_2-norm overfitting solution

$$\mathbf{V}^{\ell_2} := \arg\min_{\boldsymbol{v}} k\boldsymbol{v}k_2, \text{ subject to } \mathbf{H}\boldsymbol{v} = \boldsymbol{y}.$$

Whenever $\mathbf{V}^{\ell_2}$ exists, it can be written in closed form as

$$\mathbf{V}^{\ell_2} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\boldsymbol{y}. \tag{3}$$

The reason that we are interested in $\mathbf{V}^{\ell_2}$ is that gradient descent (GD) or stochastic gradient descent (SGD) for the NTK model in Eq. (2) is known to converge to $\mathbf{V}^{\ell_2}$ (proven in Supplementary Material, Appendix B).

Using Eq. (2) and Eq. (3), the trained model is then

$$\hat{f}^{\ell_2}(\boldsymbol{x}) := \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V}^{\ell_2}. \tag{4}$$

In the rest of the paper, we will study the generalization error of Eq. (4).

We collect some assumptions. Define the unit sphere in R^d as: $S^{d \quad 1} := \{\boldsymbol{v} \; 2 \; \mathsf{R}^d \; j \; k\boldsymbol{v}k_2 = 1\}$. Let $\mu(\,)$ denote the distribution of the input $\boldsymbol{x}$. Without loss of generality, we make the following assumptions: **(i)** the inputs $\boldsymbol{x}$ are *i.i.d.* uniformly distributed in $S^{d \quad 1}$, and the initial weights $\mathbf{V}_0[j]$'s are *i.i.d.* uniformly distributed in all directions in R^d; **(ii)** $p \quad n/d$ and $d \quad 2$; **(iii)** $\mathbf{X}_i \;\; , \;\; \mathbf{X}_j$ for any $i \; \not\in \; j$, and $\mathbf{V}_0[k] \;\; , \;\; \mathbf{V}_0[l]$ for any $k \; \not\in \; l$. We provide detailed justification of those assumptions in Supplementary Material, Appendix C.

3. Learnable Functions and Generalization Performance

We now show that the generalization performance of the overfitted NTK model in Eq. (4) crucially depends on the ground-truth function $f(\,)$, where good generalization performance only occurs when the ground-truth function is "learnable." Below, we first describe a candidate class of ground-truth functions, and explain why they may correspond to the class of "learnable functions." Then, we will give an upper-bound on the generalization performance for this class of ground-truth functions. Finally, we will give a lower-bound on the generalization performance when the ground-truth functions are outside of this class.

We first define a set F^{ℓ_2} of ground-truth functions.

Definition 1. $F^{\ell_2} := \{f \overset{a.e.}{=} f_g \mid f_g(\boldsymbol{x}) = \int_{S^{d \quad 1}} \boldsymbol{x}^T\boldsymbol{z} \frac{\pi \quad \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} g(\boldsymbol{z}) d\mu(\boldsymbol{z}), \; kgk_1 < 1\}$.

Note that in Definition 1, $\overset{a.e.}{=}$ means two functions equals almost everywhere, and $kgk_1 := \int_{S^{d \quad 1}} jg(\boldsymbol{z})jd\mu(\boldsymbol{z})$. The function $g(\boldsymbol{z})$ may be any finite-value function in $L^1(S^{d \quad 1} \; ?! \; \mathsf{R})$. Further, we also allow $g(\boldsymbol{z})$ to contain (as components) Dirac δ-functions on $S^{d \quad 1}$. Note that a δ-function $\delta_{\boldsymbol{z}_0}(\boldsymbol{z})$ has zero value for all $\boldsymbol{z} \; 2 \; S^{d \quad 1} \; n \; f\boldsymbol{z}_0g$, but $k\delta_{\boldsymbol{z}_0}k_1 := \int_{S^{d \quad 1}} \delta_{\boldsymbol{z}_0}(\boldsymbol{z}) d\mu(\boldsymbol{z}) = 1$. Thus, the function $g(\boldsymbol{z})$ may contain any sum of δ-functions and finite-value L^1-functions. [1]

To see why F^{ℓ_2} may correspond to the class of learnable functions, we can first examine what the learned function $\hat{f}^{\ell_2}$ in Eq. (4) should look like. Recall that $\mathbf{H}^T = [\mathbf{H}_1^T \quad \mathbf{H}_n^T]$. Thus, $\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T = \sum_{i=1}^n (\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}_i^T)\boldsymbol{e}_i^T$, where $\boldsymbol{e}_i \; 2 \; \mathsf{R}^n$ denotes the i-th standard basis. Combining Eq. (3) and Eq. (4), we can see that the learned function in Eq. (4) is of the form

$$\begin{aligned}\hat{f}^{\ell_2}(\boldsymbol{x}) &= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\boldsymbol{y} \\ &= \sum_{i=1}^{n} \left(\frac{1}{p}\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}_i^T\right) p\boldsymbol{e}_i^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\boldsymbol{y}. \end{aligned} \tag{5}$$

For all $\boldsymbol{x}, \boldsymbol{z} \quad 2 \quad S^{d \quad 1}$, define $C_{\boldsymbol{z},\boldsymbol{x}}^{\mathbf{V}_0} := fj \quad 2 \; f1, 2, \quad , pg \; j \; \boldsymbol{z}^T\mathbf{V}_0[j] > 0, \boldsymbol{x}^T\mathbf{V}_0[j] > 0g$, and its cardinality is given by

$$\left|C_{\boldsymbol{z},\boldsymbol{x}}^{\mathbf{V}_0}\right| = \sum_{j=1}^{p} \mathbf{1}_{f\boldsymbol{z}^T\mathbf{V}_0[j]>0, \; \boldsymbol{x}^T\mathbf{V}_0[j]>0g}. \tag{6}$$

Then, using Eq. (1), we can show $\frac{1}{p}\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}_i^T = \boldsymbol{x}^T\mathbf{X}_i\frac{jC_{\mathbf{X}_i,\boldsymbol{x}}^{\mathbf{V}_0}j}{p}$. It is not hard to show that

$$\frac{jC_{\boldsymbol{z},\boldsymbol{x}}^{\mathbf{V}_0}j}{p} \; \overset{\mathrm{P}}{?} \; \frac{\pi \quad \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}, \text{ as } p \; ! \; 1. \tag{7}$$

[1] Alternatively, we can also interpret $g(\boldsymbol{z})$ as a signed measure (Rao & Rao, 1983) on $S^{d \quad 1}$. Then, δ-functions correspond to point masses, and the condition $kgk_1 < 1$ implies that the corresponding unsigned version of the measure on $S^{d \quad 1}$ is bounded.

where $\xrightarrow{p}$ denotes converge in probability. (see Supplementary Material, Appendix D.5). Thus, if we let

$$g(\boldsymbol{z}) = \sum_{i=1}^{n} p\boldsymbol{e}_i^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{y}\delta_{\mathbf{X}_i}(\boldsymbol{z}), \tag{8}$$

then as $p \to \infty$, Eq. (5) should approach a function in F^{ℓ_2}. This explains why F^{ℓ_2} is a candidate class of "learnable functions." However, note that the above discussion only addresses the *expressiveness* of the model. It is still unclear whether any function in F^{ℓ_2} can be learned with low generalization error. The following result provides the answer.

For some $m \in \left[1, \frac{\ln n}{\ln \frac{\pi}{2}}\right]$, define (recall that d is the dimension of $\boldsymbol{x}$)

$$J_m(n,d) := 2^{1.5d+5.5} d^{0.5d} n^{\left(2+\frac{1}{m}\right)(d-1)}. \tag{9}$$

Theorem 1. *Assume a ground-truth function $f \overset{a.e.}{=} f_g \in F^{\ell_2}$ where $\|g\|_1 < 1$[2], $n \geq 2$, $m \in \left[1, \frac{\ln n}{\ln \frac{\pi}{2}}\right]$, $d \geq n^4$, and $p \geq 6J_m(n,d)\ln\left(4n^{1+\frac{1}{m}}\right)$. Then, for any $q \in [1, 1)$ and for almost every $\boldsymbol{x} \in S^{d-1}$, we must have*[3]

$$\begin{aligned}
&\Pr_{\mathbf{V}_0,\mathbf{X}}\Big\{|\hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \underbrace{n^{\frac{1}{2}\left(1-\frac{1}{q}\right)}}_{\text{Term 1}} \\
&+ \underbrace{\left(1+\sqrt{J_m(n,d)n}\right)p^{\frac{1}{2}\left(1-\frac{1}{q}\right)}}_{\text{Term 2}} + \underbrace{\sqrt{J_m(n,d)n}\epsilon\|k\|_2}_{\text{Term 3}}, \\
&\text{for all } \epsilon \in \mathbb{R}^n\Big\} \leq 2e^2\Bigg(\underbrace{\exp\left(-\frac{p^{\frac{1}{q}}n}{8\|k\|_1^2}\right)}_{\text{Term 4}} \\
&+ \underbrace{\exp\left(-\frac{p^{\frac{1}{q}}p}{8\|k\|_1^2}\right)}_{\text{Term 5}} + \underbrace{\exp\left(-\frac{p^{\frac{1}{q}}p}{8n\|k\|_1^2}\right)}_{\text{Term 6}}\Bigg) + \underbrace{\frac{4}{m}\frac{n}{p}}_{\text{Term 7}}.
\end{aligned} \tag{10}$$

To interpret Theorem 1, we can first focus on the noiseless case, where ϵ and Term 3 are zero. If we fix n and let $p \to \infty$, then Terms 2, 5, and 6 all approach zero. We can then conclude that, in the noiseless and heavily overparameterized setting ($p \to \infty$), the generalization error will converge to a small limiting value (Term 1) that depends only on n. Further, this limiting value (Term 1) will converge to zero (so do Terms 4 and 7) as $n \to \infty$, i.e., when there are sufficiently many training samples. Finally, Theorem 1 holds even when there is noise.

The parameters of q and m can be tuned to make Eq. (10) sharper when n and p are large. For example, as we increase q, Term 1 will approach $n^{-0.5}$. Although a larger q makes Terms 4, 5, and 6 bigger, as long as n and p are sufficiently large, those terms will still be close to 0. Similarly, if we increase m, then $J_m(n,d)$ will approach the order of $n^{2(d-1)}$. As a result, Term 3 approaches the order of $n^{2d-0.5}$ times $\|k\|\epsilon\|k\|_2$ and the requirement $p \geq 6J_m(n,d)\ln\left(4n^{1+\frac{1}{m}}\right)$ approaches the order of $n^{2(d-1)}\ln n$.

Remark 1. We note that (Arora et al., 2019) shows that, for two-layer neural networks whose bottom-layer weights are trained by gradient descent, the generalization error for sufficiently large p has the following upper bound: for any $\zeta > 0$,

$$\begin{aligned}
&\Pr\Bigg\{\mathop{\mathbb{E}}_{\boldsymbol{x}}|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \sqrt{\frac{2\boldsymbol{y}^T(\mathbf{H}^\infty)^{-1}\boldsymbol{y}}{n}} \\
&+ O\left(\sqrt{\frac{\log\frac{n}{\zeta\,\min\text{eig}(\mathbf{H}^\infty)}}{n}}\right)\Bigg\} \geq 1-\zeta,
\end{aligned} \tag{11}$$

where $\mathbf{H}^\infty = \lim_{p\to\infty}(\mathbf{H}\mathbf{H}^T/p) \in \mathbb{R}^{n\times n}$. For certain class of learnable functions (we will compare them with our F^{ℓ_2} in Section 4), the quantity $\boldsymbol{y}^T(\mathbf{H}^\infty)^{-1}\boldsymbol{y}$ is bounded. Thus, $\sqrt{\frac{2\boldsymbol{y}^T(\mathbf{H}^\infty)^{-1}\boldsymbol{y}}{n}}$ also decreases at the speed $1/\sqrt{p}$. The second $O(\cdot)$-term in Eq. (11) contains the minimum eigenvalue of $\mathbf{H}^\infty$, which decreases with n. (Indeed, we show that this minimum eigenvalue is upper bounded by $O(n^{-\frac{1}{d-1}})$ in Supplementary Material, Appendix G.) Thus, Eq. (11) may decrease a little bit slower than $1/\sqrt{p}$, which is consistent with Term 1 in Eq. (10) (when q is large). Note that the term $2\boldsymbol{y}^T(\mathbf{H}^\infty)^{-1}\boldsymbol{y}$ in Eq. (11) captures how the complexity of the ground-truth function affects the generalization error. Similarly, the norm of $g(\cdot)$ also captures the impact[4] of the complexity of the ground-truth function in Eq. (10). However, we caution that the GD solution in (Arora et al., 2019) is based on the original neural network, which is usually different from our min ℓ_2-norm solution based on the NTK model (even though they are close for very large p). Thus, the two results may not be directly comparable.

Theorem 1 reveals several important insights on the generalization performance when the ground-truth function belongs to F^{ℓ_2}.

(i) Descent in the overparameterized region: When p increases, both sides of Eq. (10) decreases, suggesting that the test error of the overfitted NTK model decreases with

[2]The requirement of $\|k\|_1 < \infty$ can be relaxed. We show in Supplementary Material, Appendix L that, even when g is a δ-function (so $\|k\|_1 = \infty$), we can still have a similar result of Eq. (10) but Term 1 will have a slower speed of decay $O(n^{-\frac{1}{2(d-1)}\left(1-\frac{1}{q}\right)})$ with respect to n instead of $O(n^{\frac{1}{2}\left(1-\frac{1}{q}\right)})$ shown in Eq. (10). Term 4 of Eq. (10) will also be different when g is a δ-function, but it still goes to zero when p and n are large.

[3]The notion $\Pr_M$ in Eq. (10) emphasizes that randomness is in M.

[4]Although Term 1 in Eq. (10) in its current form does not depend on $g(\cdot)$, it is possible to modify our proof so that the norm of $g(\cdot)$ also enters Term 1.

p. In Fig. 1(a), we choose a ground-truth function in F^{ℓ_2} (we will explain why this function is in F^{ℓ_2} later in Section 4). The test MSE of the aforementioned NTK model (green curve) confirms the overall trend[5] of descent in the overparameterized region. We note that while (Arora et al., 2019) provides a generalization error upper-bound for large p (i.e., Eq. (11)), the upper bound there does not capture the dependency in p and thus does not predict this descent.

More importantly, we note a significant difference between the descent in Theorem 1 and that of min ℓ_2-norm overfitting solutions for linear models with simple features (Belkin et al., 2018b; 2019; Bartlett et al., 2020; Hastie et al., 2019; Muthukumar et al., 2019; Liao et al., 2020; Jacot et al., 2020). For example, for linear models with Gaussian features, we can obtain (see, e.g., Theorem 2 of (Belkin et al., 2019)):

$$\text{MSE} = \|f\|_2^2 \left(1 - \frac{n}{p}\right) + \frac{\sigma^2 n}{p - n - 1}, \text{ for } p > n + 2 \tag{12}$$

where σ^2 denotes the variance of the noise. If we let $p \to \infty$ in Eq. (12), we can see that the MSE quickly approaches $\|f\|_2^2$, which is referred to as the "null risk" (Hastie et al., 2019), i.e., the MSE of a model that predicts zero. Note that the null-risk is at the level of the signal, and thus is quite large. In contrast, as $p \to \infty$, the test error of the NTK model converges to a value determined by n and ϵ (and is independent of the null risk). This difference is confirmed in Fig. 1(a), where the test MSE for the NTK model (green curve) is much lower than the null risk (the dashed line) when $p \to \infty$, while both the min ℓ_2-norm (the red curve) and the min ℓ_1-norm solutions (the blue curve) (Ju et al., 2020) with Fourier features rise to the null risk when $p \to \infty$. Finally, note that the descent in Theorem 1 requires p to increase much faster than n. Specifically, to keep Term 2 in Eq. (10) small, it suffices to let p increase a little bit faster than $\Omega(n^{4d-1})$. This is again quite different from the descent shown in Eq. (12) and in other related work using Fourier and Gaussian features (Liao et al., 2020; Jacot et al., 2020), where p only needs to grow proportionally with n.

(ii) Speed of the descent: Since Theorem 1 holds for finite p, it also characterizes the speed of descent. In particular, Term 2 is proportional to $p^{\frac{1}{2}(1-\frac{1}{q})}$, which approaches $1/\sqrt{p}$ when q is large. Again, such a speed of descent is not captured in (Arora et al., 2019). As we show in Fig. 1(a), the test error of the gradient descent solution under the original neural network (cyan curve) is usually quite close to that of the NTK model (green curve). Thus, our result provides useful guidance on how fast the generalization error descends with p for such neural networks.

[5]This curve oscillates at the early stage when p is small. We suspect it is because, at small p, the convergence in Eq. (7) has not occurred yet, and thus the randomness in $\mathbf{V}_0[j]$ makes the simulation results more volatile.

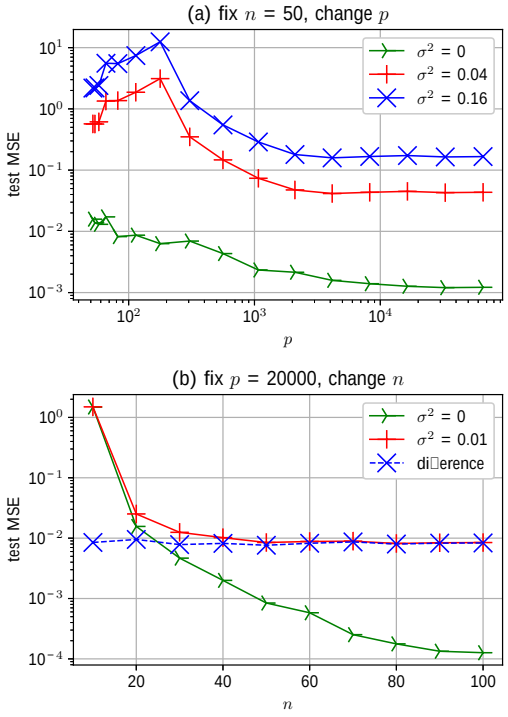

Figure 3. The test MSE of the overfitted NTK model for the same ground-truth function as Fig. 1(a). **(a)** We fix $n = 50$ and increase p for different noise level σ^2. **(b)** We fix $p = 20000$ and increase n. All data points in this figure are the average of five random simulation runs.

(iii) The effect of noise: Term 3 in Eq. (10) characterizes the impact of the noise ϵ, which does not decrease or increase with p. Notice that this is again very different from Eq. (12), i.e., results of min ℓ_2-norm overfitting solutions for simple features, where the noise term $\frac{\sigma^2 n}{p-n-1} \to 0$ when $p \to \infty$. We use Fig. 3(a) to validate this insight. In Fig. 3(a), we fix $n = 50$ and plot curves of test MSE of NTK overfitting solution as p increases. We let the noise ϵ_i in the i-th training sample be *i.i.d.* Gaussian with zero mean and variance σ^2. The green, red, and blue curves in Fig. 3(a) corresponds to the situation $\sigma^2 = 0$, $\sigma^2 = 0.04$, and $\sigma^2 = 0.16$, respectively. We can see that all three curves become flat when p is very large, and this phenomenon implies that the gap across different noise levels does not decrease when $p \to \infty$, which is in contrast to Eq. (12).

In Fig. 3(b), we instead fix $p = 20000$, and increase n). We plot the test MSE both for the noiseless setting (green curve) and for $\sigma^2 = 0.01$ (red curve). The difference between the two curves (dashed blue curve) then captures the impact of noise, which is related to Term 3 in Eq. (10). Somewhat

surprisingly, we find that the dashed blue curve is insensitive to n, which suggests that Term 3 in Eq. (10) may have room for improvement.

In summary, we have shown that any ground-truth function in F^{ℓ_2} leads to low generalization error for overfitted NTK models. It is then natural to ask what happens if the ground-truth function is not in F^{ℓ_2}. Let $\overline{F^{\ell_2}}$ denote the closure[6] of F^{ℓ_2}, and $D(f, F^{\ell_2})$ denotes the L^2-distance between f and F^{ℓ_2} (i.e., the infimum of the L^2-distance from f to every function in F^{ℓ_2}).

Proposition 2. ***(i)*** *For any given $(\mathbf{X}, \boldsymbol{y})$, there exists a function $\hat{f}_1^{\ell_2} \in F^{\ell_2}$ such that, uniformly over all $\boldsymbol{x} \in S^{d-1}$, $\hat{f}^{\ell_2}(\boldsymbol{x}) \xrightarrow{p} \hat{f}_1^{\ell_2}(\boldsymbol{x})$ as $p \to \infty$.* ***(ii)*** *Consequently, if the ground-truth function $f \notin \overline{F^{\ell_2}}$ (or equivalently, $D(f, F^{\ell_2}) > 0$), then the MSE of $\hat{f}_1^{\ell_2}$ (with respect to the ground-truth function f) is at least $D(f, F^{\ell_2})$.*

Intuitively, Proposition 2 (proven in Supplementary Material Appendix J) suggests that, if a ground-truth function is outside the closure of F^{ℓ_2}, then no matter how large n is, the test error of a NTK model with infinitely many neurons cannot be small (regardless whether or not the training samples contain noise). We validate this in Fig. 1(b), where a ground-truth function is chosen outside $\overline{F^{\ell_2}}$. The test MSE of NTK overfitting solutions (green curve) is above null risk (dashed black line) and thus is much higher compared with Fig. 1(a). We also plot the test MSE of the GD solution of the real neural network (cyan curve), which seems to show the same trend.

Comparing Theorem 1 and Proposition 2, we can clearly see that, all functions in F^{ℓ_2} are learnable by the overfitted NTK model, and all functions not in $\overline{F^{\ell_2}}$ are not.

4. What Exactly are the Functions in F^{ℓ_2}?

Our expression for learnable functions in Definition 1 is still in an indirect form, i.e., through the unknown function $g(\cdot)$. In (Arora et al., 2019), the authors show that all functions of the form $(\boldsymbol{x}^T\boldsymbol{a})^l$, $l \in \{0, 1, 2, 4, 6, \ldots\}$ are learnable by GD (assuming large p and small step size), for a similar 2-layer network with ReLU activation that has no bias. In the following, we will show that our learnable functions in Definition 1 also have a similar form. Further, we can show that any functions of the form $(\boldsymbol{x}^T\boldsymbol{a})^l$, $l \in \{3, 5, 7, \ldots\}$ are not learnable. Our characterization uses an interesting connection to harmonics and filtering on S^{d-1}, which may be of independent interest.

Towards this end, we first note that the integral form in Definition 1 can be viewed as a convolution on S^{d-1} (denoted by $\sim$). Specifically, for any $f_g \in F^{\ell_2}$, we can rewrite it as

$$f_g(\boldsymbol{x}) = g \sim h(\boldsymbol{x}) := \int_{\mathsf{SO}(d)} g(\mathbf{S}\boldsymbol{e}) h(\mathbf{S}^{-1}\boldsymbol{x}) d\mathbf{S}, \quad (13)$$

$$h(\boldsymbol{x}) := \boldsymbol{x}^T\boldsymbol{e} \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{e})}{2\pi}, \quad (14)$$

where $\boldsymbol{e} := [0\ 0\ \cdots\ 0\ 1]^T \in \mathbb{R}^d$, and $\mathbf{S}$ is a $d \times d$ orthogonal matrix that denotes a rotation in S^{d-1}, chosen from the set $\mathsf{SO}(d)$ of all rotations. An important property of the convolution Eq. (13) is that it corresponds to multiplication in the frequency domain, similar to Fourier coefficients. To define such a transformation to the frequency domain, we use a set of hyper-spherical harmonics $Y_{\mathbf{K}}^{l}$ (Vilenkin, 1968; Dokmanic & Petrinovic, 2009) when $d \geq 3$, which forms an orthonormal basis for functions on S^{d-1}. These harmonics are indexed by l and $\mathbf{K}$, where $\mathbf{K} = (k_1, k_2, \ldots, k_{d-2})$ and $l = k_0 \geq k_1 \geq k_2 \geq \cdots \geq k_{d-2} \geq 0$ (those k_i's and l are all non-negative integers). Any function $f \in L^2(S^{d-1} \to \mathbb{R})$ (including even δ-functions (Li & Wong, 2013)) can be decomposed uniquely into these harmonics, i.e., $f(\boldsymbol{x}) = \sum_l \sum_{\mathbf{K}} c_f(l, \mathbf{K})\, Y_{\mathbf{K}}^{l}(\boldsymbol{x})$, where $c_f(\cdot, \cdot)$ are projections of f onto the basis function. In Eq. (13), let $c_g(\cdot, \cdot)$ and $c_h(\cdot, \cdot)$ denote the coefficients corresponding to the decompositions of g and h, respectively. Then, we must have (Dokmanic & Petrinovic, 2009)

$$c_{f_g}(l, \mathbf{K}) = \alpha\, c_g(l, \mathbf{K}) c_h(l, \mathbf{0}), \quad (15)$$

where α is some normalization constant. Notice that in Eq. (15), the coefficient for h is $c_h(l, \mathbf{0})$ instead of $c_h(l, \mathbf{K})$, which is due to the intrinsic rotational symmetry of such convolution (Dokmanic & Petrinovic, 2009).

The above decomposition has an interesting "filtering" interpretation as follows. We can regard the function h as a "filter" or "channel," while the function g as a transmitted "signal." Then, the function f_g in Eq. (13) and Eq. (15) can be regarded as the received signal after g goes through the channel/filter h. Therefore, when coefficient $c_h(l, \mathbf{0})$ of h is non-zero, then the corresponding coefficient $c_{f_g}(l, \mathbf{K})$ for f_g can be any value (because we can arbitrarily choose g). In contrast, if a coefficient $c_h(l, \mathbf{0})$ of h is zero, then the corresponding coefficient $c_{f_g}(l, \mathbf{K})$ for f_g must also be zero for all $\mathbf{K}$.

Ideally, if h contains all "frequencies," i.e., all coefficients $c_h(l, \mathbf{0})$ are non-zero, then f_g can also contain all "frequencies," which means that F^{ℓ_2} can contain almost all functions. Unfortunately, this is not true for the function h given in Eq. (14). Specifically, using the harmonics defined in (Dokmanic & Petrinovic, 2009), the basis $Y_{\mathbf{0}}^{l}$ for $(l, \mathbf{0})$ turns out to have the form

$$Y_{\mathbf{0}}^{l}(\boldsymbol{x}) = \sum_{k=0}^{\frac{l}{2}} (-1)^k\ a_{l,k}\ (\boldsymbol{x}^T\boldsymbol{e})^{l-2k}, \quad (16)$$

[6] We consider the normed space of all functions in $L^2(S^{d-1} \to \mathbb{R})$. Notice that although $g(\boldsymbol{z})$ in Definition 1 may not be in L^2, f_g is always in L^2. Specifically, $f_g(\boldsymbol{x})$ is bounded for every $\boldsymbol{x} \in S^{d-1}$ when $\|g\|_1 < \infty$.

where $a_{l,k}$ are positive constants. Note that the expression Eq. (16) contains either only even powers of $\boldsymbol{x}^T\boldsymbol{e}$ (if l is even) or odd powers of $\boldsymbol{x}^T\boldsymbol{e}$ (if l is odd). Then, for the function h in Eq. (14), we have the following proposition (proven in Supplementary Material, Appendix K.4). We note that (Basri et al., 2019) has a similar harmonics analysis, where the expression of $c_h(l,\mathbf{0})$ is given. However, it is not obvious that the expression of $c_h(l,\mathbf{0})$ for all $l = 0,1,2,4,6,$ given in (Basri et al., 2019) must be non-zero, which is made clear by Proposition 3 as follows.

Proposition 3. *$c_h(l,\mathbf{0})$ is zero for $l = 3,5,7,$ and is non-zero for $l = 0,1,2,4,6,$.*

We are now ready to characterize what functions are in F^{ℓ_2}. By the form of Eq. (16), for any non-negative integer k, any even power $(\boldsymbol{x}^T\boldsymbol{e})^{2k}$ is a linear combination of $^{0}_{\mathbf{0}}, ^{2}_{\mathbf{0}}, \ , ^{2k}_{\mathbf{0}}$, and any odd power $(\boldsymbol{x}^T\boldsymbol{e})^{2k+1}$ is a linear combination of $^{1}_{\mathbf{0}}, ^{3}_{\mathbf{0}}, \ , ^{2k+1}_{\mathbf{0}}$. By Proposition 3, we thus conclude that any function $f_g(\boldsymbol{x}) = (\boldsymbol{x}^T\boldsymbol{e})^l$ where $l \in \{0,1,2,4,6,\ g$ can be written in the form of Eq. (15) in the frequency domain, and thus are in F^{ℓ_2}. In contrast, any function $f(\boldsymbol{x}) = (\boldsymbol{x}^T\boldsymbol{e})^l$ where $l \in \{3,5,7,\ g$ cannot be written in the form of Eq. (15), and are thus not in F^{ℓ_2}. Further, the ℓ_2-norm of any latter function will also be equal to its distance to F^1. Therefore, the generalization-error lower-bound in Proposition 2 will apply (with $D(f,F^{\ell_2}) = \|f\|_2$). Finally, by Eq. (13), F^{ℓ_2} is invariant under rotation and finite linear summation. Therefore, any finite sum of $(\boldsymbol{x}^T\boldsymbol{a})^l$, $l = 0,1,2,4,6,$ must also belong to F^{ℓ_2}.

For the special case of $d = 2$, the input $\boldsymbol{x}$ corresponds to an angle $\theta \in [-\pi,\pi]$, and the above-mentioned harmonics become Fourier series $\sin(k\theta)$ and $\cos(k\theta)$, $k = 0,1,$. We can then get similar results that frequencies of $k \in \{0,1,2,4,6,\ g$ are learnable (while others are not), which explains the learnable and not-learnable functions in Fig. 1. Details can be found in Supplementary Material, Appendix K.5.

Remark 2. We caution that the above claim on non-learnable functions critically depends on the network architecture. That is, we assume throughout this paper that the ReLU activation has no bias. It is known from an expressiveness point of view that, using ReLU without bias, a shallow network can only approximate the sum of linear functions and even functions (Ghorbani et al., 2019). Thus, it is not surprising that other odd-power (but non-linear) polynomials cannot be learned. In contrast, by adding a bias, a shallow network using ReLU becomes a universal approximator (Ji et al., 2019). The recent work of (Satpathi & Srikant, 2021) shows that polynomials with all powers can be learned by the corresponding 2-layer NTK model. These results are consistent with ours because a ReLU activation function operating on $\boldsymbol{x} \in \mathbb{R}^{d-1}$ with a bias can be equivalently viewed as one operating on a d-dimension input (with the last-dimension being fixed at $1/\sqrt{d}$) but with no bias. Even though only a subset of functions are learnable in the d-dimension space, when projected into a $(d-1)$-dimension subspace, they may already span all functions. For example, one could write $(\boldsymbol{x}^T\boldsymbol{a})^3$ as a linear combination of $(\begin{bmatrix}\tilde{\boldsymbol{x}}\\ 1/\sqrt{d}\end{bmatrix}^T\boldsymbol{b}_i)^{l_i}$, where $i \in \{1,2,\ ,5\}$, $[l_1,\ ,l_5] = [4,4,2,1,0]$, and $\boldsymbol{b}_i \in \mathbb{R}^d$ depends only on $\boldsymbol{a}$. (See Supplementary Material, Appendix K.6 for details.) It remains an interesting question whether similar difference arises for other network architectures (e.g., with more than 2 layers).

5. Proof Sketch of Theorem 1

In this section, we sketch the key steps to prove Theorem 1. Starting from Eq. (3), we have

$$\mathbf{V}^{\ell_2} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(\mathbf{F}(\mathbf{X}) + \boldsymbol{\epsilon}). \tag{17}$$

For the learned model $\hat{f}^{\ell_2}(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{V}^{\ell_2}$ given in Eq. (4), the error for any test input $\boldsymbol{x}$ is then

$$\begin{aligned}\hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x}) = &\left(\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{F}(\mathbf{X}) - f(\boldsymbol{x})\right)\\ &+ \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}.\end{aligned} \tag{18}$$

In the classical "bias-variance" analysis with respect to MSE (Belkin et al., 2018a), the first term on the right-hand-side of Eq. (18) contributes to the bias and the second term contributes to the variance. We first quantify the second term (i.e., the variance) in the following proposition.

Proposition 4. *For any $n \geq 2$, $m \in \left[1, \frac{\ln n}{\ln\frac{\pi}{2}}\right]$, $d \geq n^4$, if $p \geq 6J_m(n,d)\ln\left(4n^{1+\frac{1}{m}}\right)$, we must have $\Pr_{\mathbf{X},\mathbf{V}_0}\left\{|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}| \leq \sqrt{J_m(n,d)n}\|\boldsymbol{\epsilon}\|_2, \text{ for all } \boldsymbol{\epsilon} \in \mathbb{R}^n\right\} \geq 1 - \frac{2}{n^m}$.*

The proof is in Supplementary Material Appendix F. Proposition 4 implies that, for fixed n and d, when $p \gg 1$, with high probability the variance will not exceed a certain factor of the noise $\|\boldsymbol{\epsilon}\|_2$. In other words, the variance will not go to infinity when $p \gg 1$. The main step in the proof is to lower bound $\min\text{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$, which is given by $1/(J_m(n,d)n)$. Note that this is the main place where we used the assumption that $\boldsymbol{x}$ is uniformly distributed. We expect that our main proof techniques can be generalized to other distributions (with a different expression of $J_m(n,d)$), which we leave for future work.

Remark 3. In the upper bound in (Arora et al., 2019) (i.e., Eq. (11)), any noise added to $\boldsymbol{y}$ will at least contribute to the generalization upper bound Eq. (11) by a positive term $\boldsymbol{\epsilon}^T(\mathbf{H}^1)^{-1}\boldsymbol{\epsilon}/n$. Thus, their upper bound may also grow as $\min\text{eig}(\mathbf{H}^1)$ decreases. One of the contribution of Proposition 4 is to characterize this minimum eigenvalue.

We now bound the bias part. We first study the class of ground-truth functions that can be learned with fixed $\mathbf{V}_0$. We refer to them as *pseudo ground-truth*, to differentiate them with the set F^{ℓ_2} of learnable functions for random $\mathbf{V}_0$. They are defined with respect to the same $g(\,)$ function, so that we can later extend to the "real" ground-truth functions in F^{ℓ_2} when considering the randomness of $\mathbf{V}_0$.

Definition 2. *Given $\mathbf{V}_0$, for any learnable ground-truth function $f_g \in F^{\ell_2}$ with the corresponding function $g(\,)$, define the corresponding* ***pseudo ground-truth*** *as*

$$f^g_{\mathbf{V}_0}(\boldsymbol{x}) := \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{\mathbb{I}[\boldsymbol{z},\boldsymbol{x}]}{p} g(\boldsymbol{z}) d\mu(\boldsymbol{z}).$$

The reason that this class of functions may be the learnable functions for fixed $\mathbf{V}_0$ is similar to the discussions in Eq. (5) and Eq. (6). Indeed, using the same choice of $g(\boldsymbol{z})$ in Eq. (8), the learned function $\hat{f}^{\ell_2}$ in Eq. (5) at fixed $\mathbf{V}_0$ is always of the form in Definition 2.

The following proposition gives an upper bound of the generalization performance when the data model is based on the pseudo ground-truth and the NTK model uses exactly the same $\mathbf{V}_0$.

Proposition 5. *Assume fixed $\mathbf{V}_0$ (thus p and d are also fixed), there is no noise. If the ground-truth function is $f = f^g_{\mathbf{V}_0}$ in Definition 2 and $\|g\|_1 < 1$, then for any $\boldsymbol{x} \in S^{d-1}$ and $q \in [1, 1)$, we have* $\Pr_{\mathbf{X}}\{|\hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x})| \le n^{\frac{1}{2}(1-\frac{1}{q})}\} \ge 1 - 2e^2 \exp\left(-\frac{p}{8\|g\|_1^2 q \bar{p}}\right)$.

The proof is in Supplementary Material, Appendix H. Note that both the threshold of the probability event and the upper bound coincide with Term 1 and Term 4, respectively, in Eq. (10). Here we sketch the proof of Proposition 5. Based on the definition of the pseudo ground-truth, we can rewrite $f^g_{\mathbf{V}_0}$ as $f^g_{\mathbf{V}_0}(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^{dp}$ is given by, for all $j \in \{1, 2, \ldots, p\}$, $\mathbf{V}[j] = \int_{S^{d-1}} \mathbf{1}_{\{\boldsymbol{z}^T \mathbf{V}_0[j] > 0\}} \boldsymbol{z} \frac{g(\boldsymbol{z})}{p} d\mu(\boldsymbol{z})$. From Eq. (3) and Eq. (4), we can see that the learned model is $\hat{f}^{\ell_2}(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \mathbf{P} \mathbf{V}$ where $\mathbf{P} := \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}$. Note that $\mathbf{P}$ is an orthogonal projection to the row-space of $\mathbf{H}$. Further, it is easy to show that $\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \le \sqrt{p/\bar{p}}$. Thus, we have $|\hat{f}^{\ell_2}(\boldsymbol{x}) - f^g_{\mathbf{V}_0}(\boldsymbol{x})| = |\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} - \mathbf{I})\mathbf{V}| \le \sqrt{p/\bar{p}}\|(\mathbf{P} - \mathbf{I})\mathbf{V}\|_2$. The term $(\mathbf{P} - \mathbf{I})\mathbf{V}$ can be interpreted as the distance from $\mathbf{V}$ to the row-space of $\mathbf{H}$. Note that this distance is no greater than the distance between $\mathbf{V}$ and any point in the row-space of $\mathbf{H}$. Thus, in order to get an upper bound on $\|(\mathbf{P} - \mathbf{I})\mathbf{V}\|_2$, we only need to find a vector $\boldsymbol{a} \in \mathbb{R}^n$ that makes $\|\mathbf{V} - \mathbf{H}^T\boldsymbol{a}\|_2$ as small as possible, especially when n is large. Our proof uses the vector $\boldsymbol{a}$ such that its i-th element is $\boldsymbol{a}_i := \frac{g(\mathbf{X}_i)}{np}$. See Supplementary Material, Appendix H for the rest of the details.

The final step is to allow $\mathbf{V}_0$ to be random. Given any random $\mathbf{V}_0$, any function $f_g \in F^{\ell_2}$ can be viewed as the summation of a pseudo ground-truth function (with the same $g(\,)$) and a difference term. This difference can be viewed as a special form of "noise", and thus we can use Proposition 4 to quantify its impact. Further, the magnitude of this "noise" should decrease with p (because of Eq. (7)). Combining this argument with Proposition 5, we can then prove Theorem 1. See Supplementary Material, Appendix I for details.

6. Conclusions

In this paper, we studied the generalization performance of the min ℓ_2-norm overfitting solution for a two-layer NTK model. We provide a precise characterization of the learnable ground-truth functions for such models, by providing a generalization upper bound for all functions in F^{ℓ_2}, and a generalization lower bound for all functions not in F^{ℓ_2}. We show that, while the test error of the overfitted NTK model also exhibits descent in the overparameterized regime, the descent behavior can be quite different from the double descent of linear models with simple features.

There are several interesting directions for future work. First, based on Fig. 3(b), our estimation of the effect of noise could be further improved. Second, it would be interesting to explore whether the methodology can be extended to NTK model for other neural networks, e.g., with different activation functions and with more than two layers.

Acknowledgements

This work is partially supported by an NSF sub-award via Duke University (IIS-1932630), by NSF grants CNS-1717493, CNS-1901057, and CNS-2007231, and by Office of Naval Research under Grant N00014-17-1-241. The authors would like to thank Professor R. Srikant at the University of Illinois at Urbana-Champaign and anonymous reviewers for their valuable comments and suggestions.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The convergence rate of neural networks for learned functions of different frequencies. *arXiv preprint arXiv:1906.00425*, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018a.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549, 2018b.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

Bishop, C. M. *Pattern recognition and machine learning*. Springer, 2006.

Chaudhry, M. A., Qadir, A., Rafique, M., and Zubair, S. Extension of euler's beta function. *Journal of computational and applied mathematics*, 78(1):19–32, 1997.

Dokmanic, I. and Petrinovic, D. Convolution on the n-sphere with application to pdf modeling. *IEEE transactions on signal processing*, 58(3):1157–1170, 2009.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Dutka, J. The incomplete beta function—a historical profile. *Archive for history of exact sciences*, pp. 11–29, 1981.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Fiat, J., Malach, E., and Shalev-Shwartz, S. Decoupling gating from linearity. *arXiv preprint arXiv:1906.05032*, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Goemans, M. Chernoff bounds, and some applications. *URL http: //math.mit.edu /goemans /18310S15 /chernoff-notes.pdf*, 2015.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hayes, T. P. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

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

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020.

James, W. and Stein, C. Estimation with quadratic loss. In *Breakthroughs in Statistics*, pp. 443–460. Springer, 1992.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. *arXiv preprint arXiv:1910.06956*, 2019.

Ju, P., Lin, X., and Liu, J. Overfitting can be harmless for basis pursuit, but only to a degree. *Advances in Neural Information Processing Systems*, 33, 2020.

LeCun, Y., Kanter, I., and Solla, S. A. Second order properties of error surfaces: Learning time and generalization. In *Advances in Neural Information Processing Systems*, pp. 918–924, 1991.

Li, S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, 31: 8157–8166, 2018.

Li, Y. and Wong, R. Integral and series representations of the dirac delta function. *arXiv preprint arXiv:1303.1943*, 2013.

Liao, Z., Couillet, R., and Mahoney, M. W. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *arXiv preprint arXiv:2006.05013*, 2020.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Muthukumar, V., Vodrahalli, K., and Sahai, A. Harmless interpolation of noisy data in regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2299–2303. IEEE, 2019.

Rao, K. B. and Rao, M. B. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.

Satpathi, S. and Srikant, R. The dynamics of gradient descent for overparametrized neural networks. In *3rd Annual Learning for Dynamics and Control Conference (L4DC)*, 2021.

Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States, 1956.

Tikhonov, A. N. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pp. 195–198, 1943.

Vilenkin, N. Y. Special functions and the theory of group representations. providence: American mathematical society. *sftp*, 1968.

Wainwright, M. Uniform laws of large numbers, 2015. https://www.stat.berkeley.edu/~mjwain/stat210b/Chap4_Uniform_Feb4_2015.pdf, Accessed: Feb. 7, 2021.

Wendel, J. G. A problem in geometric probability. *Math. Scand*, 11:109–111, 1962.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

A. Extra Notations

In addition to the notations that we have introduced in the main body of this paper, we need some extra notations that are used in the following appendices. The distribution of the initial weights $\mathbf{V}_0[j]$ is denoted by the probability density $\lambda(\ )$ on R^d, and the directions of the initial weights (i.e., the normalized initial weights $\frac{\mathbf{V}_0[j]}{k\mathbf{V}_0[j]k_2}$) follows the probability density $\tilde{\lambda}(\ )$ on $S^{d\ 1}$. Let $\lambda_a(\ )$ be the Lebesgue measure on R^a where the dimension a can be, e.g., $(d\ \ 1)$ and $(d\ \ 2)$.

Let $\mathsf{Bino}(a,b)$ denote the binomial distribution, where a is the number of trials and b is the success probability. Let $I(\ ,\ )$ denote the regularized incomplete beta function (Dutka, 1981). Let $B(\ ,\ )$ denote the beta function (Chaudhry et al., 1997). Specifically,

$$B(x,y) := \int_0^1 t^{x\ 1}(1\ \ t)^{y\ 1}dt, \tag{19}$$

$$I_x(a,b) := \frac{\int_0^x t^{a\ 1}(1\ \ t)^{b\ 1}dt}{B(a,b)}. \tag{20}$$

Define a cap on a unit hyper-sphere $S^{d\ 1}$ as the intersection of $S^{d\ 1}$ with an open ball in R^d centered at $\boldsymbol{v}$ with radius r, i.e.,

$$B_{\boldsymbol{v}}^r := \left\{\boldsymbol{v}\ 2\ S^{d\ 1}\ j\ k\boldsymbol{v}\ \ \boldsymbol{v}\ k_2 < r\right\}. \tag{21}$$

Remark 4. For ease of exposition, we will sometimes neglect the subscript $\boldsymbol{v}$ of $B_{\boldsymbol{v}}^r$ and use B^r instead, when the quantity that we are estimating only depends on r but not $\boldsymbol{v}$. For example, where we are interested in the area of $B_{\boldsymbol{v}}^r$, it only depends on r but not $\boldsymbol{v}$. Thus, we write $\lambda_{d\ 1}(B^r)$ instead.

For any $\boldsymbol{x}\ 2\ \mathsf{R}^d$ such that $\boldsymbol{x}^T\boldsymbol{v}\ = 0$, define two halves of the cap $B_{\boldsymbol{v}}^r$ as

$$B_{\boldsymbol{v},+}^{r,\boldsymbol{x}} := \left\{\boldsymbol{v}\ 2\ B_{\boldsymbol{v}}^r\ \ j\ \boldsymbol{x}^T\boldsymbol{v} > 0\right\}, \quad B_{\boldsymbol{v},}^{r,\boldsymbol{x}} := \left\{\boldsymbol{v}\ 2\ B_{\boldsymbol{v}}^r\ \ j\ \boldsymbol{x}^T\boldsymbol{v} < 0\right\}. \tag{22}$$

Define the set of directions of the initial weights $\mathbf{V}_0[j]$'s as

$$A_{\mathbf{V}_0} := \left\{ \frac{\mathbf{V}_0[j]}{k\mathbf{V}_0[j]k_2} \,\middle|\, j\ 2\ f1,2,\ \ \ ,pg \right\}. \tag{23}$$

B. GD (gradient descent) Converges to Min ℓ_2-Norm Solutions

We assume that the GD algorithm for minimizing the training MSE is given by

$$\mathbf{V}_{k+1}^{\mathrm{GD}} = \ \mathbf{V}_k^{\mathrm{GD}}\ \ \gamma_k\sum_{i=1}^n(\mathbf{H}_i\ \ \mathbf{V}_k^{\mathrm{GD}}\ \ y_i)\mathbf{H}_i^T, \tag{24}$$

where $\mathbf{V}_k^{\mathrm{GD}}$ denotes the solution in the k-th GD iteration ($\mathbf{V}_0^{\mathrm{GD}} = \mathbf{0}$), and γ_k denotes the step size of the k-th iteration.

Lemma 6. *If* $\mathbf{V}^{\ell_2}$ *exists and GD in Eq.* (24) *converges to zero-training loss (i.e.,* $\mathbf{H}\ \ \mathbf{V}_1^{GD} = \boldsymbol{y}$*), then* $\mathbf{V}_1^{GD} = \ \mathbf{V}^{\ell_2}$.

Proof. Because $\mathbf{V}_0^{\mathrm{GD}} = \mathbf{0}$ and Eq. (24), we know that $\mathbf{V}_k^{\mathrm{GD}}$ is in the row space of $\mathbf{H}$ for any k. Thus, we can let $\mathbf{V}_1^{\mathrm{GD}} = \mathbf{H}^T\boldsymbol{a}$ where $\boldsymbol{a}\ 2\ \mathsf{R}^n$. When GD converges to zero training loss, we have $\mathbf{H}\ \ \mathbf{V}_1^{\mathrm{GD}} = \boldsymbol{y}$. Thus, we have $\mathbf{H}\mathbf{H}^T\boldsymbol{a} = \boldsymbol{y}$, which implies $\boldsymbol{a} = (\mathbf{H}\mathbf{H}^T)^{\ 1}\boldsymbol{y}$. Therefore, we must have $\mathbf{V}_1^{\mathrm{GD}} = \mathbf{H}^T\boldsymbol{a} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{\ 1}\boldsymbol{y} = \ \mathbf{V}^{\ell_2}$. □

C. Assumptions and Justifications

Because $\hat{f}_{\Delta\mathbf{V},\mathbf{V}_0}(a\boldsymbol{x}) = a\ \ \hat{f}_{\Delta\mathbf{V},\mathbf{V}_0}(\boldsymbol{x})$ for any $a\ 2\ \mathsf{R}$, we can always do preprocessing to normalize the input $\boldsymbol{x}$. For simplicity, we focus on the simplest situation that the randomness for the inputs and the initial weights are uniform. Nonetheless, methods and results of this paper can be readily generalized to other continuous random variable distributions, which we leave for future work. We thus make the following Assumption 1.

Assumption 1. *The input $\boldsymbol{x}$ are uniformly distributed in S^{d-1}. The initial weights $\mathbf{V}_0[j]$'s are uniform in all directions. In other words, $\mu(\cdot)$ and $\pi(\cdot)$ are both* $\mathsf{unif}(S^{d-1})$.

We study the overparameterized and overfitted setting, so in this paper we always assume $p \geq n/d$, i.e., the number of parameters pd is larger than or equal to the number of training samples n. The situation of $d = 1$ is relatively trivial, so we only consider the case $d \geq 2$. We then make Assumption 2.

Assumption 2. $p \geq n/d$ *and* $d \geq 2$.

If the input is a continuous random vector, then for any $i \neq j$, we have $\Pr\{\mathbf{X}_i = \mathbf{X}_j\} = 0$ and $\Pr\{\mathbf{X}_i = -\mathbf{X}_j\} = 0$ (because the probability that a continuous random variable equals to a given value is zero). Thus, $\Pr\{\mathbf{X}_i \parallel \mathbf{X}_j\} = 0$, and $\Pr\{\mathbf{X}_i \not\parallel \mathbf{X}_j\} = 1$. Similarly, we can show that $\Pr\{\mathbf{V}_0[k] \not\parallel \mathbf{V}_0[l]\} = 1$. We thus make Assumption 3.

Assumption 3. $\mathbf{X}_i \not\parallel \mathbf{X}_j$ *for any* $i \neq j$, *and* $\mathbf{V}_0[k] \not\parallel \mathbf{V}_0[l]$ *for any* $k \neq l$.

With these assumptions, the following lemma says that when p is large enough, with high probability $\mathbf{H}$ has full row-rank (and thus $\mathbf{V}^{\ell_2}$ exists).

Lemma 7. $\lim_{p \to \infty} \Pr_{\mathbf{V}_0}\{\mathrm{rank}(\mathbf{H}) = n\} = 1$.

Proof. See Appendix E. □

D. Some Useful Supporting Results

Here we collect some useful lemmas that are needed for proofs in other appendices, many of which are estimations of certain quantities that we will use later.

D.1. Quantities related to the area of a cap on a hyper-sphere

The following lemma is introduced by (Li, 2011), which gives the area of a cap on a hyper-sphere with respect to the colatitude angle.

Lemma 8. *Let $\phi \in [0, \frac{\pi}{2}]$ denote the colatitude angle of the smaller cap on S^{d-1}, then the area (in the measure of λ_{d-1}) of this hyper-spherical cap is*

$$\frac{1}{2}\lambda_{d-1}(S^{d-1}) I_{\sin^2\phi}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

The following lemma is another representation of the area of the cap with respect to the radius r (recall the definition of B^r in Eq. (21) and Remark 4).

Lemma 9. *If $r \leq \sqrt{2}$, then we have*

$$\lambda_{d-1}(B^r) = \frac{1}{2}\lambda_{d-1}(S^{d-1}) I_{r^2(1-\frac{r^2}{4})}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Proof. Let ϕ denote the colatitude angle. By the law of cosines, we have

$$\cos\phi = 1 - \frac{r^2}{2}.$$

Thus, we have

$$\sin^2\phi = 1 - \cos^2\phi = 1 - \left(1 - \frac{r^2}{2}\right)^2 = r^2\left(1 - \frac{r^2}{4}\right).$$

By Lemma 8, the result of this lemma thus follows. Notice that we require $r \leq \sqrt{2}$ to make sure that $\phi \in [0, \frac{\pi}{2}]$, which is required by Lemma 8. □

The area of a cap can be interpreted as the probability of the event that a uniformly-distributed random vector falls into that cap. We have the following lemma.

Lemma 10. *Suppose that a random vector $\boldsymbol{b} \in S^{d-1}$ follows uniform distribution in all directions. Given any $\boldsymbol{a} \in S^{d-1}$ and for any $c \in (0,1)$, we have*

$$\Pr_{\boldsymbol{b}}\left\{|\boldsymbol{a}^T\boldsymbol{b}| > c\right\} = I_{1-c^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Proof. Notice that $\left\{\boldsymbol{b} \mid \boldsymbol{a}^T\boldsymbol{b} > c\right\}$ is a hyper-spherical cap. Define its colatitude angle as ϕ. We have $\cos\phi = \boldsymbol{a}^T\boldsymbol{b} = c$. Thus, we have $\sin^2\phi = 1-c^2$. By Lemma 8, we then have

$$\lambda_{d-1}\left(\left\{\boldsymbol{b} \mid \boldsymbol{a}^T\boldsymbol{b} > c\right\}\right) = \frac{1}{2}\lambda_{d-1}(S^{d-1})I_{1-c^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Further, by symmetry, we have

$$\lambda_{d-1}\left(\left\{\boldsymbol{b} \mid |\boldsymbol{a}^T\boldsymbol{b}| > c\right\}\right) = 2\lambda_{d-1}\left(\left\{\boldsymbol{b} \mid \boldsymbol{a}^T\boldsymbol{b} > c\right\}\right) = \lambda_{d-1}(S^{d-1})I_{1-c^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Because $\boldsymbol{b}$ follows uniform distribution in all directions, we have

$$\Pr_{\boldsymbol{b}}\left\{|\boldsymbol{a}^T\boldsymbol{b}| > c\right\} = \frac{\lambda_{d-1}\left(\left\{\boldsymbol{b} \mid |\boldsymbol{a}^T\boldsymbol{b}| > c\right\}\right)}{\lambda_{d-1}(S^{d-1})} = I_{1-c^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

□

D.2. Estimation of certain norms

In this subsection, we will show $\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \le \sqrt{p}$ in Lemma 11. We also upper bound the norm of the product of two matrices by the product of their norms in Lemma 12. At last, Lemma 13 states that if two vector differ a lot, then the sum of their norm cannot be too small.

Lemma 11. *$\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \le \sqrt{p}$ for any $\boldsymbol{x} \in S^{d-1}$.*

Proof. This follows because the input $\boldsymbol{x}$ is normalized. Specifically, by Eq. (1), we have

$$\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 = \sqrt{\sum_{j=1}^{p}\left\|\mathbf{1}_{\{\boldsymbol{x}^T\mathbf{V}_0[j]>0\}}\boldsymbol{x}^T\right\|_2^2} \le \sqrt{p}. \tag{25}$$

□

Lemma 12. *If $\mathbf{C} = \mathbf{AB}$, then $\|\mathbf{C}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{B}\|_2$. Here $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ could be scalars, vectors, or matrices.*

Proof. This lemma directly follows the definition of matrix norm. □

Remark 5. Note that the (ℓ_2) matrix-norm (i.e., spectral norm) of a vector is exactly its ℓ_2 vector-norm (i.e., Euclidean norm)[7]. Therefore, when applying Lemma 12, we do not need to worry about whether $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are matrices or vectors.

Lemma 13. *For any $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$, we have*

$$\|\boldsymbol{v}_1\|_2^2 + \|\boldsymbol{v}_2\|_2^2 \ge \frac{1}{2}\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2^2.$$

[7] To see this, consider a (row or column) vector $\boldsymbol{a}$. The matrix norm of $\boldsymbol{a}$ is

$$\max_{\|x\|_2=1}\|\boldsymbol{a}x\|_2 \text{ (when } \boldsymbol{a} \text{ is a column vector)},$$
$$\text{or } \max_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{a}\boldsymbol{x}\|_2 \text{ (when } \boldsymbol{a} \text{ is a row vector)}.$$

In both cases, the value of the matrix-norm equals to $\sqrt{\sum a_i^2}$, which is exactly the ℓ_2-norm (Euclidean norm) of $\boldsymbol{a}$.

Proof. It is easy to prove that $k \quad k_2^2$ is convex. Thus, we have

$$\begin{aligned} k\boldsymbol{v}_1 k_2^2 + k\boldsymbol{v}_2 k_2^2 &= k\boldsymbol{v}_1 k_2^2 + k \quad \boldsymbol{v}_2 k_2^2 \\ &\quad 2\left\| \frac{\boldsymbol{v}_1 \quad \boldsymbol{v}_2}{2} \right\|_2^2 \text{ (apply Jensen's inequality on the convex function } k \quad k_2^2) \\ &= \frac{1}{2} k\boldsymbol{v}_1 \quad \boldsymbol{v}_2 k_2^2. \end{aligned}$$

□

D.3. Estimates of certain tail probabilities

The following is the (restated) Corollary 5 of (Goemans, 2015).

Lemma 14. *If the random variable X follows* $\mathsf{Bino}(a, b)$*, then for all* $0 < \delta < 1$*, we have*

$$\Pr\{jX \quad abj > \delta abg \quad 2e^{\ ab\delta^2/3}.$$

The following lemma is the (restated) Theorem 1.8 of (Hayes, 2005).

Lemma 15 (Azuma–Hoeffding inequality for random vectors)**.** *Let $X_1, X_2, \quad , X_k$ be* i.i.d. *random vectors with zero mean (of the same dimension) in a real Euclidean space such that $kX_i k_2 \quad 1$ for all $i = 1, 2, \quad , k$. Then, for every $a > 0$,*

$$\Pr\left\{ \left\| \sum_{i=1}^{k} X_i \right\|_2 \quad a \right\} < 2e^2 \exp\left(\quad \frac{a^2}{2k} \right).$$

In the following lemma, we use Azuma–Hoeffding inequality to upper bound the deviation of the empirical mean value of a bounded random vector from its expectation.

Lemma 16. *Let $X_1, X_2, \quad , X_k$ be* i.i.d. *random vectors (of the same dimension) in a real Euclidean space such that $kX_i k_2 \quad U$ for all $i = 1, 2, \quad , k$. Then, for any $q \ 2 \ [1, \ 1)$,*

$$\Pr\left\{ \left\| \left(\frac{1}{k} \sum_{i=1}^{k} X_i \right) \quad \mathsf{E}\, X_1 \right\|_2 \quad k^{\frac{1}{2q} \quad \frac{1}{2}} \right\} < 2e^2 \exp\left(\quad \frac{p_q \overline{k}}{8U^2} \right).$$

Proof. Because $kX_i k_2 \quad U$, we have $\mathsf{E}\, kX_i k_2 \quad U$. By triangle inequality, we have $kX_i \quad \mathsf{E}\, X_i k_2 \quad kX_i k_2 + \mathsf{E}\, kX_i k_2$ $2U$, i.e.,

$$\left\| \frac{X_i \quad \mathsf{E}\, X_i}{2U} \right\|_2 \quad 1. \tag{26}$$

We also have

$$\mathsf{E}\left[\frac{X_i \quad \mathsf{E}\, X_i}{2U} \right] = \frac{\mathsf{E}\, X_i \quad \mathsf{E}\, X_i}{2U} = \mathbf{0}. \tag{27}$$

We then have

$$\begin{aligned} &\Pr\left\{ \left\| \left(\frac{1}{k} \sum_{i=1}^{k} X_i \right) \quad \mathsf{E}\, X_1 \right\|_2 \quad k^{\frac{1}{2q} \quad \frac{1}{2}} \right\} \\ =\ &\Pr\left\{ \left\| \sum_{i=1}^{k} (X_i \quad \mathsf{E}\, X_i) \right\|_2 \quad k^{\frac{1}{2q} + \frac{1}{2}} \right\} \\ =\ &\Pr\left\{ \left\| \sum_{i=1}^{k} \left(\frac{X_i \quad \mathsf{E}\, X_i}{2U} \right) \right\|_2 \quad \frac{k^{\frac{1}{2q} + \frac{1}{2}}}{2U} \right\} \\ <\ &2e^2 \exp\left(\quad \frac{p_q \overline{k}}{8U^2} \right) \text{ (by Eqs. (26)(27) and letting } a = \frac{k^{\frac{1}{2q} + \frac{1}{2}}}{2U} \text{ in Lemma 15).} \end{aligned}$$

□

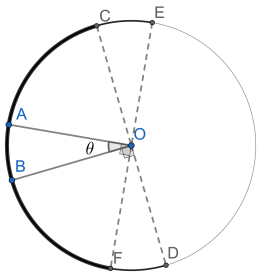

Figure 4. The arc $\overset{\frown}{\mathsf{CBF}}$ is $\frac{\pi - \theta}{2\pi}$ of the perimeter of the circle O.

D.4. Calculation of certain integrals

The following lemma calculates the ratio between the intersection area of two hyper-hemispheres and the area of the whole hyper-sphere.

Lemma 17.

$$\int_{S^{d-1}} \mathbf{1}_{\{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0\}} d\lambda(\boldsymbol{v}) = \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}. \tag{28}$$

(Recall that $\lambda(\cdot)$ denotes the distribution of the normalized version of $\mathbf{V}_0[j]$ on S^{d-1} and is assumed to be uniform in all directions.)

Before we give the proof of Lemma 17, we give its geometric explanation.

Geometric explanation of Eq. (28): Indeed, since λ is uniform on S^{d-1}, the integral on the left-hand-side of Eq. (28) represents the probability that a random point falls into the intersection of two hyper-hemispheres that are represented by $\{\boldsymbol{v} \in S^{d-1} \mid \boldsymbol{z}^T\boldsymbol{v} > 0\}$ and $\{\boldsymbol{v} \in S^{d-1} \mid \boldsymbol{x}^T\boldsymbol{v} > 0\}$, respectively. We can calculate that probability by

$$\frac{\text{measure of a hyper-spherical lune with angle } \pi - \theta(\boldsymbol{z},\boldsymbol{x})}{\text{measure of a unit hyper-sphere}} = \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}, \tag{29}$$

where $\theta(\cdot,\cdot)$ denote the angle (in radians) between two vectors, which would lead to Eq. (28). To help readers understand Eq. (29), we give examples for 2D and 3D in Fig. 4 and Fig. 5, respectively. In the 2D case depicted in Fig. 4, $\overrightarrow{\mathsf{OA}}$ denotes $\boldsymbol{z}$, $\overrightarrow{\mathsf{OB}}$ denotes $\boldsymbol{x}$. Thus, the arc $\overset{\frown}{\mathsf{EAF}}$ denotes $\{\boldsymbol{v} \mid \boldsymbol{z}^T\boldsymbol{v} > 0\}$, and the arc $\overset{\frown}{\mathsf{CBD}}$ denotes $\{\boldsymbol{v} \mid \boldsymbol{x}^T\boldsymbol{v} > 0\}$. The intersection of $\overset{\frown}{\mathsf{EAF}}$ and $\overset{\frown}{\mathsf{CBD}}$, i.e., the arc $\overset{\frown}{\mathsf{CBF}}$, represents $\{\boldsymbol{v} \mid \boldsymbol{z}^T\boldsymbol{v} > 0, \boldsymbol{x}^T\boldsymbol{v} > 0\}$. Notice that the angle of $\overset{\frown}{\mathsf{CBF}}$ equals $\pi - \theta$, where θ denotes the angle between $\boldsymbol{z}$ and $\boldsymbol{x}$. Therefore, ratio of the length of $\overset{\frown}{\mathsf{CBF}}$ to the perimeter of the circle equals to $\frac{\angle\mathrm{COF}}{2\pi} = \frac{\pi - \theta}{2\pi}$. Similarly, in the 3D case depicted in Fig. 5, the spherical lune ICHF denotes the intersection of the semi-sphere in the direction of $\overrightarrow{\mathsf{OA}}$ and the semi-sphere in the direction of $\overrightarrow{\mathsf{OB}}$. We can see that the area of the spherical lune ICHF is still proportional to the angle $\angle\mathsf{COF}$. Thus, we still have the result that the area of the spherical lune ICHF is $\frac{\pi - \theta}{2\pi}$ of the area of the whole sphere. The proof below, on the other hand, applies to arbitrary dimensions.

Proof. Due to symmetry, we know that the integral of Eq. (28) only depends on the angle between $\boldsymbol{x}$ and $\boldsymbol{z}$. Thus, without loss of generality, we let

$$\boldsymbol{x} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_d] = [0\ 0\ \cdots\ 0\ 1\ 0]^T,\ \boldsymbol{z} = [0\ 0\ \cdots\ 0\ \cos\theta\ \sin\theta]^T,$$

where

$$\theta = \arccos(\boldsymbol{x}^T\boldsymbol{z}) \in [0,\ \pi]. \tag{30}$$

Thus, for any $\boldsymbol{v} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_d]^T$ that makes $\boldsymbol{z}^T\boldsymbol{v} > 0$ and $\boldsymbol{x}^T\boldsymbol{v} > 0$, it only needs to satisfy

$$[\cos\theta\ \sin\theta] \begin{bmatrix} \boldsymbol{v}_{d-1} \\ \boldsymbol{v}_d \end{bmatrix} > 0, \quad [1\ 0] \begin{bmatrix} \boldsymbol{v}_{d-1} \\ \boldsymbol{v}_d \end{bmatrix} > 0. \tag{31}$$

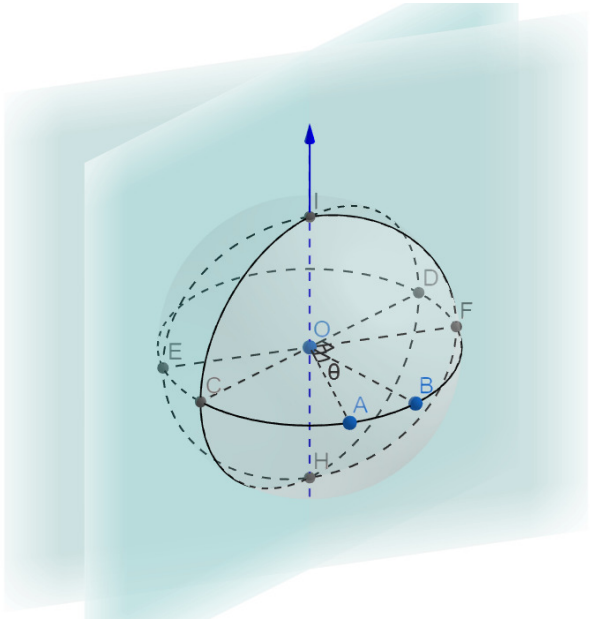

Figure 5. The area of the spherical lune ICHF is $\frac{\pi - \theta}{2\pi}$ of the area of the whole sphere.

We compute the spherical coordinates $\boldsymbol{\varphi_x} = [\varphi_1^{\boldsymbol{x}}\ \varphi_2^{\boldsymbol{x}}\ \cdots\ \varphi_{d-1}^{\boldsymbol{x}}]^T$ where $\varphi_1^{\boldsymbol{x}}, \cdots, \varphi_{d-2}^{\boldsymbol{x}} \in [0, \pi]$ and $\varphi_{d-1}^{\boldsymbol{x}} \in [0, 2\pi)$ with the convention that

$$
\begin{aligned}
\boldsymbol{x}_1 &= \cos(\varphi_1^{\boldsymbol{x}}), \\
\boldsymbol{x}_2 &= \sin(\varphi_1^{\boldsymbol{x}})\cos(\varphi_2^{\boldsymbol{x}}), \\
\boldsymbol{x}_3 &= \sin(\varphi_1^{\boldsymbol{x}})\sin(\varphi_2^{\boldsymbol{x}})\cos(\varphi_3^{\boldsymbol{x}}), \\
&\vdots \\
\boldsymbol{x}_{d-1} &= \sin(\varphi_1^{\boldsymbol{x}})\sin(\varphi_2^{\boldsymbol{x}})\cdots\sin(\varphi_{d-2}^{\boldsymbol{x}})\cos(\varphi_{d-1}^{\boldsymbol{x}}), \\
\boldsymbol{x}_d &= \sin(\varphi_1^{\boldsymbol{x}})\sin(\varphi_2^{\boldsymbol{x}})\cdots\sin(\varphi_{d-2}^{\boldsymbol{x}})\sin(\varphi_{d-1}^{\boldsymbol{x}}).
\end{aligned}
$$

Thus, we have $\boldsymbol{\varphi_x} = [\pi/2\ \pi/2\ \cdots\ \pi/2\ 0]^T$. Similarly, the spherical coordinates for $\boldsymbol{z}$ is $\boldsymbol{\varphi_z} = [\pi/2\ \pi/2\ \cdots\ \pi/2\ \theta]^T$. Let the spherical coordinates for $\boldsymbol{v}$ be $\boldsymbol{\varphi_v} = [\varphi_1^{\boldsymbol{v}}\ \varphi_2^{\boldsymbol{v}}\ \cdots\ \varphi_{d-1}^{\boldsymbol{v}}]^T$. Thus, Eq. (31) is equivalent to

$$
\sin(\varphi_1^{\boldsymbol{v}})\sin(\varphi_2^{\boldsymbol{v}})\cdots\sin(\varphi_{d-2}^{\boldsymbol{v}})\left(\cos\theta\cos(\varphi_{d-1}^{\boldsymbol{v}}) + \sin\theta\sin(\varphi_{d-1}^{\boldsymbol{v}})\right) > 0, \tag{32}
$$

$$
\sin(\varphi_1^{\boldsymbol{v}})\sin(\varphi_2^{\boldsymbol{v}})\cdots\sin(\varphi_{d-2}^{\boldsymbol{v}})\cos(\varphi_{d-1}^{\boldsymbol{v}}) > 0. \tag{33}
$$

Because $\varphi_1^{\boldsymbol{v}}, \cdots, \varphi_{d-2}^{\boldsymbol{v}} \in [0, \pi]$ (by the convention of spherical coordinates), we have

$$
\sin(\varphi_1^{\boldsymbol{v}})\sin(\varphi_2^{\boldsymbol{v}})\cdots\sin(\varphi_{d-2}^{\boldsymbol{v}}) \geq 0.
$$

Thus, for Eq. (32) and Eq. (33), we have

$$
\cos(\theta - \varphi_{d-1}^{\boldsymbol{v}}) > 0, \quad \cos(\varphi_{d-1}^{\boldsymbol{v}}) > 0,
$$

i.e., $\varphi_{d-1}^{\boldsymbol{v}} \in (-\pi/2, \pi/2] \setminus (\theta - \pi/2, \theta + \pi/2] \pmod{2\pi}$. We have

$$
\begin{aligned}
&\int_{S^{d-1}} \mathbf{1}_{\{\boldsymbol{z}^T\boldsymbol{v}>0,\, \boldsymbol{x}^T\boldsymbol{v}>0\}} d\lambda(\boldsymbol{v}) \\
&= \frac{\int_{(-\frac{\pi}{2}, \frac{\pi}{2}] \setminus (\theta - \frac{\pi}{2}, \theta+\frac{\pi}{2}]} \int_0^{\pi} \cdots \int_0^{\pi} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \cdots \sin(\varphi_{d-2})\, d\varphi_1\, d\varphi_2 \cdots d\varphi_{d-1}}{\int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \cdots \sin(\varphi_{d-2})\, d\varphi_1\, d\varphi_2 \cdots d\varphi_{d-1}} \\
&= \frac{\int_{(-\frac{\pi}{2}, \frac{\pi}{2}] \setminus (\theta - \frac{\pi}{2}, \theta+\frac{\pi}{2}]} A \cdots d\varphi_{d-1}}{\int_0^{2\pi} A \cdots d\varphi_{d-1}} \\
&\quad \text{(by defining } A := \int_0^{\pi} \cdots \int_0^{\pi} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \cdots \sin(\varphi_{d-2})\, d\varphi_1\, d\varphi_2) \\
&= \frac{\text{length of the interval } (-\frac{\pi}{2}, \frac{\pi}{2}] \setminus (\theta - \frac{\pi}{2}, \theta + \frac{\pi}{2}]}{2\pi} \\
&= \frac{\pi - \theta}{2\pi} \text{ (because } \theta \in [0, \pi] \text{ by Eq. (30))} \\
&= \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} \text{ (by Eq. (30)).}
\end{aligned}
$$

The result of this lemma thus follows. □

D.5. Limits of $|C^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}|/p$ when $p \to \infty$

We introduce some notions given by (Wainwright, 2015).

Glivenko-Cantelli class. Let $\mathcal{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $X_1^k = \{X_1, \ldots, X_k\}$ be a collection of *i.i.d.* samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable

$$\|\mathbb{P}_k - \mathbb{P}\|_{\mathcal{F}} := \sup_{\tilde{f} \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} \tilde{f}(X_k) - \mathbb{E}[\tilde{f}] \right|,$$

which measures the maximum deviation (over the class $\mathcal{F}$) between the sample average $\frac{1}{k}\sum_{i=1}^{k} \tilde{f}(X_i)$ and the population average $\mathbb{E}[\tilde{f}] = \mathbb{E}[\tilde{f}(X)]$. We say that $\mathcal{F}$ is a *Glivenko-Cantelli* class for $\mathbb{P}$ if $\|\mathbb{P}_k - \mathbb{P}\|_{\mathcal{F}}$ converges to zero in probability as $k \to \infty$.

Polynomial discrimination. A class $\mathcal{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$ if for each positive integer k and collection $X_1^k = \{X_1, \ldots, X_k\}$ of k points in $\mathcal{X}$, the set $\mathcal{F}(X_1^k)$ has cardinality upper bounded by

$$\text{card}(\mathcal{F}(X_1^k)) \leq (k+1)^{\nu}.$$

The following lemma is shown in Page 108 of (Wainwright, 2015).

Lemma 18. *Any bounded function class with polynomial discrimination is Glivenko-Cantelli.*

For our case, we care about the following value.

$$\left| \frac{|C^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} \right| = \left| \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{\boldsymbol{x}^T\mathbf{V}_0[j]>0, \boldsymbol{z}^T\mathbf{V}_0[j]>0\}} - \mathop{\mathbb{E}}_{\boldsymbol{v} \sim \lambda(\cdot)} [\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}>0, \boldsymbol{z}^T\boldsymbol{v}>0\}}] \right| \text{ (by Lemma 17).}$$

In the language of Glivenko-Cantelli class, the function class $\mathcal{F}$ consists of functions $\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}>0, \boldsymbol{z}^T\boldsymbol{v}>0\}}$ that map $\boldsymbol{v} \in S^{d-1}$ to 0 or 1, where every $\boldsymbol{x} \in S^{d-1}$ and $\boldsymbol{z} \in S^{d-1}$ corresponds to a distinct function in $\mathcal{F}$. According to Lemma 18, we need to calculate the order of the polynomial discrimination for this $\mathcal{F}$. Towards this end, we need the following lemma, which can be derived from the quantity $Q_{n,N}$ in (Wendel, 1962) (which is the quantity $Q_{d,k}$ in the following lemma).

Lemma 19. *Given $\boldsymbol{v}_1, \boldsymbol{v}_2, \quad , \boldsymbol{v}_k \ 2 \ S^{d \ 1}$, the number of different values (i.e., the cardinality) of the set $\left\{\left(\mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_1>0g}, \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_2>0g}, \quad , \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_k>0g}\right) \mid \boldsymbol{x} \ 2 \ S^{d \ 1}\right\}$ is at most $Q_{d,k}$, where*

$$Q_{d,k} := \begin{cases} 2\sum_{i=0}^{d \ 1} \binom{k \ 1}{i}, & \text{if } k > d, \\ 2^k, & \text{if } k \quad d. \end{cases}$$

Intuitively, Lemma 19 states the number of different regions that k hyper-planes through the origin (i.e., the kernel of the inner product with each $\boldsymbol{v}_i$) can cut $S^{d \ 1}$ into, because all $\boldsymbol{x}$ in one region corresponds to the same value of the tuple $\left(\mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_1>0g}, \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_2>0g}, \quad , \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_k>0g}\right)$. For example, in the 2D case (i.e., $d = 2$), k diameters of a circle can at most cut the whole circle into $2k$ (which equals to $Q_{2,k}$) parts. Notice that if some $\boldsymbol{v}_i$'s are parallel (thus some diameters are overlapped), then the total number of different parts can only be smaller. That is why Lemma 19 states that the cardinality is "at most" $Q_{d,k}$.

The following lemma shows that the cardinality in Lemma 19 is polynomial in k.

Lemma 20. *Recall the definition $Q_{d,k}$ in Lemma 19. For any integer $k \quad 1$ and $d \quad 2$, we must have $Q_{d,k} \quad (k + 1)^{d+1}$.*

Proof. When $k > d$, because $\binom{k \ 1}{i} \quad (k \quad 1)^{d \ 1}$ when $i \quad d \quad 1$, we have $Q_{d,k} = 2\sum_{i=0}^{d \ 1} \binom{k \ 1}{i} \quad 2d(k + 1)^{d \ 1} \quad (k + 1)^{d+1}$ (the last step uses $k \quad 1$ and $k > d$). When $k \quad d$, because $k \quad 1$, we have $Q_{d,k} = 2^k \quad (k + 1)^k \quad (k + 1)^d$. In summary, for any integer $k \quad 1$ and $d \quad 2$, the result $Q_{d,k} \quad (k + 1)^{d+1}$ always holds. □

We can now calculate the order of the polynomial discrimination for the function class F. Because

$$\begin{aligned} &\text{card}\left(\left\{\left(\mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_1>0,\boldsymbol{z}^T\boldsymbol{v}_1>0g}, \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_2>0,\boldsymbol{z}^T\boldsymbol{v}_2>0g}, \quad , \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_k>0,\boldsymbol{z}^T\boldsymbol{v}_k 0g}\right) \mid \boldsymbol{x} \ 2 \ S^{d \ 1}, \boldsymbol{z} \ 2 \ S^{d \ 1}\right\}\right) \\ &\text{card}\left(\left\{\left(\mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_1>0g}, \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_2>0g}, \quad , \mathbf{1}_{f\boldsymbol{x}^T\boldsymbol{v}_k>0g}\right) \mid \boldsymbol{x} \ 2 \ S^{d \ 1}\right\}\right) \\ &\quad \text{card}\left(\left\{\left(\mathbf{1}_{f\boldsymbol{z}^T\boldsymbol{v}_1>0g}, \mathbf{1}_{f\boldsymbol{z}^T\boldsymbol{v}_2>0g}, \quad , \mathbf{1}_{f\boldsymbol{z}^T\boldsymbol{v}_k>0g}\right) \mid \boldsymbol{z} \ 2 \ S^{d \ 1}\right\}\right), \end{aligned}$$

by Lemma 19 and Lemma 20, we know that

$$\text{card}(F \ (X_1^k)) \quad (Q_{d,k})^2 \quad (k + 1)^{2(d+1)}.$$

(Here X_1^k means $f\mathbf{V}_0[1], \quad , \mathbf{V}_0[k]g$.)

Thus, F has polynomial discrimination with order at most $2(d + 1)$. Notice that all functions in F is bounded because their outputs can only be 0 or 1. Therefore, by Lemma 18 (i.e., any bounded function class with polynomial discrimination is Glivenko-Cantelli), we know that F is Glivenko-Cantelli. In other words, we have shown the following lemma.

Lemma 21.

$$\sup_{\boldsymbol{x},\boldsymbol{z}2S^{d \ 1}} \left| \frac{jC^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}j}{p} \quad \frac{\pi \quad \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} \right| \xrightarrow{P} 0, \ as \ p \ ! \ 1. \tag{34}$$

E. Proof of Lemma 7 (H has full row-rank with high probability as $p \ ! \ 1$)

In this section, we prove Lemma 7, i.e., the matrix $\mathbf{H}$ has full row-rank with high probability when $p \ ! \ 1$. We first introduce two useful lemmas as follows.

The following lemma states that, given $\mathbf{X}$ (that satisfies Assumption 3) and $k \ 2 \ f1, 2, \quad , ng$, there always exists a vector $\boldsymbol{v} \ 2 \ S^{d \ 1}$ that is only orthogonal to one training input $\mathbf{X}_k$ but not orthogonal to other training inputs $\mathbf{X}_i$ for all $i \ 6= \ k$. An intuitive explanation is that, because no training inputs are parallel (as stated in Assumption 3), the total set of vectors that are orthogonal to at least two training inputs is too small. That gives us many options to pick such a vector $\boldsymbol{v}$ that is only orthogonal to one input but not others.

Lemma 22. *For all $k \ 2 \ f1, 2, \quad , ng$ we have*

$$T_k := \left\{\boldsymbol{v} \ 2 \ S^{d \ 1} \ j \ \boldsymbol{v}^T\mathbf{X}_k = 0, \boldsymbol{v}^T\mathbf{X}_i \ 6= 0, \text{for all } i \ 2 \ f1, 2, \quad , ng \ n \ fkg\right\} \ 6= \ ?.$$

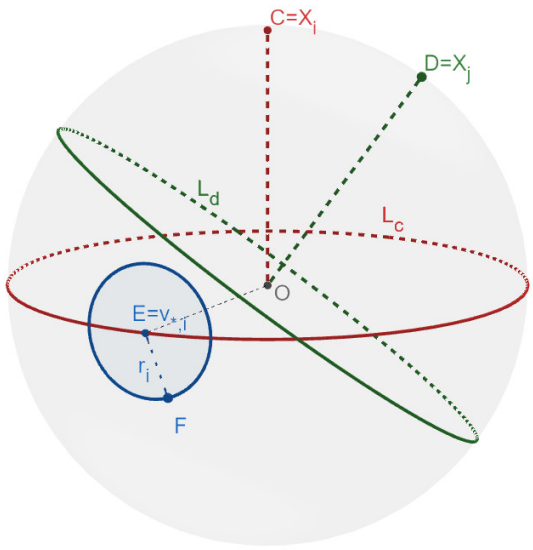

Figure 6. Geometric interpretation of $B_{\boldsymbol{v}_{,i,+}}^{r_i,\mathbf{X}_i}$ and $B_{\boldsymbol{v}_{,i,}}^{r_i,\mathbf{X}_i}$ on a sphere (i.e., S^2).

Proof. We have

$$T_k = S^{d\ 1} \setminus \ker(\mathbf{X}_k) \cap \left(\bigcup_{i2f1,2,\ ,ngnfkg} \ker(\mathbf{X}_i) \right)$$
$$= S^{d\ 1} \setminus \ker(\mathbf{X}_k) \cap \left(\bigcup_{i2f1,2,\ ,ngnfkg} \left(S^{d\ 1} \setminus \ker(\mathbf{X}_k) \setminus \ker(\mathbf{X}_i)\right) \right).$$

Because

$$\begin{aligned} &\dim(S^{d\ 1} \setminus \ker(\mathbf{X}_k)) = d \quad 2, \\ &\dim(S^{d\ 1} \setminus \ker(\mathbf{X}_k) \setminus \ker(\mathbf{X}_i)) = d \quad 3 \text{ for all } i\ 2\ f1,2, \quad ,ng\,n\,fkg \text{ (because } \mathbf{X}_i\ ,\ \mathbf{X}_k), \end{aligned} \tag{35}$$

we have

$$\begin{aligned} &\lambda_{d\ 2}(S^{d\ 1} \setminus \ker(\mathbf{X}_k)) = \lambda_{d\ 2}(S^{d\ 2}) > 0, \\ &\lambda_{d\ 2}\left(S^{d\ 1} \setminus \ker(\mathbf{X}_k) \setminus \ker(\mathbf{X}_i)\right) = 0 \text{ for all } i\ 2\ f1,2, \quad ,ng\,n\,fkg. \end{aligned} \tag{36}$$

(When $d = 2$, the set in Eq. (35) is not defined. Nonetheless, Eq. (36) still holds when $d = 2$.) Thus, we have

$$\begin{aligned} \lambda_{d\ 2}(T_k) = \lambda_{d\ 2}\left(S^{d\ 1} \setminus \ker(\mathbf{X}_k)\right) \quad & \lambda_{d\ 2}\left(\bigcup_{i2f1,2,\ ,ngnfkg} \left(S^{d\ 1} \setminus \ker(\mathbf{X}_k) \setminus \ker(\mathbf{X}_i)\right) \right) \\ \lambda_{d\ 2}\left(S^{d\ 1} \setminus \ker(\mathbf{X}_k)\right) \quad & \sum_{i2f1,2,\ ,ngnfkg} \lambda_{d\ 2}\left(S^{d\ 1} \setminus \ker(\mathbf{X}_k) \setminus \ker(\mathbf{X}_i)\right) \\ = \lambda_{d\ 2}(S^{d\ 2}) & \\ > 0. & \end{aligned}$$

Therefore, T_k ~~6~~ ?. □

The following lemma plays an important role in answering whether $\mathbf{H}$ has full row-rank. Further, it is also closely related to our estimation on the $\min \mathrm{eig}(\mathbf{H}\mathbf{H}^T)$ later in Appendix F.

Lemma 23. *Consider any $i \in \{1,2,\ldots,n\}$. For any $\boldsymbol{v}_{\setminus i} \in S^{d-1}$ satisfying $\boldsymbol{v}_{\setminus i}^T \mathbf{X}_i = 0$, we define*

$$r_i := \min_{j\in\{1,2,\ldots,n\}\setminus\{i\}} \left|\boldsymbol{v}_{\setminus i}^T \mathbf{X}_j\right|. \tag{37}$$

If there exist $k, l \in \{1,\ldots,p\}$ such that

$$\frac{\mathbf{V}_0[k]}{\|\mathbf{V}_0[k]\|_2} \in B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},+}, \quad \frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2} \in B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},-}, \tag{38}$$

then we must have

$$\mathbf{H}_j[k] = \mathbf{H}_j[l], \text{ for all } j \in \{1,2,\ldots,n\}\setminus\{i\}, \tag{39}$$

$$\mathbf{H}_i[k] = \mathbf{X}_i^T, \tag{40}$$

$$\mathbf{H}_i[l] = \mathbf{0}. \tag{41}$$

(Notice that Eq. (38) implies $r_i > 0$.)

Remark 6. We first give an intuitive geometric interpretation of Lemma 23. In Fig. 6, the sphere centered at O denotes S^{d-1}, the vector $\overrightarrow{\mathsf{OC}}$ denotes $\mathbf{X}_i$, the vector $\overrightarrow{\mathsf{OD}}$ denotes one of other $\mathbf{X}_j$'s, the vector $\overrightarrow{\mathsf{OE}}$ denotes $\boldsymbol{v}_{\setminus i}$, which is perpendicular to $\mathbf{X}_i$ (i.e., $\mathbf{X}_i^T \boldsymbol{v}_{\setminus i} = 0$). The upper half of the cap E denotes $B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},+}$, the lower half of the cap E denotes $B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},-}$. The great circle L_c cuts the sphere into two semi-spheres. The semi-sphere in the direction of $\overrightarrow{\mathsf{OC}}$ corresponds to all vectors $\boldsymbol{v}$ on the sphere that have positive inner product with $\mathbf{X}_i$ (i.e., $\boldsymbol{v}_{\setminus i}^T \mathbf{X}_i > 0$), and the semi-sphere in the opposite direction of $\overrightarrow{\mathsf{OC}}$ corresponds to all vectors $\boldsymbol{v}$ on the sphere that have negative inner product with $\mathbf{X}_i$ (i.e., $\boldsymbol{v}^T \mathbf{X}_i < 0$). The great circle L_d is similar to the great circle L_c, but is perpendicular to the direction $\overrightarrow{\mathsf{OD}}$ (i.e., $\mathbf{X}_j$). By choosing the radius of the cap E in Eq. (37), we can ensure that all great circles that are perpendicular to other $\mathbf{X}_j$'s do not pass the cap E. In other words, for the two semi-spheres cut by the great circle perpendicular to $\mathbf{X}_j$, $j \neq i$, the cap E must be contained in one of them. Therefore, vectors on the upper half of the cap E and the vectors on the lower half of the cap E must have the same sign when calculating the inner product with all $\mathbf{X}_j$'s, for all $j \neq i$.

Now, let us consider the meaning of Eq. (38) in this geometric setup depicted in Fig. 6. The expression $\frac{\mathbf{V}_0[k]}{\|\mathbf{V}_0[k]\|_2} \in B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},+}$ means that the direction of $\mathbf{V}_0[k]$ is in the upper half of the cap E. By the definition of $\mathbf{H}_i = \boldsymbol{h}_{\mathbf{V}_0,\mathbf{X}_i}$ in Eq. (1), we must then have $\mathbf{H}_i[k] = \mathbf{X}_i^T$. Similarly, the expression $\frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2} \in B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{\setminus i},-}$ means that the direction of $\mathbf{V}_0[l]$ is in the lower half of the cap E, and thus $\mathbf{H}_i[l] = \mathbf{0}$. Then, based on the discussions in the previous paragraph, we know that $\mathbf{V}_0[k]$ and $\mathbf{V}_0[l]$ has the same activation pattern under ReLU for all $\mathbf{X}_j$'s that $j \neq i$, which implies that $\mathbf{H}_j[k] = \mathbf{H}_j[l]$. These are precisely the conclusions in Eqs. (39)(40)(41).

Later in Appendix F, Lemma 23 plays an important role in estimating $\min_{\boldsymbol{a}\in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2$. To see this, let a_j denotes the j-th element of $\boldsymbol{a}$. By Eq. (39), we have $\sum_{j\in\{1,2,\ldots,n\}\setminus\{i\}} ((\mathbf{H}^T a_j)[k] - (\mathbf{H}^T a_j)[l]) = \mathbf{0}$. By Eq. (40) and Eq. (41), we have $(\mathbf{H}^T a_i)[k] - (\mathbf{H}^T a_i)[l] = \mathbf{X}_i$. Combining them together, we have $(\mathbf{H}^T\boldsymbol{a})[k] - (\mathbf{H}^T\boldsymbol{a})[l] = a_i\mathbf{X}_i$. As long as a_i is not zero, then regardless values of other elements in $\boldsymbol{a}$, we always obtain that $\mathbf{H}^T\boldsymbol{a}$ is a non-zero vector. This implies $\|\mathbf{H}^T\boldsymbol{a}\|_2 > 0$, which will be useful for estimating $\min \mathrm{eig}(\mathbf{H}\mathbf{H}^T)/p$ in Appendix F.

Proof. By the definition of r_i, we have

$$\left|\boldsymbol{v}_{\setminus i}^T \mathbf{X}_j\right| \geq r_i > 0, \text{ for all } j \in \{1,2,\ldots,n\}\setminus\{i\}. \tag{42}$$

For any $j \in \{1,2,\ldots,n\}\setminus\{i\}$ and any $\boldsymbol{v} \in B^{r_i}_{\boldsymbol{v}_{\setminus i}}$, since $\|\boldsymbol{v} - \boldsymbol{v}_{\setminus i}\|_2 < r_i$, we have

$$\begin{aligned}
(\boldsymbol{v}^T\mathbf{X}_j)(\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j) &= \left((\boldsymbol{v} - \boldsymbol{v}_{\setminus i})^T\mathbf{X}_j + \boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right)(\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j) \\
&= (\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j)^2 + (\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j)\left((\boldsymbol{v} - \boldsymbol{v}_{\setminus i})^T\mathbf{X}_j\right) \\
&\geq (\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j)^2 - \left|\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right| \left|(\boldsymbol{v} - \boldsymbol{v}_{\setminus i})^T\mathbf{X}_j\right| \\
&\geq (\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j)^2 - \left|\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right| \|\boldsymbol{v} - \boldsymbol{v}_{\setminus i}\|_2 \|\mathbf{X}_j\|_2 \\
&> (\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j)^2 - \left|\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right| r_i \text{ (by Eq. (21))} \\
&= \left|\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right|\left(\left|\boldsymbol{v}_{\setminus i}^T\mathbf{X}_j\right| - r_i\right) \\
&\geq 0 \text{ (by Eq. (42)).}
\end{aligned}$$

Thus, for any $\boldsymbol{v}_1 \in B^{r_i}_{\boldsymbol{v}_{,i}}$, $\boldsymbol{v}_2 \in B^{r_i}_{\boldsymbol{v}_{,i}}$, $j \in \{1,2,\ldots,n\} \setminus \{i\}$, we have $(\boldsymbol{v}_1^T\mathbf{X}_j)(\boldsymbol{v}^T_{,i}\mathbf{X}_j) > 0$ and $(\boldsymbol{v}_2^T\mathbf{X}_j)(\boldsymbol{v}^T_{,i}\mathbf{X}_j) > 0$. It implies that

$$\operatorname{sign}(\boldsymbol{v}_1^T\mathbf{X}_j) = \operatorname{sign}(\boldsymbol{v}^T_{,i}\mathbf{X}_j) = \operatorname{sign}(\boldsymbol{v}_2^T\mathbf{X}_j). \tag{43}$$

By Eq. (38), we know that both $\mathbf{V}_0[k]$ and $\mathbf{V}_0[l]$ are in $B^{r_i}_{\boldsymbol{v}_{,i}}$. Applying Eq. (43), we have

$$\operatorname{sign}(\mathbf{X}_j^T\mathbf{V}_0[k]) = \operatorname{sign}(\mathbf{X}_j^T\mathbf{V}_0[l]), \text{ for all } j \in \{1,2,\ldots,n\} \setminus \{i\}.$$

Thus, by Eq. (1), we have

$$\mathbf{H}_j[k] = \mathbf{1}_{\{\mathbf{X}_j^T\mathbf{V}_0[k]>0\}}\mathbf{X}_j^T = \mathbf{1}_{\{\mathbf{X}_j^T\mathbf{V}_0[l]>0\}}\mathbf{X}_j^T = \mathbf{H}_j[l], \text{ for all } j \in \{1,2,\ldots,n\} \setminus \{i\}.$$

By Eq. (22), we have

$$\mathbf{X}_i^T\mathbf{V}_0[k] > 0, \ \mathbf{X}_i^T\mathbf{V}_0[l] < 0.$$

Thus, by Eq. (1), we have

$$\mathbf{H}_i[k] = \mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[k]>0\}}\mathbf{X}_i^T = \mathbf{X}_i^T, \quad \mathbf{H}_i[l] = \mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[l]>0\}}\mathbf{X}_i^T = \mathbf{0}.$$

□

Now, we are ready to prove Lemma 7.

Proof. We prove by contradiction. Suppose on the contrary that with some nonzero probability, the design matrix is not full row-rank as $p \geq 1$. Note that when the design matrix is not full row-rank, there exists a set of indices $I \subseteq \{1,\ldots,n\}$ such that

$$\sum_{i\in I} b_i\mathbf{H}_i = \mathbf{0}, \ b_i \neq 0 \text{ for all } i \in I. \tag{44}$$

The proof will be finished by two steps: 1) find an event J that happens almost surely when $p \geq 1$; 2) prove this event J contradicts Eq. (44).

Step 1:

Consider each $i \in \{1,2,\ldots,n\}$. By Lemma 22, we know that there exists a $\boldsymbol{v}_{,i} \in S^{d-1}$ such that

$$\boldsymbol{v}^T_{,i}\mathbf{X}_i = 0, \ \boldsymbol{v}^T_{,i}\mathbf{X}_j \neq 0, \text{ for all } j \in \{1,2,\ldots,n\} \setminus \{i\}. \tag{45}$$

Define

$$r_i = \min_{j\in\{1,2,\ldots,n\}\setminus\{i\}} \left|\boldsymbol{v}^T_{,i}\mathbf{X}_j\right| > 0. \tag{46}$$

For all $i = 1,2,\ldots,n$, we define several events as follows.

$$\begin{aligned}
J_i &:= \left\{A_{\mathbf{V}_0} \setminus B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{,i},+} \neq \emptyset, \ A_{\mathbf{V}_0} \setminus B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{,i},-} \neq \emptyset\right\},\\
J_{i,+} &= \left\{A_{\mathbf{V}_0} \setminus B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{,i},+} \neq \emptyset\right\},\\
J_{i,-} &= \left\{A_{\mathbf{V}_0} \setminus B^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{,i},-} \neq \emptyset\right\},\\
J &:= \bigcap_{i=1}^n J_i.
\end{aligned}$$

(Recall the geometric interpretation in Remark 6. The events $J_{i,+}$ and $J_{i,-}$ mean that there exists $\mathbf{V}_0[j]/\|\mathbf{V}_0[j]\|_2$ in the upper half and the lower half of the cap E, respectively. The event $J_i = J_{i,+} \setminus J_{i,-}$ means that there exist $\mathbf{V}_0[j]/\|\mathbf{V}_0[j]\|_2$

in both halves of the cap E. Finally, the event J occurs when J_i occurs for all i, although the vector $\mathbf{V}_0[j]/k\mathbf{V}_0[j]k_2$ that falls into the two halves may differ across i. As we will show later, whenever the event J occurs, the matrix $\mathbf{H}$ will have the full row-rank, which is why we are interesting in the probability of the event J.)

Those definitions implies that

$$J_i^c = J_{i,+}^c [J_{i,}^c \quad \text{for all } i = 1, 2, \quad , n, \tag{47}$$

$$J^c = \bigcup_{i=1}^{n} J_i^c. \tag{48}$$

Thus, we have

$$\begin{aligned} \Pr_{\mathbf{V}_0}[J] = 1 \quad & \Pr_{\mathbf{V}_0}[J^c] \\ 1 \quad & \sum_{i=1}^{n} \Pr_{\mathbf{V}_0}[J_i^c] \text{ (by Eq. (48) and the union bound).} \end{aligned} \tag{49}$$

For a fixed i, recall that by Eq. (46), we have $r_i > 0$. Because $B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,+}$ and $B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,}$ are two halves of $B_{\boldsymbol{v}}^{r_i}{}_{,i}$, we have

$$\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,+}) = \lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,}) = \frac{1}{2}\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i}{}_{,i}). \tag{50}$$

Therefore, we have

$$\begin{aligned} \Pr_{\mathbf{V}_0}[J_i^c] \quad & \Pr_{\mathbf{V}_0}[J_{i,+}^c] + \Pr_{\mathbf{V}_0}[J_{i,}^c \ ] \text{ (by Eq. (47) and the union bound)} \\ & = \left(1 \quad \frac{\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,+})}{\lambda_{d\ \ 1}(S^{d\ \ 1})}\right)^p + \left(1 \quad \frac{\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,})}{\lambda_{d\ \ 1}(S^{d\ \ 1})}\right)^p \\ & \quad \text{(all } \mathbf{V}_0[i]\text{'s are independent and Assumption 1)} \\ & = 2\left(1 \quad \frac{\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_i}{}_{,i})}{2\lambda_{d\ \ 1}(S^{d\ \ 1})}\right)^p \text{ (by Eq. (50)).} \end{aligned}$$

Notice that r_i is determined only by $\mathbf{X}$, and is independent of $\mathbf{V}_0$ and p. Therefore, we have

$$\lim_{p\ !\ 1} \Pr_{\mathbf{V}_0}[J_i^c] = 0. \tag{51}$$

Plugging Eq. (51) into Eq. (49), we have

$$\lim_{p\ !\ 1} \Pr_{\mathbf{V}_0}[J] = 1 \text{ (because } n \text{ is finite).}$$

Step 2:

To complete the proof, it remains to show that the event J contradicts Eq. (44). Towards this end, we assume the event J happens. By Eq. (44), we can pick one $i \ 2 \ I$. Further, by the definition of J, there exists r_i such that $A_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,+} \ 6= \ ?$ and $A_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,} \ 6= \ ?$. In other words, there must exist $k, l \ 2 \ f1, \quad , pg$ such that

$$\frac{\mathbf{V}_0[k]}{k\mathbf{V}_0[k]k_2} \ 2 \ B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,+}, \quad \frac{\mathbf{V}_0[l]}{k\mathbf{V}_0[l]k_2} \ 2 \ B_{\boldsymbol{v}}^{r_i,\mathbf{X}_i}{}_{,i,} \ .$$

By Lemma 23, we have

$$\mathbf{H}_j[k] = \mathbf{H}_j[l], \text{ for all } j \ 2 \ f1, 2, \quad , ng \, n \, fig, \tag{52}$$

$$\mathbf{H}_i[k] = \mathbf{X}_i^T, \quad \mathbf{H}_i[l] = \mathbf{0}. \tag{53}$$

We now show that $\mathbf{H}$ restricted to the columns corresponding to k and l cannot be linearly dependent. Specifically, we have

$$\begin{aligned}\sum_{j \neq i,l} b_j \mathbf{H}_j[k] &= b_i \mathbf{H}_i[k] + \sum_{j \neq i,l, j \in [n]} b_j \mathbf{H}_j[k] \text{ (as we have picked } i \neq l\text{)} \\ &= b_i \mathbf{H}_i[k] \quad b_j \mathbf{H}_i[l] + \sum_{j \neq i,l} b_j \mathbf{H}_j[l] \text{ (by Eq. (52))} \\ &= b_i \mathbf{X}_i^T + \sum_{j \neq i,l} b_j \mathbf{H}_j[l] \text{ (by Eq. (53))} \\ &\neq \sum_{j \neq i,l} b_j \mathbf{H}_j[l] \text{ (because } b_i \neq 0\text{).}\end{aligned}$$

This contradicts the assumption Eq. (44) that

$$\sum_{j \neq i,l} b_j \mathbf{H}_j[k] = \sum_{j \neq i,l} b_j \mathbf{H}_j[l] = \mathbf{0}.$$

The result thus follows. □

F. Proof of Proposition 4 (the upper bound of the variance)

The following lemma shows the relationship between the variance term and $\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$.

Lemma 24.

$$|h_{\mathbf{V}_0,\boldsymbol{x}} \mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon| \leq \frac{\sqrt{p}\,\bar{p}\|\epsilon\|_2}{\sqrt{\min \operatorname{eig}(\mathbf{H}\mathbf{H}^T)}}.$$

Proof. We have

$$\|\mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon\|_2 = \sqrt{(\mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon)^T \mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon} = \sqrt{\epsilon^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon} \leq \frac{\|\epsilon\|_2}{\sqrt{\min \operatorname{eig}(\mathbf{H}\mathbf{H}^T)}}. \tag{54}$$

Thus, we have

$$\begin{aligned}&|h_{\mathbf{V}_0,\boldsymbol{x}} \mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon| \\ =&\|h_{\mathbf{V}_0,\boldsymbol{x}} \mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon\|_2 \text{ (ℓ_2-norm of a number equals to its absolute value)} \\ \leq&\|h_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \cdot \|\mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1} \epsilon\|_2 \text{ (by Lemma 12)} \\ \leq&\frac{\sqrt{p}\,\bar{p}\|\epsilon\|_2}{\sqrt{\min \operatorname{eig}(\mathbf{H}\mathbf{H}^T)}} \text{ (by Lemma 11 and Eq. (54)).}\end{aligned}$$

□

The following lemma shows our estimation on $\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$.

Lemma 25. *For any* $n \geq 2$, $m \geq \left[1, \frac{\ln n}{\ln \frac{n}{2}}\right]$, $d \geq n^4$, *if* $p \geq 6 J_m(n,d) \ln\left(4n^{1+\frac{1}{m}}\right)$, *we must have*

$$\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\frac{\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)}{p} \geq \frac{1}{J_m(n,d)n}\right\} \geq 1 - \frac{2}{n^{\frac{m}{2}}}.$$

Proposition 4 directly follows from Lemma 25 and Lemma 24. [8]

In rest of this section, we will show how to prove Lemma 25. The following lemma shows that, to estimate $\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$, it is equivalent to estimate $\min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T \boldsymbol{a}\|_2^2/p$.

[8] We can see that the key part during the proof of Proposition 4 is to estimate $\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$. Lemma 25 shows a lower bound of $\min \operatorname{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$ which is almost $\Omega(n^{-2d})$ when p is large. However, our estimation of this value may be loose. We will show a upper bound which is $O(n^{-\frac{1}{d-1}})$ (see Appendix G).

Lemma 26.

$$\min \mathrm{eig}\left(\mathbf{H}\mathbf{H}^T\right) = \min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2.$$

Proof. Do the singular value decomposition (SVD) of $\mathbf{H}^T$ as $\mathbf{H}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T$, where

$$\mathbf{\Sigma} \in \mathbb{R}^{(dp)\times n} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n).$$

By properties of singular values, we have

$$\min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2 = \min_{i \in \{1,2,\ldots,n\}} \sigma_i^2.$$

We also have

$$\begin{aligned}\mathbf{H}\mathbf{H}^T &= \mathbf{W}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{W}^T \\ &= \mathbf{W}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{W}^T \text{ (because } \mathbf{U}^T\mathbf{U} = \mathbf{I}) \\ &= \mathbf{W}\mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)\mathbf{W}^T.\end{aligned}$$

This equation is indeed the eigenvalue decomposition of $\mathbf{H}\mathbf{H}^T$, which implies that its eigenvalues are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Thus, we have

$$\min \mathrm{eig}\left(\mathbf{H}\mathbf{H}^T\right) = \min_{i \in \{1,2,\ldots,n\}} \sigma_i^2 = \min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2.$$

□

Therefore, to finish the proof of Proposition 4, it only remains to estimate $\min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2$.

By Lemma 7 and its proof in Appendix E, we have already shown that $\mathbf{H}^T\boldsymbol{a}$ is not likely to be zero (i.e. $\min_{\boldsymbol{a} \in S^{n-1}} \|\mathbf{H}^T\boldsymbol{a}\|_2^2 > 0$) when $p \gg 1$. Here, we basically use the similar method as in Appendix E, but with more precise quantification.

Recall the definitions in Eqs. (21)(22)(23). For any $i \in \{1, 2, \ldots, n\}$, we choose one

$$\boldsymbol{v}_{\cdot,i} \in S^{d-1} \text{ independently of } \mathbf{X}_j,\ j \neq i, \text{ such that } \boldsymbol{v}_{\cdot,i}^T\mathbf{X}_i = 0. \tag{55}$$

(Note that here, unlike in Eq. (45), we do not require $\boldsymbol{v}_{\cdot,i}^T\mathbf{X}_j \neq 0$ for all $j \neq i$. This is important as we would like $\mathbf{X}_j$ to be independent of $\boldsymbol{v}_{\cdot,i}$ for all $j \neq i$.) Further, for any $0 < r_0 < 1$, we define

$$c_{r_0}^i := \min\left\{\left|A_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}_{\cdot,i},+}^{r_0,\mathbf{X}_i}\right|, \left|A_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}_{\cdot,i},-}^{r_0,\mathbf{X}_i}\right|\right\}. \tag{56}$$

Then, we define

$$r_i := \min_{j \in \{1,2,\ldots,n\} \setminus \{i\}} \left|\boldsymbol{v}_{\cdot,i}^T\mathbf{X}_j\right|, \tag{57}$$

$$\hat{r} := \min_{i \in \{1,2,\ldots,n\}} r_i. \tag{58}$$

(Note that here r_i or $\hat{r}$ may be zero. Later we will show that they can be lower bounded with high probability.) Define

$$D_{\mathbf{X}} := \frac{\lambda_{d-1}(B^{\hat{r}})}{8n\lambda_{d-1}(S^{d-1})}. \tag{59}$$

Similar to Remark 6, these definitions have their geometric interpretation in Fig. 6. The value $c_{r_0}^i$ denotes the number of distinct pairs $\left(\frac{\mathbf{V}_0[k]}{\|\mathbf{V}_0[k]\|_2}, \frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2}\right)$[9] such that $\frac{\mathbf{V}_0[k]}{\|\mathbf{V}_0[k]\|_2}$ is in the upper half of the cap $\mathbb{E}$, and $\frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2}$ is in the lower half of the cap $\mathbb{E}$. The quantities r_0, r_i, and $\hat{r}$ can all be used as the radius of the cap $\mathbb{E}$. The ratio $D_{\mathbf{X}}$ is proportional to the area of the cap $\mathbb{E}$ with radius $\hat{r}$ (or equivalently, the probability that the normalized $\mathbf{V}_0[j]$ falls in the cap $\mathbb{E}$).

The following lemma gives an estimation on $\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$ when $\mathbf{X}$ is given. We put its proof in Appendix F.1.

[9] Here, "distinct" means that any normalized version of $\mathbf{V}_0[j]$ can appear at most in one pair.

Lemma 27. *Given* $\mathbf{X}$*, we have*

$$\Pr_{\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad pD_{\mathbf{X}}, \text{ for all } \boldsymbol{a} \in S^{n-1}\right\} \geq 1 - 4ne^{-npD_{\mathbf{X}}/6}.$$

Notice that $D_{\mathbf{X}}$ only depends on $\mathbf{X}$ and it may even be zero if $\hat{\lambda}$ is zero. However, after we introduce the randomness of $\mathbf{X}$, we can show that $\hat{\lambda}$ is lower bounded with high probability. We can then obtain the following lemma. We put its proof in Appendix F.2.

Define

$$C_d := \frac{2^{p-\overline{2}}}{B(\frac{d-1}{2}, \frac{1}{2})}, \tag{60}$$

$$D(n,d,\delta) := \frac{1}{16n} I_{\frac{\delta^2}{n^4 C_d^2}}\left(1 - \frac{\delta^2}{4n^4 C_d^2}\right)\left(\frac{d-1}{2}, \frac{1}{2}\right). \tag{61}$$

Lemma 28. *For any* $\delta \in \left(0, \frac{2}{\pi}\right]$*, we have*

$$\Pr_{\mathbf{X},\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1}\right\} \geq 1 - 4ne^{-npD(n,d,\delta)/6} - \delta.$$

Notice that Lemma 28 is already very close to Lemma 25, and we put the final steps of the proof of Lemma 25 in Appendix F.3.

F.1. Proof of Lemma 27

Proof. Define events as follows.

$$J := \left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad pD_{\mathbf{X}}, \text{ for all } \boldsymbol{a} \in S^{n-1}\right\},$$

$$J_i := \left\{\text{there exists } \boldsymbol{a} \in S^{n-1} \text{ that } i \in \arg\max_{j \in \{1,2,\ldots,n\}} |a_j|, \text{ and } k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad pD_{\mathbf{X}}\right\},$$

$$K_i := \left\{c^i_{r_i} \quad 2npD_{\mathbf{X}}\right\}, \text{ for } i = 1, 2, \ldots, n.$$

Those definitions directly imply that

$$J^c = \bigcup_{i=1}^n J_i. \tag{62}$$

Step 1: prove $J_i \quad K_i$

To show $J_i \quad K_i$, we only need to prove that J_i implies K_i. To that end, it suffices to show $k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad \frac{c^i_{r_i}}{2n}$ for the vector $\boldsymbol{a}$ defined in J_i. Because $i \in \arg\max_{j=1}^n |a_j|$ and $k\boldsymbol{a}k_2 = 1$, we have

$$|a_i| \quad \frac{1}{\sqrt{n}}. \tag{63}$$

By Eq. (56), we can construct $c^i_{r_i}$ pairs (k_j, l_j) for $j = 1, 2, \ldots, c^i_{r_i}$ (all k_j's are different and all l_j's are different), such that

$$\frac{\mathbf{V}_0[k_j]}{k\mathbf{V}_0[k_j]k_2} \in B^{r_i,\mathbf{x}_i}_{\boldsymbol{v}_{,i},+}, \quad \frac{\mathbf{V}_0[l_j]}{k\mathbf{V}_0[l_j]k_2} \in B^{r_i,\mathbf{x}_i}_{\boldsymbol{v}_{,i},-}.$$

Thus, we have

$$\begin{aligned}(\mathbf{H}^T\boldsymbol{a})[k_j] \quad (\mathbf{H}^T\boldsymbol{a})[l_j] &= \sum_{k=1}^n a_k\left(\mathbf{H}_k[k_j] \quad \mathbf{H}_k[l_j]\right) \\ &= a_i\left(\mathbf{H}_i[k_j] \quad \mathbf{H}_i[l_j]\right) + \sum_{k \in \{1,2,\ldots,n\} \setminus \{i\}} a_k\left(\mathbf{H}_k[k_j] \quad \mathbf{H}_k[l_j]\right) \\ &= a_i\mathbf{X}_i \text{ (by Lemma 23).}\end{aligned}$$

We then have

$$k(\mathbf{H}^T\boldsymbol{a})[k_j]k_2^2 + k(\mathbf{H}^T\boldsymbol{a})[l_j]k_2^2 \quad \frac{1}{2}ka_i\mathbf{X}_ik_2^2 \text{ (by Lemma 13)}$$
$$\frac{1}{2n} \text{ (by Eq. (63)).}$$

Further, we have

$$k\mathbf{H}^T\boldsymbol{a}k_2^2 = \sum_{j=1}^{p} k(\mathbf{H}^T\boldsymbol{a})[j]k_2^2 \quad \sum_{j=1}^{c_{r_i}^i} k(\mathbf{H}^T\boldsymbol{a})[k_j]k_2^2 + k(\mathbf{H}^T\boldsymbol{a})[l_j]k_2^2 = \frac{c_{r_i}^i}{2n}. \tag{64}$$

Clearly, if the event J_i occurs, then $k\mathbf{H}\boldsymbol{a}k_2^2 \quad pD_{\mathbf{X}}$. Combining with Eq. (64), we then have $c_{r_i}^i \quad 2npD_{\mathbf{X}}$. In other words, the event K_i must occur. Hence, we have shown that $J_i \quad K_i$.

Step 2: estimate the probability of K_i

For all $j \ 2 \ f1, 2, \quad , pg$, because $\mathbf{V}_0[j]$ is uniformly distributed in all directions, for any fixed $0 \quad r_0 \quad 1$, we have

$$\Pr_{\mathbf{V}_0}\left\{\frac{\mathbf{V}_0[j]}{k\mathbf{V}_0[j]k_2} \ 2 \ B_{\boldsymbol{v}\ ,i,+}^{r_0,\mathbf{x}_i} \middle| \ i\right\} = \frac{\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{2\lambda_{d\ \ 1}(S^{d\ \ 1})}.$$

Thus, $jA_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}\ ,i,+}^{r_0,\mathbf{x}_i} j$ follows the distribution $\mathsf{Bino}\left(p, \ \frac{\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{2\lambda_{d\ \ 1}(S^{d\ \ 1})}\right)$ given i and $\mathbf{X}$. By Lemma 14 (with $\delta = \frac{1}{2}$), we have

$$\Pr_{\mathbf{V}_0}\left\{jA_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}\ ,i,+}^{r_0,\mathbf{x}_i} j < \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{4\lambda_{d\ \ 1}(S^{d\ \ 1})} \middle| \ i\right\} \quad 2\exp\left(\ \ \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{48\lambda_{d\ \ 1}(S^{d\ \ 1})}\right). \tag{65}$$

Similarly, we have

$$\Pr_{\mathbf{V}_0}\left\{jA_{\mathbf{V}_0} \setminus B_{\boldsymbol{v}\ ,i,}^{r_0,\mathbf{x}_i} j < \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{4\lambda_{d\ \ 1}(S^{d\ \ 1})} \middle| \ i\right\} \quad 2\exp\left(\ \ \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{48\lambda_{d\ \ 1}(S^{d\ \ 1})}\right). \tag{66}$$

By plugging Eq. (65) and Eq. (66) into Eq. (56) and applying the union bound, we have

$$\Pr_{\mathbf{V}_0}\left\{c_{r_0}^i < \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{4\lambda_{d\ \ 1}(S^{d\ \ 1})} \middle| \ i\right\} \quad 4\exp\left(\ \ \frac{p\lambda_{d\ \ 1}(B_{\boldsymbol{v}}^{r_0})}{48\lambda_{d\ \ 1}(S^{d\ \ 1})}\right).$$

By letting $r_0 = \text{ŕ}$ and by Eq. (59), we thus have

$$\Pr_{\mathbf{V}_0}\left\{c_{r_i}^i \quad 2npD_{\mathbf{X}} \middle| \ i\right\} \quad 4\exp\left(\ \ \frac{1}{6}npD_{\mathbf{X}}\right),$$

i.e.,

$$\Pr_{\mathbf{V}_0}[K_i] \quad 4\exp\left(\ \ \frac{1}{6}npD_{\mathbf{X}}\right), \text{ for all } i = 1, 2, \quad , n. \tag{67}$$

Step3: estimate the probability of J

We have

$$\Pr_{\mathbf{V}_0}[J^c] \quad \sum_{i=1}^{n} \Pr_{\mathbf{V}_0}[J_i] \text{ (by Eq. (62) and the union bound)}$$
$$\sum_{i=1}^{n} \Pr_{\mathbf{V}_0}[K_i] \text{ (by } J_i \quad K_i \text{ proven in Step 1)}$$
$$4n\exp\left(\ \ \frac{1}{6}npD_{\mathbf{X}}\right) \text{ (by Eq. (67)).}$$

Thus, we have

$$\Pr_{\mathbf{V}_0}[J] = 1 \quad \Pr_{\mathbf{V}_0}[J^c] \quad 1 \quad 4n\exp\left(\ \ \frac{1}{6}npD_{\mathbf{X}}\right).$$

The result of this lemma thus follows. □

F.2. Proof of Lemma 28

Based on Lemma 27, it remains to estimate $\hat{\Lambda}$, which will then allow us to bound $D_{\mathbf{X}}$. Towards this end, we need a few lemmas to estimate $B\left(\frac{d-1}{2}, \frac{1}{2}\right)$ and $I_x\left(\frac{d-1}{2}, \frac{1}{2}\right)$.

Lemma 29. *For any $x \geq \mathbb{R}$, we must have $x + 1 \leq e^x$.*

Proof. Consider a function $g(x) = e^x - x - 1$. It remains to show that $g(x) \geq 0$ for all x. We have $g'(x) = e^x - 1$. In other words, $g'(x) \leq 0$ when $x \leq 0$, and $g'(x) \geq 0$ when $x \geq 0$. Thus, $g(x)$ is monotone decreasing when $x \leq 0$, and is monotone increasing when $x \geq 0$. Hence, we know that $g(x)$ achieves its minimum value at $x = 0$, i.e., $g(x) \geq g(0) = 0$ for any x. The conclusion of this lemma thus follows. □

Lemma 30. *For any $d \geq 5$, we have*

$$\left(1 - \frac{1}{d-3}\right)^{d-3} \leq \frac{1}{e^2}.$$

Proof. By letting $x = \frac{1}{d-4}$ in Lemma 29, we have

$$\frac{d-3}{d-4} = \frac{1}{d-4} + 1 \leq \exp\left(\frac{1}{d-4}\right),$$

i.e.,

$$\frac{d-4}{d-3} \leq \exp\left(-\frac{1}{d-4}\right). \tag{68}$$

Thus, we have

$$\begin{aligned}\left(1 - \frac{1}{d-3}\right)^{d-3} &= \left(\frac{d-4}{d-3}\right)^{d-3} \\ &\leq \exp\left(-\frac{d-3}{d-4}\right) \\ &= \exp\left(-1 - \frac{1}{d-4}\right) \\ &\leq \exp(-2) \text{ (because } \exp(\cdot) \text{ is monotone increasing and } d \geq 5).\end{aligned}$$

□

Lemma 31. *For any $d \geq 5$, we must have*

$$\frac{2}{e}\sqrt{\frac{1}{d-3}} \geq \frac{1}{\sqrt{d}}$$

Proof. Because $1 - \frac{4}{e^2} \geq 0.46 \geq 0.6$, we have

$$\begin{aligned}&\frac{3}{5} - 1 - \frac{4}{e^2} \\ \Leftarrow\ &\frac{3}{d} - 1 - \frac{4}{e^2} \text{ (because } d \geq 5) \\ \Leftarrow\ &1 - \frac{3}{d} - \frac{4}{e^2} \\ \Leftarrow\ &\frac{d-3}{d} - \frac{4}{e^2} \\ \Leftarrow\ &\frac{4}{e^2}\frac{d}{d-3} - 1 \\ \Leftarrow\ &\frac{2}{e}\sqrt{\frac{1}{d-3}} - \frac{1}{\sqrt{d}}.\end{aligned}$$

□

Lemma 32.

$$B\left(\frac{d}{2}, \frac{1}{2}\right) \geq \left[\frac{1}{\sqrt{d}}, \pi\right].$$

Further, if $d \geq 5$*, we have*

$$B\left(\frac{d}{2}, \frac{1}{2}\right) \geq \left[\frac{1}{\sqrt{d}}, \frac{4}{\sqrt{d-3}}\right].$$

Proof. When $d = 2$, we have $B\left(\frac{d}{2}, \frac{1}{2}\right) = \pi$. When $d = 3$, we have $B\left(\frac{d}{2}, \frac{1}{2}\right) = 2$. When $d = 4$, we have $B\left(\frac{d}{2}, \frac{1}{2}\right) \geq 1.57$. It is easy to verify that the statement of the lemma holds for $d = 2, 3$, and 4. It remains to validate the case of $d \geq 5$. We first prove the lower bound. For any $m \in (0, 1)$, we have

$$\begin{aligned}
B\left(\frac{d}{2}, \frac{1}{2}\right) &= \int_0^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \\
&\geq \int_m^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \text{ (because } t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}} \geq 0) \\
&\geq m^{\frac{d}{2}-\frac{3}{2}}\int_m^1 (1-t)^{\frac{1}{2}}dt \\
&\quad \text{(because } t^{\frac{d}{2}-\frac{3}{2}} \text{ is monotone increasing with respect to } t \text{ when } d \geq 5) \\
&= m^{\frac{d}{2}-\frac{3}{2}}\left(-\frac{2}{3}(1-t)^{\frac{3}{2}}\Big|_m^1\right) \\
&= m^{\frac{d}{2}-\frac{3}{2}}\cdot\frac{2}{3}(1-m)^{\frac{3}{2}}.
\end{aligned}$$

By letting $m = 1 - \frac{1}{d-3}$, we thus have

$$B\left(\frac{d}{2}, \frac{1}{2}\right) \geq \left(1-\frac{1}{d-3}\right)^{\frac{d}{2}-\frac{3}{2}}\cdot\frac{2}{3}\sqrt{\frac{1}{d-3}}.$$

Then, applying Lemma 30, we have

$$B\left(\frac{d}{2}, \frac{1}{2}\right) \geq \frac{2}{e}\cdot\frac{1}{3}\sqrt{\frac{1}{d-3}}.$$

Thus, by Lemma 31, we have

$$B\left(\frac{d}{2}, \frac{1}{2}\right) \geq \frac{1}{\sqrt{d}}.$$

Now we prove the upper bound. For any $m \in (0, 1)$, we have

$$\begin{aligned}
B\left(\frac{d}{2}, \frac{1}{2}\right) &= \int_0^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \\
&= \int_0^m t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt + \int_m^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \\
&\leq \int_0^m t^{\frac{d}{2}-\frac{3}{2}}(1-m)^{\frac{1}{2}}dt + \int_m^1 (1-t)^{\frac{1}{2}}dt \\
&= \frac{2}{d-1}m^{\frac{d}{2}-\frac{1}{2}}(1-m)^{\frac{1}{2}} + \frac{2}{3}(1-m)^{\frac{3}{2}} \\
&\leq \frac{2}{d-1}(1-m)^{\frac{1}{2}} + \frac{2}{3}(1-m)^{\frac{3}{2}} \text{ (because } m < 1 \text{ and } d \geq 5).
\end{aligned}$$

By letting $m = 1 \quad \frac{1}{d \quad 3}$, we thus have

$$B\left(\frac{d \quad 1}{2}, \frac{1}{2}\right) \quad \frac{2^{p\overline{d \quad 3}}}{d \quad 1} + \not{p}\frac{2}{\overline{d \quad 3}} \quad \not{p}\frac{4}{\overline{d \quad 3}}.$$

Notice that $\not{p}\frac{4}{\overline{5 \quad 3}} = 2^{p\overline{2}} < \pi$. The result of this lemma thus follows.

□

Lemma 33. *Recall C_d is defined in Eq. (60). If $d \quad n^4$ and $\delta \quad 1$, then*

$$\left(1 \quad \frac{\delta^2}{4n^4 C_d^2}\right)^{\frac{d \quad 1}{2}} \quad \frac{1}{2}.$$

Proof. We have

$$\begin{aligned}
&\left(1 \quad \frac{\delta^2}{4n^4 C_d^2}\right)^{\frac{d \quad 1}{2}} \quad \left(1 \quad \frac{\delta^2}{4n^4 C_d^2}\right)^{d \quad 1} \\
&\quad 1 \quad \frac{(d \quad 1)\delta^2}{4n^4 C_d^2} \text{ (by Bernoulli's inequality } (1 + x)^a \quad 1 + ax) \\
&= 1 \quad \frac{(d \quad 1)\left(B\left(\frac{d \quad 1}{2}, \frac{1}{2}\right)\right)^2}{4n^4 \quad 8} \text{ (by } \delta \quad 1 \text{ and Eq. (60))} \\
&\quad 1 \quad \frac{(d \quad 1)\pi^2}{32n^4} \text{ (by Lemma 32)} \\
&\quad 1 \quad \frac{d}{n^4} \quad \frac{\pi^2}{32} \\
&\quad \frac{1}{2} \text{ (because } n^4 \quad d \text{ and } \pi \quad 4).
\end{aligned}$$

□

Lemma 34. *For any $\delta \; 2 \; (0, \frac{2}{\pi}]$, we must have $\frac{\delta}{n^2 C_d} \quad \not{p}\frac{1}{\overline{2}}$.*

Proof. Because Eq. (60), $\delta \quad \frac{2}{\pi}$, and $n \quad 1$, this lemma directly follows by Lemma 32. □

Lemma 35. *For any $x \; 2 \; [0, 1]$, we must have*

$$I_x\left(\frac{d \quad 1}{2}, \frac{1}{2}\right) \quad \not{p}\frac{C_d}{\overline{2}(d \quad 1)} x^{\frac{d \quad 1}{2}},$$

and

$$\lim_{x \; ! \; 0} \frac{I_x\left(\frac{d \quad 1}{2}, \frac{1}{2}\right)}{x^{\frac{d \quad 1}{2}}} = \not{p}\frac{C_d}{\overline{2}(d \quad 1)}.$$

Proof. we have

$$
\begin{aligned}
I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right) &= \frac{\int_0^x t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt}{B\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)} \\
&= \frac{C_d}{2\cancel{p}\overline{2}}\int_0^x t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \text{ (by Eq. (60))} \\
&\in 2\left[\frac{C_d}{2\cancel{p}\overline{2}}\int_0^x t^{\frac{d}{2}-\frac{3}{2}}dt,\ \frac{C_d}{2\cancel{p}\overline{2}\cancel{p}\overline{1-x}}\int_0^x t^{\frac{d}{2}-\frac{3}{2}}dt\right] \\
&\quad \text{(because } (1-t)^{-1/2} \in 2\left[1, \cancel{p}\frac{1}{\overline{1-x}}\right]\text{)} \\
&\in 2\left[\cancel{p}\frac{C_d}{\overline{2}(d-1)}x^{\frac{d}{2}-\frac{1}{2}},\ \cancel{p}\frac{C_d}{\overline{2}(d-1)\cancel{p}\overline{1-x}}x^{\frac{d}{2}-\frac{1}{2}}\right].
\end{aligned}
$$

Thus, we have

$$
\frac{I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}{x^{\frac{d}{2}-\frac{1}{2}}} \in 2\left[\cancel{p}\frac{C_d}{\overline{2}(d-1)},\ \cancel{p}\frac{C_d}{\overline{2}(d-1)\cancel{p}\overline{1-x}}\right],
$$

which implies

$$
\lim_{x\to 0}\frac{I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}{x^{\frac{d}{2}-\frac{1}{2}}} = \cancel{p}\frac{C_d}{\overline{2}(d-1)}.
$$

□

Lemma 36. *For any $x \in \left[\frac{1}{2}, 1\right)$ and for any $d \in \{2, 3, \ldots\}$, we have*

$$
I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right) \le \frac{2\sqrt{2(1-x)}}{B\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}.
$$

We also have

$$
\lim_{(1-x)\to 0^+}\frac{1-I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}{\cancel{p}\overline{1-x}} = \frac{2}{B\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}.
$$

Proof. By the definition of regularized incomplete beta function in Eq. (20), we have

$$
I_x\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right) = \frac{\int_0^x t^{\frac{d}{2}-\frac{1}{2}-1}(1-t)^{\frac{1}{2}}dt}{B\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)} = 1 - \frac{\int_x^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt}{B\left(\frac{d}{2}-\frac{1}{2},\frac{1}{2}\right)}.
$$

Thus, it remains to show that

$$
\int_x^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt \le 2\sqrt{2(1-x)}, \text{ and} \tag{69}
$$

$$
\lim_{(1-x)\to 0^+}\frac{\int_x^1 t^{\frac{d}{2}-\frac{3}{2}}(1-t)^{\frac{1}{2}}dt}{\cancel{p}\overline{1-x}} = 2. \tag{70}
$$

First, we prove Eq. (69). Case 1: $d = 2$. We have

$$\begin{aligned}
&\int_x^1 t^{\frac{d-3}{2}}(1-t)^{\frac{1}{2}}dt \\
=&\int_x^1 t^{-\frac{1}{2}}(1-t)^{\frac{1}{2}}dt \\
\le&\frac{1}{\sqrt{x}}\int_x^1 (1-t)^{\frac{1}{2}}dt \text{ (because } t^{-\frac{1}{2}} \text{ is monotone decreasing in } [x,\,1]) \\
=&2\sqrt{\frac{1-x}{x}} \\
\le&2\sqrt{2(1-x)} \text{ (because } x \ge \frac{1}{2}).
\end{aligned}$$

Case 2: $d \ge 3$. Then $t^{\frac{d-3}{2}}$ is monotone increasing in $[x,\,1]$. Thus, we have

$$\int_x^1 t^{\frac{d-3}{2}}(1-t)^{\frac{1}{2}}dt \le \int_x^1 (1-t)^{\frac{1}{2}}dt = 2\sqrt{1-x} \le 2\sqrt{2(1-x)}.$$

To conclude, for all $d \in \{2, 3, \ldots\}$, Eq. (69) holds.

Second, we prove Eq. (70). We have

$$\begin{aligned}
\frac{\int_x^1 t^{\frac{d-3}{2}}(1-t)^{\frac{1}{2}}dt}{\sqrt{1-x}} \in& \left[\frac{\min\{1,\, x^{\frac{d-3}{2}}\}\int_x^1(1-t)^{\frac{1}{2}}dt}{\sqrt{1-x}},\ \frac{\max\{1,\, x^{\frac{d-3}{2}}\}\int_x^1(1-t)^{\frac{1}{2}}dt}{\sqrt{1-x}}\right] \\
=&\left[2\min\{1,\, x^{\frac{d-3}{2}}\},\ 2\max\{1,\, x^{\frac{d-3}{2}}\}\right].
\end{aligned}$$

Since $\lim_{x \to 1} x^{\frac{d-3}{2}} = 1$, Eq. (70) thus follows. □

Now we are ready to prove Lemma 28.

Recall $\boldsymbol{v}_{\cdot,i}$ defined in Eq. (55). For any $b \in \left(0, \frac{1}{\sqrt{2}}\right]$, we have, for $\boldsymbol{x}$ independent of $\boldsymbol{v}_{\cdot,i}$ and with distribution μ,

$$\begin{aligned}
\Pr_{\boldsymbol{x}\sim\mu}\left\{|\boldsymbol{v}_{\cdot,i}^T\boldsymbol{x}| \le b\right\} =& I_{1-b^2}\left(\frac{d-1}{2}, \frac{1}{2}\right) \text{ (because Lemma 10)} \\
\ge& \frac{2\sqrt{2(1-(1-b^2))}}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)} \text{ (by Lemma 36)} \\
\ge& C_d b \text{ (by the definition of } C_d \text{ in Eq. (60))}.
\end{aligned} \tag{71}$$

Since each of the $\mathbf{X}_j$, $j \ne i$, is independent of $\boldsymbol{v}_{\cdot,i}$, we have

$$\begin{aligned}
&\Pr_{\mathbf{X}}\left\{\min_{j\in\{1,2,\ldots,n\}\setminus\{i\}}|\boldsymbol{v}_{\cdot,i}^T\mathbf{X}_j| \le b\right\} \\
=&\left(\Pr_{\boldsymbol{x}\sim\mu}\left\{|\boldsymbol{v}_{\cdot,i}^T\boldsymbol{x}| \le b\right\}\right)^{n-1} \text{ (because each } \mathbf{X}_j, j \ne i, \text{ is } \textit{i.i.d.} \text{ and independent of } \boldsymbol{v}_{\cdot,i}) \\
\ge&(1-C_d b)^{n-1} \text{ (by Eq. (71))} \\
\ge&1-(n-1)C_d b \text{ (by Bernoulli's inequality)} \\
\ge&1-nC_d b.
\end{aligned}$$

Or, equivalently,

$$\Pr_{\mathbf{X}}\left\{\min_{i\in\{1,2,\ldots,n\}\setminus\{i\}}|\boldsymbol{v}_{\cdot,i}^T\mathbf{X}_i| < b\right\} \le nC_d b. \tag{72}$$

Recall the definition of r_i and $\clubsuit$ in Eqs. (57)(58). Thus, we then have

$$
\begin{aligned}
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\clubsuit < \frac{\delta}{n^2 C_d}\right\} \\
&n \Pr_{\mathbf{X},\mathbf{V}_0}\left\{r_i < \frac{\delta}{n^2 C_d}\right\} \text{ (by Eq. (58) and the union bound)} \\
=&n \Pr_{\mathbf{X}}\left\{r_i < \frac{\delta}{n^2 C_d}\right\} \text{ (because } r \text{ is independent of } \mathbf{V}_0) \\
=&n \Pr_{\mathbf{X}}\left\{\min_{j \in \{1,2,\ldots,n\}} \left|\boldsymbol{v}_{,i}^T \mathbf{X}_j\right| < \frac{\delta}{n^2 C_d}\right\} \text{ (by Eq. (57))} \\
&n \cdot nC_d \cdot \frac{\delta}{n^2 C_d} \text{ (by letting } b = \frac{\delta}{n^2 C_d} \text{ in Eq. (72) and } b \leq \frac{1}{2} \text{ because of Lemma 34)} \\
=&\delta.
\end{aligned} \tag{73}
$$

By Lemma 9 and Eq. (61), we have

$$
\lambda_{d-1}\left(B^{\frac{\delta}{n^2 C_d}}\right) = \frac{1}{2}\lambda_{d-1}\left(S^{d-1}\right) I_{\frac{\delta^2}{n^4 C_d^2}\left(1-\frac{\delta^2}{4n^4 C_d^2}\right)}\left(\frac{d-1}{2}, \frac{1}{2}\right) = 8n\lambda_{d-1}\left(S^{d-1}\right) D(n,d,\delta).
$$

Thus, we have

$$
D(n,d,\delta) = \frac{\lambda_{d-1}\left(B^{\frac{\delta}{n^2 C_d}}\right)}{8n\lambda_{d-1}\left(S^{d-1}\right)}. \tag{74}
$$

By Eq. (59) and Eq. (74), we have

$$
D_{\mathbf{X}} \geq D(n,d,\delta), \text{ when } \clubsuit \leq \frac{\delta}{n^2 C_d}.
$$

Notice that $\clubsuit$ only depends on $\mathbf{X}$ and is independent of $\mathbf{V}_0$. By Lemma 27, for any $\mathbf{X}$ that makes $\clubsuit \leq \frac{\delta}{n^2 C_d}$, we must have

$$
\Pr_{\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1}\right\} \geq 1 - 4ne^{-npD(n,d,\delta)/6}.
$$

In other words,

$$
\Pr_{\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1} \,\middle|\, \text{any given } \mathbf{X} \text{ such that } \clubsuit \leq \frac{\delta}{n^2 C_d}\right\} \geq 1 - 4ne^{-npD(n,d,\delta)/6}.
$$

We thus have

$$
\Pr_{\mathbf{X},\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1} \,\middle|\, \clubsuit \leq \frac{\delta}{n^2 C_d}\right\} \geq 1 - 4ne^{-npD(n,d,\delta)/6}. \tag{75}
$$

Thus, we have

$$
\begin{aligned}
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1}\right\} \\
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\clubsuit \leq \frac{\delta}{n^2 C_d}, \text{ and } k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1}\right\} \\
=&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a} \in S^{n-1} \,\middle|\, \clubsuit \leq \frac{\delta}{n^2 C_d}\right\} \Pr_{\mathbf{X},\mathbf{V}_0}\left\{\clubsuit \leq \frac{\delta}{n^2 C_d}\right\} \\
&(1 - 4ne^{-npD(n,d,\delta)/6})(1-\delta) \text{ (by Eq. (73) and Eq. (75))} \\
&\geq 1 - 4ne^{-npD(n,d,\delta)/6} - \delta.
\end{aligned}
$$

The result of this lemma thus follows.

F.3. Proof of Lemma 25

Based on Lemma 28, it only remains to estimate $D(n,d,\delta)$. We start with a lemma.

Lemma 37. *If $\delta \quad 1$ and $d \quad n^4$, we must have*

$$D(n,d,\delta) \quad 2^{\ 1.5d\ \ 5.5}d^{\ 0.5d}n^{\ \ 2d+1}\delta^{d\ \ 1}. \tag{76}$$

For any given $\delta \quad 0$ and d, we must have

$$\lim_{n\ !\ 1}\frac{D(n,d,\delta)}{n^{2d\ \ 1}} = 2^{\ 1.5d\ \ 1.5}\left(B\left(\frac{d\quad 1}{2},\frac{1}{2}\right)\right)^{d\ \ 2}\frac{1}{d\quad 1}\delta^{d\ \ 1}.$$

Proof. We start from

$$\begin{aligned}\frac{1}{(d\quad 1)C_d^{d\ \ 2}} &= \frac{\left(B\left(\frac{d\quad 1}{2},\frac{1}{2}\right)\right)^{d\ \ 2}}{(d\quad 1)(2\pi)^{d\ \ 2}} \text{ (by Eq. (60))}\\ &\quad \frac{1}{(d\quad 1)d^{\frac{d}{2}\ \ 1}(2\pi)^{d\ \ 2}} \text{ (by Lemma 32)}\\ &\quad \frac{1}{d^{\frac{d}{2}}(2\pi)^d}\\ &= (8d)^{\ \frac{d}{2}}.\end{aligned} \tag{77}$$

Thus, we have

$$\begin{aligned}D(n,d,\delta) &\quad \frac{1}{16n}\pi\frac{C_d}{2(d\quad 1)}\left(\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}} \text{ (by Eq. (61) and Lemma 35)}\\ &= \frac{1}{16}\pi\frac{1}{2}\frac{1}{(d\quad 1)C_d^{d\ \ 2}}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)^{\frac{d\ \ 1}{2}}\frac{\delta^{d\ \ 1}}{n^{2d\ \ 1}}\\ &\quad \frac{1}{32}\pi\frac{1}{2}(8d)^{\ \frac{d}{2}}\frac{\delta^{d\ \ 1}}{n^{2d\ \ 1}} \text{ (by Lemma 33 and Eq. (77))}\\ &= 2^{\ 1.5d\ \ 5.5}d^{\ 0.5d}n^{\ \ 2d+1}\delta^{d\ \ 1}.\end{aligned}$$

For any given d and $\delta \quad 0$, we have

$$\begin{aligned}\lim_{n\ !\ 1}\frac{D(n,d,\delta)}{n^{2d\ \ 1}} &= \lim_{n\ !\ 1}\frac{1}{16n^{2d\ \ 2}}I_{\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)}\left(\frac{d\quad 1}{2},\frac{1}{2}\right) \text{ (by Eq. (61))}\\ &= \lim_{n\ !\ 1}\frac{\left(\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}}{16n^{2d\ \ 2}}\ \frac{I_{\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)}\left(\frac{d\quad 1}{2},\frac{1}{2}\right)}{\left(\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}}\\ &= \frac{1}{16}\lim_{n\ !\ 1}\left(\frac{\delta^2}{C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}\ \frac{I_{\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)}\left(\frac{d\quad 1}{2},\frac{1}{2}\right)}{\left(\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}}\\ &= \frac{1}{16}\lim_{n\ !\ 1}\left(\frac{\delta^2}{C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}\ \lim_{n\ !\ 1}\frac{I_{\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)}\left(\frac{d\quad 1}{2},\frac{1}{2}\right)}{\left(\frac{\delta^2}{n^4C_d^2}\left(1\quad\frac{\delta^2}{4n^4C_d^2}\right)\right)^{\frac{d\ \ 1}{2}}}\\ &= \frac{1}{16}\frac{\delta^{d\ \ 1}}{C_d^{d\ \ 1}}\pi\frac{C_d}{2(d\quad 1)} \text{ (by Lemma 35)}\\ &= 2^{\ 1.5d\ \ 1.5}\left(B\left(\frac{d\quad 1}{2},\frac{1}{2}\right)\right)^{d\ \ 2}\frac{1}{d\quad 1}\delta^{d\ \ 1} \text{ (by Eq. (60)).}\end{aligned}$$

□

Now we are ready to finish our proof of Lemma 25.

We have

$$D(n,d,\delta)\Big|_{\delta=\frac{1}{m}\sqrt{\frac{2}{\pi n}}} \quad \frac{1}{2^{1.5d+5.5}d^{0.5d}n^{2d\ \ 1}n^{\frac{d\ 1}{m}}} \text{ (by Eq. (76))}$$
$$= \frac{1}{2^{1.5d+5.5}d^{0.5d}n^{(2+\frac{1}{m})(d\ \ 1)}n}$$
$$= \frac{1}{J_m(n,d)n} \text{ (by Eq. (9)).}$$

Thus, when $p \quad 6J_m(n,d)\ln\left(4n^{1+\frac{1}{m}}\right)$, we have

$$1 \quad 4ne^{\ npD(n,d,\delta)/6} \quad \delta\Big|_{\delta=\frac{1}{m}\sqrt{\frac{2}{\pi n}}} \quad 1 \quad \frac{2}{m}\sqrt{\frac{p}{\pi n}}.$$

Then, we have

$$m\ 2\left[1,\frac{\ln n}{\ln\frac{\pi}{2}}\right] = \left(\frac{\pi}{2}\right)^m \quad n = n^{\frac{1}{m}} \quad \frac{\pi}{2} = \frac{1}{m}\sqrt{\frac{p}{\pi n}} \quad \frac{2}{\pi} = \delta \quad \frac{2}{\pi}.$$

By Lemma 26 and Lemma 28, the conclusion of Lemma 25 thus follows.

G. Upper bound of $\min\text{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$

By Lemma 26, to get an upper bound of $\min\text{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$, it is equivalent to get an upper bound of $\min_{\boldsymbol{a}2S^{n\ 1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$. To that end, we only need to construct a vector $\boldsymbol{a}$ and calculate the value of $\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$, which automatically becomes an upper bound $\min_{\boldsymbol{a}2S^{n\ 1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$.

The following lemma shows that, for given $\mathbf{X}$, if two input training data $\mathbf{X}_i$ and $\mathbf{X}_k$ are close to each other, then $\min_{\boldsymbol{a}2S^{n\ 1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$ is unlikely to be large.

Lemma 38. *If there exist $\mathbf{X}_i$ and $\mathbf{X}_k$ such that $i \neq k$ and $\theta := \arccos(\mathbf{X}_i^T\mathbf{X}_k)$, then*

$$\Pr_{\mathbf{V}_0}\left\{\min_{\boldsymbol{a}2S^{n\ 1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \quad \frac{3p\theta^2}{8}+\frac{3p\theta}{4\pi}\right\} \quad 2\exp\left(\ \frac{p}{24}\right)+2\exp\left(\ \frac{p\theta}{12}\right).$$

Intuitively, Lemma 38 is true because, when $\mathbf{X}_i$ and $\mathbf{X}_k$ are similar, $\mathbf{H}_i$ and $\mathbf{H}_k$ (the i-th and k-th row of $\mathbf{H}$, respectively) will also likely be similar, i.e., $\|\mathbf{H}_i \quad \mathbf{H}_k\|_2$ is not likely to be large. Thus, we can construct $\boldsymbol{a}$ such that $\mathbf{H}^T\boldsymbol{a}$ is proportional to $\mathbf{H}_i \quad \mathbf{H}_k$, which will lead to the result of Lemma 38. We put the proof of Lemma 38 in Appendix G.1.

The next step is to estimate such difference between $\mathbf{X}_i$ and $\mathbf{X}_k$ (or equivalently, the angle θ between them). We have the following lemma.

Lemma 39. *When $n \quad \pi(d \quad 1)$, there must exist two different $\mathbf{X}_i$'s such that the angle between them is at most*

$$\theta = \pi\left((d \quad 1)B(\frac{d\ \ 1}{2},\frac{1}{2})\right)^{\frac{1}{d\ 1}} n^{\ \frac{1}{d\ 1}}.$$

Lemma 39 is intuitive because $S^{d\ 1}$ has limited area. When there are many $\mathbf{X}_i$'s on $S^{d\ 1}$, there must exist at least two $\mathbf{X}_i$'s that are relatively close. We put the proof of Lemma 39 in Appendix G.2.

Finally, we have the following lemma.

Lemma 40. *When n $\pi(d$ 1), we have*

$$\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\frac{\min \mathrm{eig}(\mathbf{H}\mathbf{H}^T)}{p} \quad \frac{3\pi^2}{8}\left((d \quad 1)B(\frac{d}{2}\ \frac{1}{2},\frac{1}{2})\right)^{\frac{2}{d\ 1}} n^{\ \frac{2}{d\ 1}} + \frac{3}{4}\left((d \quad 1)B(\frac{d}{2}\ \frac{1}{2},\frac{1}{2})\right)^{\frac{1}{d\ 1}} n^{\ \frac{1}{d\ 1}}\right\}$$
$$1 \quad 2\exp\left(\ \frac{p}{24}\right) \quad 2\exp\left(\ \frac{p}{12}\pi\left((d \quad 1)B(\frac{d}{2}\ \frac{1}{2},\frac{1}{2})\right)^{\frac{1}{d\ 1}} n^{\ \frac{1}{d\ 1}}\right).$$

Proof. This lemma directly follows by combining Lemma 26, Lemma 38, and Lemma 39. □

By Lemma 40, we can conclude that when p is much larger than $n^{\frac{1}{d\ 1}}$, $\frac{\min \mathrm{eig}(\mathbf{H}\mathbf{H}^T)}{p} = O(n^{\ \frac{1}{d\ 1}})$ with high probability.

G.1. Proof of Lemma 38

We first prove a useful lemma.

Lemma 41. *For any φ 2 $[0, 2\pi]$, we must have* $\sin\varphi$ φ. *For any φ 2 $[0, \pi/2]$, we must have φ* $\frac{\pi}{2}\sin\varphi$.

Proof. To prove the first part of the lemma, note that

$$\frac{d(\varphi \quad \sin\varphi)}{d\varphi} = 1 \quad \cos\varphi \quad 0.$$

Thus, the function $(\varphi \quad \sin\varphi)$ is monotone increasing with respect to φ 2 $[0, 2\pi]$. Thus, we have

$$\min_{\varphi 2[0,2\pi]}(\varphi \quad \sin\varphi) = (\varphi \quad \sin\varphi)\big|_{\varphi=0} = 0.$$

In other words, we have $\sin\varphi$ φ for any φ 2 $[0, 2\pi]$.

To prove the second part of the lemma, note that when φ 2 $[0, \pi/2]$, we have

$$\frac{d^2(\varphi \quad \frac{\pi}{2}\sin\varphi)}{d\varphi^2} = \frac{\pi}{2}\sin\varphi \quad 0.$$

Thus, the function φ $\frac{\pi}{2}\sin\varphi$ is convex with respect to φ 2 $[0, \pi/2]$. Because the maximum of a convex function must be attained at the endpoint of the domain interval, we have

$$\max_{\varphi 2[0,\pi/2]}(\varphi \quad \frac{\pi}{2}\sin\varphi) = \max_{\varphi 2 f0,\pi/2g}(\varphi \quad \frac{\pi}{2}\sin\varphi) = 0.$$

Thus, we have φ $\frac{\pi}{2}\sin\varphi$ for any φ 2 $[0, \pi/2]$. □

Now we are ready to prove Lemma 38.

Proof. Through the proof, we fix $\mathbf{X}_i$ and $\mathbf{X}_k$, and only consider the randomness of $\mathbf{V}_0$. Because θ is the angle between $\mathbf{X}_i$ and $\mathbf{X}_k$ and because of Assumption 1, we have

$$\begin{aligned} k\mathbf{X}_i \quad \mathbf{X}_k k_2 &= 2\sin\frac{\theta}{2} \\ &\quad 2 \quad \frac{\theta}{2} \text{ (by Lemma 41)} \\ &= \theta. \end{aligned} \tag{78}$$

Let $\boldsymbol{a} = \frac{1}{2}(\boldsymbol{e}_i \quad \boldsymbol{e}_k)$, where $\boldsymbol{e}_q$ denotes the q-th standard basis vector, $q = 1, 2, \quad , n$. Then, we have

$$\begin{aligned}
k\mathbf{H}^T\boldsymbol{a}k_2^2 &= \frac{1}{2}k\mathbf{H}_i^T \quad \mathbf{H}_k^T k_2^2 \\
&= \frac{1}{2}\sum_{j=1}^{p}\left\|\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}}\mathbf{X}_i \quad \mathbf{1}_{\{\mathbf{X}_k^T\mathbf{V}_0[j]>0\}}\mathbf{X}_k\right\|_2^2 \text{ (by Eq. (1))} \\
&= \frac{1}{2}\sum_{j=1}^{p}\left(\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0,\ \mathbf{X}_k^T\mathbf{V}_0[j]>0\}}k\mathbf{X}_i \quad \mathbf{X}_k k_2^2 + \mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}\right) \text{ (by } k\mathbf{X}_i k_2^2 = k\mathbf{X}_k k_2^2 = 1) \\
&\quad \frac{1}{2}\sum_{j=1}^{p}\left(\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0,\ \mathbf{X}_k^T\mathbf{V}_0[j]>0\}}\theta^2 + \mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}\right) \text{ (by Eq. (78))} \\
&\quad \frac{\theta^2}{2}\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}} + \frac{1}{2}\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}. \qquad (79)
\end{aligned}$$

Since $\mathbf{X}_i$ is fixed and the direction of $\mathbf{V}_0[j]$ is uniformly distributed, we have $\Pr_{\mathbf{V}_0}\{\mathbf{X}_i^T\mathbf{V}_0[j] > 0\} = \frac{1}{2}$ and

$$\begin{aligned}
\Pr_{\mathbf{V}_0}\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j]) < 0\} &= 2\Pr_{\mathbf{V}_0}\{\mathbf{X}_i^T\mathbf{V}_0[j] > 0,\ \mathbf{X}_k^T\mathbf{V}_0[j] < 0\} \\
&= 2\Pr_{\mathbf{V}_0}\{\mathbf{X}_i^T\mathbf{V}_0[j] > 0, \quad \mathbf{X}_k^T\mathbf{V}_0[j] > 0\} \\
&= 2\int_{S^{d\ \ 1}}\mathbf{1}_{\{\mathbf{X}_i^T\boldsymbol{v}>0, \quad \mathbf{X}_k^T\boldsymbol{v}>0\}}d\lambda(\boldsymbol{v}) \\
&= 2\ \ \frac{\pi \quad (\pi \quad \theta)}{2\pi} \text{ (by Lemma 17)} \\
&= \frac{\theta}{\pi}.
\end{aligned}$$

Thus, based on the randomness of $\mathbf{V}_0$, when $\mathbf{X}$ are given, we have

$$\begin{aligned}
&\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}} \quad \text{Bino}\left(p,\ \frac{1}{2}\right), \\
&\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}} \quad \text{Bino}\left(p,\ \frac{\theta}{\pi}\right).
\end{aligned}$$

By letting $\delta = \frac{1}{2}$, $a = p$, $b = \frac{1}{2}$ in Lemma 14, we then have

$$\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}} \quad \frac{3p}{4}\right\} \quad 2\exp\left(\quad \frac{p}{24}\right), \qquad (80)$$

$$\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}} \quad \frac{3p\theta}{2\pi}\right\} \quad 2\exp\left(\quad \frac{p\theta}{12\pi}\right). \qquad (81)$$

Thus, we have

$$\begin{aligned}
&\Pr_{\mathbf{V}_0}\left\{k\mathbf{H}^T\boldsymbol{a}k_2^2 \quad \frac{3p\theta^2}{8}+\frac{3p\theta}{4\pi}\right\}\\
&\Pr_{\mathbf{V}_0}\left\{\frac{\theta^2}{2}\sum_{j=1}^{p}\mathbf{1}_{f\mathbf{X}_i^T\mathbf{V}_0[j]>0g}+\frac{1}{2}\sum_{j=1}^{p}\mathbf{1}_{f(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0g} \quad \frac{3p\theta^2}{8}+\frac{3p\theta}{4\pi}\right\}\\
&\text{(by Eq. (79))}\\
&\Pr_{\mathbf{V}_0}\left\{\left\{\sum_{j=1}^{p}\mathbf{1}_{f\mathbf{X}_i^T\mathbf{V}_0[j]>0g}>\frac{3p}{4}\right\}\ [\ \left\{\sum_{j=1}^{p}\mathbf{1}_{f(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0g} \quad \frac{3p\theta}{2\pi}\right\}\right\}\\
&\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{f\mathbf{X}_i^T\mathbf{V}_0[j]>0g}>\frac{3p}{4}\right\}+\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{f(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0g} \quad \frac{3p\theta}{2\pi}\right\}\\
&\text{(by the union bound)}\\
&2\exp\left(\quad\frac{p}{24}\right)+2\exp\left(\quad\frac{p\theta}{12}\right)\ \text{(by Eq. (80) and Eq. (81)).}
\end{aligned}$$

The result of Lemma 38 thus follows. □

G.2. Proof of Lemma 39

We first prove a useful lemma. Recall the definition of C_d in Eq. (60).

Lemma 42. *We have*

$$\frac{2^{p}\overline{2}(d\quad 1)}{nC_d}\ 2\ \left[\frac{d}{n}p\frac{1}{\overline{d}},\ \frac{\pi(d\quad 1)}{n}\right].$$

Proof. By Lemma 32 and Eq. (60), we have

$$C_d\ 2\ \left[\frac{2^{p}\overline{2}}{\pi},\ 2^{p}\overline{2d}\right].$$

Thus, we have

$$\frac{2^{p}\overline{2}(d\quad 1)}{nC_d}\ 2\ \left[\frac{d}{n}p\frac{1}{\overline{d}},\ \frac{\pi(d\quad 1)}{n}\right].$$

□

Now we are ready to proof Lemma 39.

Proof. Recall the definition of θ in Lemma 39. Draw n caps on $S^{d\ \ 1}$ centered at $\mathbf{X}_1,\ \mathbf{X}_2,\quad ,\mathbf{X}_n$ with the colatitude angle φ where

$$\varphi=\frac{\theta}{2}=\frac{\pi}{2}\left(\frac{2^{p}\overline{2}(d\quad 1)}{nC_d}\right)^{\frac{1}{d\ 1}}\ \text{(by Eq. (60)).} \tag{82}$$

By Lemma 42 and $n\quad \pi(d\quad 1)$, we have $\varphi\ 2\ [0,\pi/2]$. Thus, by Lemma 41, we have

$$\sin\varphi\quad\frac{2\varphi}{\pi}=\left(\frac{2^{p}\overline{2}(d\quad 1)}{nC_d}\right)^{\frac{1}{d\ 1}}. \tag{83}$$

By Lemma 8, the area of each cap is

$$A = \frac{1}{2}\lambda_{d\ \ 1}(\mathbb{S}^{d\ \ 1})I_{\sin^2\varphi}\left(\frac{d\quad 1}{2},\frac{1}{2}\right).$$

Applying Lemma 35 and Eq. (83), we thus have

$$A\quad \frac{1}{2}\lambda_{d\ \ 1}(\mathbb{S}^{d\ \ 1})\,p\frac{C_d}{2(d\quad 1)}(\sin^2\varphi)^{\frac{d\ \ 1}{2}} = \frac{1}{n}\lambda_{d\ \ 1}(\mathbb{S}^{d\ \ 1}).$$

In other words, we have

$$\frac{\lambda_{d\ \ 1}(\mathbb{S}^{d\ \ 1})}{A}\quad n.$$

By the pigeonhole principle, we know there exist at least two different caps that overlap, i.e., the angle between them is at most 2φ. The result of this lemma thus follows by Eq. (82). □

H. Proof of Proposition 5

We follow the sketch of proof in Section 5. Recall the definition of the pseudo ground-truth function $f^g_{\mathbf{V}_0}$ in Definition 2, and the corresponding $\mathbf{V}\ \ 2\ \mathbb{R}^{dp}$ that

$$\mathbf{V}\ [j] = \int_{\mathbb{S}^{d\ \ 1}} \mathbf{1}_{\{\boldsymbol{z}^T\mathbf{V}_0[j]>0\}}\boldsymbol{z}\frac{g(\boldsymbol{z})}{p}d\mu(\boldsymbol{z}),\ \text{for all } j\ 2\ \{1,2,\quad ,p\}. \tag{84}$$

We first show that the pseudo ground-truth can be written in a linear form.

Lemma 43. *$h_{\mathbf{V}_0,\boldsymbol{x}}\quad \mathbf{V}\ = f^g_{\mathbf{V}_0}(\boldsymbol{x})$ for all $\boldsymbol{x}\ 2\ \mathbb{S}^{d\ \ 1}$.*

Proof. For all $\boldsymbol{x}\ 2\ \mathbb{S}^{d\ \ 1}$, we have

$$\begin{aligned}
h_{\mathbf{V}_0,\boldsymbol{x}}\quad \mathbf{V}\ &= \sum_{j=1}^{p} h_{\mathbf{V}_0,\boldsymbol{x}}[j]\quad \mathbf{V}\ [j]\\
&= \sum_{j=1}^{p}\mathbf{1}_{\{\boldsymbol{x}^T\mathbf{V}_0[j]>0\}}\quad \boldsymbol{x}^T\int_{\mathbb{S}^{d\ \ 1}}\mathbf{1}_{\{\boldsymbol{z}^T\mathbf{V}_0[j]>0\}}\boldsymbol{z}\frac{g(\boldsymbol{z})}{p}d\mu(\boldsymbol{z}) \text{ (by Eq. (1) and Eq. (84))}\\
&= \int_{\mathbb{S}^{d\ \ 1}}\sum_{j=1}^{p}\mathbf{1}_{\{\boldsymbol{x}^T\mathbf{V}_0[j]>0\}}\quad \boldsymbol{x}^T\mathbf{1}_{\{\boldsymbol{z}^T\mathbf{V}_0[j]>0\}}\boldsymbol{z}\frac{g(\boldsymbol{z})}{p}d\mu(\boldsymbol{z})\\
&= \int_{\mathbb{S}^{d\ \ 1}}\boldsymbol{x}^T\boldsymbol{z}\frac{j\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}j}{p}g(\boldsymbol{z})d\mu(\boldsymbol{z}) \text{ (by Eq. (6))}\\
&= f^g_{\mathbf{V}_0}(\boldsymbol{x}) \text{ (by Definition 2).}
\end{aligned}$$

□

Let $\mathbf{P} := \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)\ \ ^1\mathbf{H}$. Since $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P} = \mathbf{P}^T$, we know that $\mathbf{P}$ is an orthogonal projection to the row-space of $\mathbf{H}$. Next, we give an expression for the test error. Note that even though Proposition 4 assumes no noise, below we state a more general version below with noise (which will be useful later).

Lemma 44. *If the ground-truth is $f(\boldsymbol{x}) = h_{\mathbf{V}_0,\boldsymbol{x}}\quad \mathbf{V}$ for all $\boldsymbol{x}$, then we have*

$$\hat{f}^{\ell_2}(\boldsymbol{x})\quad f(\boldsymbol{x}) = h_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P}\quad \mathbf{I})\quad \mathbf{V}\ + h_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)\ \ ^1\boldsymbol{\epsilon},\ \textit{for all } \boldsymbol{x}.$$

Proof. Because $f(\boldsymbol{x}) = h_{\mathbf{V}_0,\boldsymbol{x}}\quad \mathbf{V}$, we have $\boldsymbol{y} = \mathbf{H}\quad \mathbf{V}\ + \boldsymbol{\epsilon}$. Thus, we have

$$\begin{aligned}
\mathbf{V}^{\ell_2} &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)\ \ ^1\boldsymbol{y} \text{ (by Eq. (3))}\\
&= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)\ \ ^1(\mathbf{H}\quad \mathbf{V}\ + \boldsymbol{\epsilon}).
\end{aligned}$$

Further, we have

$$\begin{aligned}\mathbf{V}^{\ell_2} \quad \mathbf{V} &= \left(\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H} \quad \mathbf{I}\right) \quad \mathbf{V} + \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon} \\ &= (\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} + \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}.\end{aligned}$$

Finally, using Eq. (4), we have

$$\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V}^{\ell_2} \quad \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V} = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} + \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}.$$

□

When there is no noise, Lemma 44 reduces to $\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}$. As we described in Section 5, $(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}$ has the interpretation of the distance from $\mathbf{V}$ to the row-space of $\mathbf{H}$. We then have the following.

Lemma 45. *For all $\boldsymbol{a} \in \mathbb{R}^n$, we have*

$$\left\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\|_2 \quad \sqrt{\frac{p}{\bar{p}}}\left\|\mathbf{V} \quad \mathbf{H}\boldsymbol{a}\right\|_2.$$

Proof. Recall that $\mathbf{P} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}$. Thus, we have

$$\mathbf{P}\mathbf{H}^T = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{H}^T = \mathbf{H}^T. \tag{85}$$

We then have

$$\begin{aligned}\left\|(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\|_2 &= \left\|\mathbf{P} \quad \mathbf{V} \quad \mathbf{V}\right\|_2 \\ &= \left\|\mathbf{P}(\mathbf{H}^T\boldsymbol{a} + \quad \mathbf{V} \quad \mathbf{H}^T\boldsymbol{a}) \quad \mathbf{V}\right\|_2 \\ &= \left\|\mathbf{P}\mathbf{H}^T\boldsymbol{a} + \mathbf{P}(\quad \mathbf{V} \quad \mathbf{H}^T\boldsymbol{a}) \quad \mathbf{V}\right\|_2 \\ &= \left\|\mathbf{H}^T\boldsymbol{a} + \mathbf{P}(\quad \mathbf{V} \quad \mathbf{H}^T\boldsymbol{a}) \quad \mathbf{V}\right\|_2 \text{ (by Eq. (85))} \\ &= \left\|(\mathbf{P} \quad \mathbf{I})(\quad \mathbf{V} \quad \mathbf{H}^T\boldsymbol{a})\right\|_2 \\ &\quad \left\|\mathbf{V} \quad \mathbf{H}^T\boldsymbol{a}\right\|_2 \text{ (because } \mathbf{P} \text{ is an orthogonal projection).}\end{aligned}$$

Therefore, we have

$$\begin{aligned}\left\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\| &= \left\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\|_2 \\ &\quad \left\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\right\|_2 \quad \left\|(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\|_2 \text{ (by Lemma 12)} \\ &\quad \sqrt{\frac{p}{\bar{p}}}\left\|\mathbf{V} \quad \mathbf{H}\boldsymbol{a}\right\|_2 \text{ (by Lemma 11).}\end{aligned}$$

□

Now we are ready to prove Proposition 5.

Proof. Because there is no noise, we have $\boldsymbol{\epsilon} = \mathbf{0}$. Thus, by Lemma 44, we have

$$\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} . \tag{86}$$

We then have, for all $\boldsymbol{a} \in \mathbb{R}^n$,

$$\begin{aligned}&\Pr_{\mathbf{X}}\left\{\left|\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x})\right| \quad n^{\frac{1}{2}(1 \quad \frac{1}{q})}\right\} \\ =&\Pr_{\mathbf{X}}\left\{\left\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V}\right\| \quad n^{\frac{1}{2}(1 \quad \frac{1}{q})}\right\} \\ &\Pr_{\mathbf{X}}\left\{\sqrt{\frac{p}{\bar{p}}}\left\|\mathbf{H}^T\boldsymbol{a} \quad \mathbf{V}\right\|_2 \quad n^{\frac{1}{2}(1 \quad \frac{1}{q})}\right\} \text{ (by Lemma 45).}\end{aligned} \tag{87}$$

It only remains to find the vector $\boldsymbol{a}$. Define $\mathbf{K}_i \in \mathbb{R}^{dp}$ for $i = 1, 2, \ldots, n$ as

$$\mathbf{K}_i[j] := \mathbf{1}_{\{\mathbf{X}_i^T \mathbf{v}_0[j] > 0\}} \mathbf{X}_i \frac{g(\mathbf{X}_i)}{p}, \ j = 1, 2, \ldots, p.$$

By Eq. (84), for all $j = 1, 2, \ldots, p$, we have

$$\mathop{\mathbb{E}}_{\mathbf{X}_i}[\mathbf{K}_i[j]] = \mathbf{v}[j]. \tag{88}$$

Because $\|\mathbf{X}_i\|_2 = 1$, we have

$$\|\mathbf{K}_i[j]\|_2 \leq \frac{\|g\|_1}{p}.$$

Thus, we have

$$\|\mathbf{K}_i\|_2 = \sqrt{\sum_{j=1}^{p} \|\mathbf{K}_i[j]\|_2^2} \leq \frac{\|g\|_1}{\sqrt{p}},$$

i.e.,

$$\sqrt{p}\|\mathbf{K}_i\|_2 \leq \|g\|_1. \tag{89}$$

We now construct the vector $\boldsymbol{a}$. Define $\boldsymbol{a} \in \mathbb{R}^n$ whose i-th element is $\boldsymbol{a}_i = \frac{g(\mathbf{X}_i)}{np}$, $i = 1, 2, \ldots, n$. Notice that $\boldsymbol{a}$ is well-defined because $\|g\|_1 < 1$. Then, for all $j \in \{1, 2, \ldots, p\}$, we have

$$\begin{aligned}
(\mathbf{H}^T \boldsymbol{a})[j] &= \sum_{i=1}^{n} \mathbf{H}_i^T[j] \boldsymbol{a}_i \\
&= \sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{X}_i^T \mathbf{v}_0[j] > 0\}} \mathbf{X}_i \frac{g(\mathbf{X}_i)}{np} \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbf{K}_i[j],
\end{aligned}$$

i.e.,

$$\mathbf{H}^T \boldsymbol{a} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{K}_i. \tag{90}$$

Thus, by Eq. (89) and Lemma 16 (with $X_i = \sqrt{p}\mathbf{K}_i$, $U = \|g\|_1$, and $k = n$), we have

$$\Pr_{\mathbf{X}} \left\{ \sqrt{p} \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{K}_i \right) - \mathop{\mathbb{E}}_{\mathbf{X}} \mathbf{K}_1 \right\|_2 \geq n^{\frac{1}{2}(1 - \frac{1}{q})} \right\} \leq 2e^2 \exp\left(- \frac{p^{\frac{1}{q}} n}{8\|g\|_1^2} \right).$$

Further, by Eq. (90) and Eq. (88), we have

$$\Pr_{\mathbf{X}} \left\{ \sqrt{p} \|\mathbf{H}^T \boldsymbol{a} - \mathbf{v}\|_2 \geq n^{\frac{1}{2}(1 - \frac{1}{q})} \right\} \leq 2e^2 \exp\left(- \frac{p^{\frac{1}{q}} n}{8\|g\|_1^2} \right). \tag{91}$$

Plugging Eq. (91) into Eq. (87), we thus have

$$\Pr_{\mathbf{X}} \left\{ \left| \hat{f}^{\mathrm{ntk}_2}(\boldsymbol{x}) - f(\boldsymbol{x}) \right| \geq n^{\frac{1}{2}(1 - \frac{1}{q})} \right\} \leq 2e^2 \exp\left(- \frac{p^{\frac{1}{q}} n}{8\|g\|_1^2} \right).$$

□

I. Proof of Theorem 1

We first prove a useful lemma.

Lemma 46. *If $\|g\|_1 < 1$, then for any $\boldsymbol{x}$, we must have*

$$\int_{S^{d-1}}\int_{S^{d-1}} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})d\mu(\boldsymbol{z})d\lambda(\boldsymbol{v}) = \int_{S^{d-1}} \boldsymbol{x}^T\boldsymbol{z}\frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}g(\boldsymbol{z})d\mu(\boldsymbol{z}).$$

Proof. This follows from Fubini's Theorem and by a change of order of the integral. Specifically, because $\|g\|_1 < 1$, we have

$$\int_{S^{d-1}} |g(\boldsymbol{z})|d\mu(\boldsymbol{z}) < 1.$$

Thus, we have

$$\int_{S^{d-1}\times S^{d-1}} |g(\boldsymbol{z})|d\mu(\boldsymbol{z})\lambda(\boldsymbol{v}) < 1.$$

Because $\left|\boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}\right| \le 1$ when $\boldsymbol{x} \in S^{d-1}$ and $\boldsymbol{z} \in S^{d-1}$, we have

$$\int_{S^{d-1}\times S^{d-1}} \left|\boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})\right| d\mu(\boldsymbol{z})\lambda(\boldsymbol{v}) \le \int_{S^{d-1}\times S^{d-1}} |g(\boldsymbol{z})|d\mu(\boldsymbol{z})\lambda(\boldsymbol{v}) < 1.$$

Thus, by Fubini's theorem, we can exchange the sequence of integral, i.e., we have

$$\begin{aligned}
&\int_{S^{d-1}}\int_{S^{d-1}} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})d\mu(\boldsymbol{z})d\lambda(\boldsymbol{v}) \\
=&\int_{S^{d-1}}\int_{S^{d-1}} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})d\lambda(\boldsymbol{v})d\mu(\boldsymbol{z}) \\
=&\int_{S^{d-1}}\left(\int_{S^{d-1}} \mathbf{1}_{\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}d\lambda(\boldsymbol{v})\right)\boldsymbol{x}^T\boldsymbol{z}g(\boldsymbol{z})d\mu(\boldsymbol{z}) \\
=&\int_{S^{d-1}} \boldsymbol{x}^T\boldsymbol{z}\frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}g(\boldsymbol{z})d\mu(\boldsymbol{z}) \text{ (by Lemma 17).}
\end{aligned}$$

□

The following proposition characterizes generalization performance when $\epsilon = 0$, i.e., the bias term in Eq. (18).

Proposition 47. *Assume no noise ($\epsilon = 0$), a ground truth $f = f_g \in F^{\ell_2}$ where $\|g\|_1 < 1$, $n \ge 2$, $m \in \left[1, \frac{\ln n}{\ln\frac{\pi}{2}}\right]$, $d \ge n^4$, and $p \ge 6J_m(n,d)\ln\left(4n^{1+\frac{1}{m}}\right)$. Then, for any $q \in [1, 1)$ and for almost every $\boldsymbol{x} \in S^{d-1}$, we must have*

$$\begin{aligned}
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{|\hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x})| \le n^{-\frac{1}{2}(1-\frac{1}{q})}\right. \\
&\left. + \left(1 + \sqrt{J_m(n,d)n}\right)p^{-\frac{1}{2}(1-\frac{1}{q})}\right\} \\
&\ge 2e^2\left(\exp\left(-\frac{p^{\frac{1}{q}}n}{8\|g\|_1^2}\right) + \exp\left(-\frac{p^{\frac{1}{q}}p}{8\|g\|_1^2}\right)\right. \\
&\left. + \exp\left(-\frac{p^{\frac{1}{q}}p}{8n\|g\|_1^2}\right)\right) + \frac{2}{m}\frac{p}{n}.
\end{aligned}$$

Proof. We split the whole proof into 5 steps as follows.

Step 1: use pseudo ground-truth as a "intermediary"

Recall Definition 2 where we define the pseudo ground-truth $f^g_{\mathbf{V}_0}$. We then define the output of the pseudo ground-truth for training input as

$$\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) := [f^g_{\mathbf{V}_0}(\mathbf{X}_1)\ f^g_{\mathbf{V}_0}(\mathbf{X}_2) \quad \cdots \quad f^g_{\mathbf{V}_0}(\mathbf{X}_n)]^T.$$

The rest of the proof will use the pseudo ground-truth as a "intermediary" to connect the ground-truth f and the model output $\hat{f}^{\ell_2}$. Specifically, we have

$$\begin{aligned}
\hat{f}^{\ell_2}(\boldsymbol{x}) &= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}} \quad \mathbf{V}^{\ell_2} \\
&= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\mathbf{F}(\mathbf{X}) \text{ (by Eq. (17) and } \boldsymbol{\epsilon} = \mathbf{0}) \\
&= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) + \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\left(\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right). \qquad (92)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
& j\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x})j \\
&= \left|\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f^g_{\mathbf{V}_0}(\boldsymbol{x}) + f^g_{\mathbf{V}_0}(\boldsymbol{x}) \quad f(\boldsymbol{x})\right| \\
&= \left|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad f^g_{\mathbf{V}_0}(\boldsymbol{x}) + \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\left(\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right)\right. \\
&\quad \left. + f^g_{\mathbf{V}_0}(\boldsymbol{x}) \quad f(\boldsymbol{x})\right| \text{ (by Eq. (92))} \\
&\quad \underbrace{\left|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad f^g_{\mathbf{V}_0}(\boldsymbol{x})\right|}_{\text{term } A} + \underbrace{\left|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T) \quad {}^1\left(\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right)\right|}_{\text{term } B} \\
&\quad + \underbrace{\left|f^g_{\mathbf{V}_0}(\boldsymbol{x}) \quad f(\boldsymbol{x})\right|}_{\text{term } C}. \qquad (93)
\end{aligned}$$

In Eq. (93), we can see that the term A corresponds to the test error of the pseudo ground-truth, the term B corresponds to the impact of the difference between the pseudo ground-truth and the real ground-truth in the training data, and the term C corresponds to the impact of the difference between pseudo ground-truth and real ground-truth in the test data. Using the terminology of bias-variance decomposition, we refer to term A as the "pseudo bias" and term B as the "pseudo variance".

Step 2: estimate term A

We have

$$\begin{aligned}
\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\text{term } A \quad n \quad {}^{\frac{1}{2}}(1 \quad {}^{\frac{1}{q}})\right\} &= \int_{\mathbf{V}_0 2\,\mathrm{R}^{dp}} \Pr_{\mathbf{X}}\left\{\text{term } A \quad n \quad {}^{\frac{1}{2}}(1 \quad {}^{\frac{1}{q}}) \,\middle|\, \mathbf{V}_0\right\} d\lambda(\mathbf{V}_0) \\
&\int_{\mathbf{V}_0 2\,\mathrm{R}^{dp}} 2e^2 \exp\left(\quad \frac{p_q \overline{n}}{8kgk_1^2}\right) d\lambda(\mathbf{V}_0) \text{ (by Proposition 5)} \\
&= 2e^2 \exp\left(\quad \frac{p_q \overline{n}}{8kgk_1^2}\right). \qquad (94)
\end{aligned}$$

Step 3: estimate term C

For all $j = 1, 2, \quad , p$, define

$$K^{\boldsymbol{x}}_j := \int_{S^d \ 1} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{f\boldsymbol{z}^T\mathbf{V}_0[j]>0,\ \boldsymbol{x}^T\mathbf{V}_0[j]>0}g(\boldsymbol{z})d\mu(\boldsymbol{z}).$$

We now show that $K^{\boldsymbol{x}}_j$ is bounded and with mean equal to f_g, where $f_g = \int_{S^d \ 1} \boldsymbol{x}^T\boldsymbol{z}\frac{\pi \ \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}g(\boldsymbol{z})d\mu(\boldsymbol{z})$ defined by Definition 1. Specifically, we have

$$\begin{aligned}
\mathop{\mathsf{E}}_{\mathbf{V}_0} K^{\boldsymbol{x}}_j &= \mathop{\mathsf{E}}_{\boldsymbol{v}\ \tilde{\lambda}}\left(\int_{S^d \ 1} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{f\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})d\mu(\boldsymbol{z})\right) \\
&= \int_{S^d \ 1}\int_{S^d \ 1} \boldsymbol{x}^T\boldsymbol{z}\mathbf{1}_{f\boldsymbol{z}^T\boldsymbol{v}>0,\ \boldsymbol{x}^T\boldsymbol{v}>0}g(\boldsymbol{z})d\mu(\boldsymbol{z})d\tilde{\lambda}(\boldsymbol{v}) \\
&= \int_{S^d \ 1} \boldsymbol{x}^T\boldsymbol{z}\frac{\pi \quad \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}g(\boldsymbol{z})d\mu(\boldsymbol{z}) \text{ (by Lemma 46)} \\
&= f_g(\boldsymbol{x}) \text{ (by Definition 1).} \qquad (95)
\end{aligned}$$

From Definition 2, we have

$$\begin{aligned} f^g_{\mathbf{V}_0}(\boldsymbol{x}) &= \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{j C^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}} j}{p} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \text{ (by Definition 2)} \\ &= \frac{1}{p} \sum_{j=1}^{p} \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{f \boldsymbol{z}^T \mathbf{V}_0[j] > 0,\ \boldsymbol{x}^T \mathbf{V}_0[j] > 0 g} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \text{ (by Eq. (6))} \\ &= \frac{1}{p} \sum_{j=1}^{p} K^{\boldsymbol{x}}_j. \end{aligned} \tag{96}$$

Because $\mathbf{V}_0[j]$'s are *i.i.d.*, $K^{\boldsymbol{x}}_j$'s are also *i.i.d.*. Thus, we have

$$\mathop{\mathsf{E}}_{\mathbf{V}_0} f^g_{\mathbf{V}_0}(\boldsymbol{x}) = f_g(\boldsymbol{x}). \tag{97}$$

Further, for any $j \ 2 \ f1, 2, \quad , pg$, we have

$$\begin{aligned} k K^{\boldsymbol{x}}_j k_2 &= j K^{\boldsymbol{x}}_j j \text{ (because } K^{\boldsymbol{x}}_j \text{ is a scalar)} \\ &= \left| \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{f \boldsymbol{z}^T \mathbf{V}_0[j] > 0,\ \boldsymbol{x}^T \mathbf{V}_0[j] > 0 g} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \right| \\ &\quad \int_{S^{d-1}} \left| \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{f \boldsymbol{z}^T \mathbf{V}_0[j] > 0,\ \boldsymbol{x}^T \mathbf{V}_0[j] > 0 g} g(\boldsymbol{z}) \right| d\mu(\boldsymbol{z}) \\ &\quad \int_{S^{d-1}} \left| \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{f \boldsymbol{z}^T \mathbf{V}_0[j] > 0,\ \boldsymbol{x}^T \mathbf{V}_0[j] > 0 g} \right| \ j g(\boldsymbol{z}) j \, d\mu(\boldsymbol{z}) \\ &\quad \int_{S^{d-1}} j g(\boldsymbol{z}) j \, d\mu(\boldsymbol{z}) \\ &= k g k_1. \end{aligned} \tag{98}$$

Thus, by Lemma 16, we have

$$\Pr_{\mathbf{V}_0} \left\{ \left\| \left(\frac{1}{p} \sum_{j=1}^{p} K^{\boldsymbol{x}}_j \right) \quad \mathop{\mathsf{E}}_{\mathbf{V}_0} K_1 \right\|_2 \quad p^{\frac{1}{2}(1 \quad \frac{1}{q})} \right\} \quad 2e^2 \exp\left(\quad \frac{p_q \overline{p}}{8 k g k_1^2} \right).$$

Further, by Eq. (96) and Eq. (95), we have

$$\Pr_{\mathbf{V}_0} \left\{ \left| f^g_{\mathbf{V}_0}(\boldsymbol{x}) \quad f_g(\boldsymbol{x}) \right| \quad p^{\frac{1}{2}(1 \quad \frac{1}{q})} \right\} \quad 2e^2 \exp\left(\quad \frac{p_q \overline{p}}{8 k g k_1^2} \right).$$

Because $f \overset{\text{a.e.}}{=} f_g$, we have

$$\Pr_{\mathbf{V}_0} \left\{ \left| f^g_{\mathbf{V}_0}(\boldsymbol{x}) \quad f(\boldsymbol{x}) \right| \quad p^{\frac{1}{2}(1 \quad \frac{1}{q})} \right\} \quad 2e^2 \exp\left(\quad \frac{p_q \overline{p}}{8 k g k_1^2} \right).$$

Because $f^g_{\mathbf{V}_0}$ does not change with $\mathbf{X}$, we thus have

$$\Pr_{\mathbf{V}_0, \mathbf{X}} \left\{ \text{term } C \quad p^{\frac{1}{2}(1 \quad \frac{1}{q})} \right\} \quad 2e^2 \exp\left(\quad \frac{p_q \overline{p}}{8 k g k_1^2} \right). \tag{99}$$

Step 4: estimate term B

Our idea is to treat $\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})$ as a special form of noise, and then apply Proposition 4. We first bound the magnitude of this special noise. For $j = 1, 2, \quad , p$, we define

$$\mathbf{K}_j := [K^{\mathbf{X}_1}_j \ K^{\mathbf{X}_2}_j \quad K^{\mathbf{X}_n}_j]^T.$$

Then, we have

$$\|\mathbf{K}_j\|_2 = \sqrt{\sum_{i=1}^{n} \|K_j^{\mathbf{x}_i}\|_2^2} \quad p\overline{n}\|g\|_1 \text{ (by Eq. (98)).}$$

Similar to how we get Eq. (99) in Step 3, we have

$$\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right\|_2 \quad p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \quad 2e^2 \exp\left(\ \frac{p_q \overline{p}}{8n\|g\|_1^2}\right). \tag{100}$$

Thus, we have

$$\begin{aligned}
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } B \quad \sqrt{J_m(n,d)np} \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \\
=&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } B \quad \sqrt{J_m(n,d)np} \ ^{\frac{1}{2}(1 \ \frac{1}{q})}, \left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right\|_2 \quad p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \\
&+\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } B \quad \sqrt{J_m(n,d)np} \ ^{\frac{1}{2}(1 \ \frac{1}{q})}, \left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right\|_2 < p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \\
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right\|_2 \quad p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \\
&+\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } B \quad \sqrt{J_m(n,d)n}\left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X}) \quad \mathbf{F}(\mathbf{X})\right\|_2\right\} \\
&2e^2 \exp\left(\ \frac{p_q \overline{p}}{8n\|g\|_1^2}\right) + \frac{2}{p_m \overline{n}} \text{ (by Eq. (100) and Proposition 4).}
\end{aligned} \tag{101}$$

Step 5: estimate $\|\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x})\|$

We have

$$\begin{aligned}
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\|\hat{f}^{\ell_2}(\boldsymbol{x}) \quad f(\boldsymbol{x})\| \quad n \ ^{\frac{1}{2}(1 \ \frac{1}{q})} + \frac{1+\sqrt{J_m(n,d)n}}{p\overline{p}}\right\} \\
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } A + \text{term } B + \text{term } C \quad n \ ^{\frac{1}{2}(1 \ \frac{1}{q})} + \frac{1+\sqrt{J_m(n,d)n}}{p\overline{p}}\right\} \text{ (by Eq. (93))} \\
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\left\{\text{term } A \quad n \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \left[\ \left\{\text{ term } B \quad \sqrt{J_m(n,d)np} \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\}\right. \right. \\
&\left.\left[\ \left\{\text{ term } C \quad p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\}\right\}\right. \\
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\text{term } A \quad n \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} + \Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } B \quad \sqrt{J_m(n,d)np} \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \\
&+\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term } C \quad p \ ^{\frac{1}{2}(1 \ \frac{1}{q})}\right\} \text{ (by the union bound)} \\
&2e^2\left(\exp\left(\ \frac{p_q \overline{n}}{8\|g\|_1^2}\right) + \exp\left(\ \frac{p_q \overline{p}}{8\|g\|_1^2}\right) + \exp\left(\ \frac{p_q \overline{p}}{8n\|g\|_1^2}\right)\right) + \frac{2}{p_m \overline{n}} \\
&\text{(by Eqs. (94)(99)(101)).}
\end{aligned}$$

The last step exactly gives the conclusion of this proposition.

□

Theorem 1 thus follows by Proposition 4, Proposition 47, Eq. (18), and the union bound.

J. Proof of Proposition 2 (lower bound for ground-truth functions outside $\overline{F^{\ell_2}}$)

We first show what $f_1^{\ell_2}$ looks like. Define $\mathbf{H}^1 \ 2 \ \mathbb{R}^{n \ n}$ where its (i,j)-th element is

$$\mathbf{H}^1_{i,j} = \mathbf{X}_i^T \mathbf{X}_j \frac{\pi \quad \arccos(\mathbf{X}_i^T \mathbf{X}_j)}{2\pi}.$$

Notice that

$$\left(\frac{\mathbf{H}\mathbf{H}^T}{p}\right)_{i,j} = \frac{1}{p}\sum_{k=1}^{p}\mathbf{X}_i^T\mathbf{X}_j\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[k]>0,\mathbf{X}_j^T\mathbf{V}_0[k]>0\}} = \mathbf{X}_i^T\mathbf{X}_j\frac{|\mathcal{C}_{\mathbf{X}_i,\mathbf{X}_j}^{\mathbf{V}_0}|}{p}.$$

By Lemma 21, we have that $\left(\frac{\mathbf{H}\mathbf{H}^T}{p}\right)_{i,j}$ converges in probability to $(\mathbf{H}^{\infty})_{i,j}$ as $p \to \infty$ uniformly in i,j. In other words,

$$\max_{i,j}\left|\left(\frac{\mathbf{H}\mathbf{H}^T}{p}\right)_{i,j} - (\mathbf{H}^{\infty})_{i,j}\right| \xrightarrow{p} 0, \text{ as } p \to \infty. \tag{102}$$

Let $\{\boldsymbol{e}_i\}_{i=1}^n$ denote the standard basis in $\mathbb{R}^n$. For $i = 1, 2, \ldots, n$, define

$$g_{i,p} := np\boldsymbol{e}_i^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{y}, \tag{103}$$

which is a number. Further, define

$$[g_{1,p}\ g_{2,p}\ \ldots\ g_{n,p}]^T = np(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{y}.$$

Further, define the number

$$g_{i,\infty} := n\boldsymbol{e}_i^T(\mathbf{H}^{\infty})^{-1}\boldsymbol{y},$$

and

$$[g_{1,\infty}\ g_{2,\infty}\ \ldots\ g_{n,\infty}]^T = n(\mathbf{H}^{\infty})^{-1}\boldsymbol{y}.$$

Notice that $(\mathbf{H}^{\infty})^{-1}$ exists because of Eq. (102) and Lemma 7.

By Eq. (102), we have

$$\max_{i\in\{1,2,\ldots,n\}}|g_{i,p} - g_{i,\infty}| \xrightarrow{p} 0, \text{ as } p \to \infty. \tag{104}$$

For any given $\mathbf{X}$, we define $f_{\infty}^{(\ell_2)}(\cdot): S^{d-1} \mapsto \mathbb{R}$ as

$$f_{\infty}^{(\ell_2)}(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^T\mathbf{X}_i\frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi}g_{i,\infty}. \tag{105}$$

By the definition of the Dirac delta function $\delta_a(\cdot)$ with peak position at a, we can write $f_{\infty}^{(\ell_2)}(\boldsymbol{x})$ as an integral

$$f_{\infty}^{(\ell_2)}(\boldsymbol{x}) = \int_{S^{d-1}}\boldsymbol{x}^T\boldsymbol{z}\frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}\frac{1}{n}\sum_{i=1}^{n}g_{i,\infty}\,\delta_{\mathbf{X}_i}(\boldsymbol{z})d\mu(\boldsymbol{z}).$$

Notice that $g_{i,\infty}$ only depends on the training data and does not change with p (and thus is finite). Therefore, we have $f_{\infty}^{(\ell_2)} \in F^{(\ell_2)}$. It remains to show why $f^{(\ell_2)}$ converges to $f_{\infty}^{(\ell_2)}$ in probability. The following lemma shows what $f^{(\ell_2)}$ looks like.

Lemma 48. $f^{(\ell_2)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^T\mathbf{X}_i\frac{|\mathcal{C}_{\mathbf{X}_i,\boldsymbol{x}}^{\mathbf{V}_0}|}{p}g_{i,p} = \int_{S^{d-1}}\boldsymbol{x}^T\boldsymbol{z}\frac{|\mathcal{C}_{\boldsymbol{z},\boldsymbol{x}}^{\mathbf{V}_0}|}{p}\frac{1}{n}\sum_{i=1}^{n}g_{i,p}\delta_{\mathbf{X}_i}(\boldsymbol{z})d\mu(\boldsymbol{z})$.

Proof. For any $\boldsymbol{x} \in S^{d-1}$, we have

$$\begin{aligned}
f^{(\ell_2)}(\boldsymbol{x}) &= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{V}^{\ell_2} \\
&= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{y} \text{ (by Eq. (3))} \\
&= \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\sum_{i=1}^{n}\mathbf{H}_i^T\boldsymbol{e}_i^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{y} \\
&= \frac{1}{np}\sum_{i=1}^{n}\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}_i^T g_{i,p} \text{ (by Eq. (103))} \\
&= \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\boldsymbol{x}^T\mathbf{X}_i\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0,\ \boldsymbol{x}^T\mathbf{V}_0[j]>0\}}g_{i,p}.
\end{aligned}$$

By Eq. (6), we thus have

$$\hat{f}^{\ell_2}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^T \mathbf{X}_i \frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p}. \tag{106}$$

By the definition of the Dirac delta function, we have

$$\hat{f}^{\ell_2}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^T \mathbf{X}_i \frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p} = \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{|C^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}|}{p} \frac{1}{n}\sum_{i=1}^{n} g_{i,p}\delta_{\mathbf{X}_i}(\boldsymbol{z}) d\mu(\boldsymbol{z}).$$

□

Now we are ready to prove the statement of Proposition 2, i.e., uniformly over all $\boldsymbol{x} \in S^{d-1}$, $\hat{f}^{\ell_2}(\boldsymbol{x}) \xrightarrow{P} \hat{f}^{\ell_2}_1(\boldsymbol{x})$ as $p \to \infty$ (notice that we have already shown that $\hat{f}^{\ell_2}_1 \in F^{\ell_2}$). To be more specific, we restate that uniform convergence as the following lemma.

Lemma 49. *For any given* $\mathbf{X}$, $\sup_{\boldsymbol{x}\in S^{d-1}} |\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}^{\ell_2}_1(\boldsymbol{x})| \xrightarrow{P} 0$ *as* $p \to \infty$.

Proof. For any $\zeta > 0$, define two events:

$$J_1 := \left\{ \sup_{\boldsymbol{x},\boldsymbol{z}\in S^{d-1}} \left| \frac{|C^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} \right| < \zeta \right\},$$
$$J_2 := \left\{ \max_{i\in\{1,2,\ldots,n\}} |g_{i,p} - g_{i,1}| < \zeta \right\}.$$

By Lemma 21, there exists a threshold p_0 such that for any $p > p_0$,

$$\Pr[J_1] > 1 - \zeta.$$

By Eq. (104), there exists a threshold p_1 such that for any $p > p_1$,

$$\Pr[J_2] > 1 - \zeta.$$

Thus, by the union bound, when $p > \max\{p_0, p_1\}$, we have

$$\Pr[J_1 \cap J_2] > 1 - 2\zeta. \tag{107}$$

When $J_1 \cap J_2$ happens, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in S^{d-1}} |\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}^{\ell_2}_1(\boldsymbol{x})| \\
=& \sup_{\boldsymbol{x}\in S^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^T \mathbf{X}_i \left(\frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} g_{i,1} \right) \right| \\
&\text{(by Lemma 48 and Eq. (105))} \\
&\sup_{\boldsymbol{x}\in S^{d-1}, i\in\{1,2,\ldots,n\}} \left| \left(\frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} g_{i,1} \right) \right| \text{ (because } |\boldsymbol{x}^T\mathbf{X}_i| \le 1) \\
=& \sup_{\boldsymbol{x}\in S^{d-1}, i\in\{1,2,\ldots,n\}} \left| \left(\frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \right) g_{i,1} + (g_{i,p} - g_{i,1}) \frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \right| \\
&\sup_{\boldsymbol{x}\in S^{d-1}, i\in\{1,2,\ldots,n\}} \left| \left(\frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \right) g_{i,1} \right| + \left| (g_{i,p} - g_{i,1}) \frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \right| \\
&\zeta \left(\max_i |g_{i,1}| + 1 \right) \text{ (because } J_1 \cap J_2 \text{ happens, } \frac{|C^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \in [0,1] \text{, and } \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \in [0, 0.5]).
\end{aligned}$$

Because $\max_i |g_{i,1}|$ is fixed when $\mathbf{X}$ is given, $\zeta (\max_i |g_{i,1}| + 1)$ can be arbitrarily small as long as ζ is small enough. The conclusion of this lemma thus follows by Eq. (107). □

If the ground-truth function $f \notin \overline{F^{\ell_2}}$ (or equivalently, $D(f, F^{\ell_2}) > 0$), then the MSE of $\hat{f}_1^{\ell_2}$ (with respect to the ground-truth function f) is at least $D(f, F^{\ell_2})$ (because $\hat{f}_1^{\ell_2} \in F^{\ell_2}$). Therefore, we have proved Proposition 2. Below we state an even stronger result than part (ii) of Proposition 2, i.e., it captures not only the MSE of $\hat{f}_1^{\ell_2}$, but also that of $\hat{f}^{\ell_2}$ for sufficiently large p.

Lemma 50. *For any given* $\mathbf{X}$ *and* $\zeta > 0$*, there exists a threshold* p_0 *such that for all* $p > p_0$, $\Pr_f^p\left[\overline{MSE} \leq D(f, F^{\ell_2}) + \zeta\right] > 1 - \zeta$.

Proof. By Lemma 49, for any $\zeta > 0$, there must exist a threshold p_0 such that for all $p > p_0$,

$$\Pr\left\{\sup_{\boldsymbol{x} \in S^{d-1}} |\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}_1^{\ell_2}(\boldsymbol{x})| < \zeta\right\} > 1 - \zeta.$$

When $\sup_{\boldsymbol{x} \in S^{d-1}} |\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}_1^{\ell_2}(\boldsymbol{x})| < \zeta$, we have

$$D(\hat{f}^{\ell_2}, \hat{f}_1^{\ell_2}) = \sqrt{\int_{S^{d-1}} \left(\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}_1^{\ell_2}(\boldsymbol{x})\right)^2 d\mu(\boldsymbol{x})} \leq \zeta.$$

Because $\hat{f}_1^{\ell_2} \in F^{\ell_2}$, we have $D(\hat{f}_1^{\ell_2}, f) \geq D(f, F^{\ell_2})$. Thus, by the triangle inequality, we have $D(f, \hat{f}^{\ell_2}) \geq D(f, \hat{f}_1^{\ell_2}) - D(\hat{f}^{\ell_2}, \hat{f}_1^{\ell_2}) \geq D(f, F^{\ell_2}) - \zeta$. Putting these together, we have

$$\Pr\left\{D(f, \hat{f}^{\ell_2}) \geq D(f, F^{\ell_2}) - \zeta\right\} > 1 - \zeta.$$

Notice that $\mathrm{MSE} = (D(f, \hat{f}^{\ell_2}))^2$. The result of this lemma thus follows. □

K. Details for Section 4 (hyper-spherical harmonics decomposition on S^{d-1})

K.1. Convolution on S^{d-1}

First, we introduce the definition of the convolution on S^{d-1}. In (Dokmanic & Petrinovic, 2009), the convolution on S^{d-1} is defined as follows.

$$f_1 * f_2(\boldsymbol{x}) := \int_{\mathrm{SO}(d)} f_1(\mathbf{S}\boldsymbol{e}) f_2(\mathbf{S}^{-1}\boldsymbol{x}) d\mathbf{S},$$

where $\mathbf{S}$ is a $d \times d$ orthogonal matrix that denotes a rotation in S^{d-1}, chosen from the set $\mathrm{SO}(d)$ of all rotations. In the following, we will show Eq. (13). To that end, we have

$$g * h(\boldsymbol{x}) = \int_{\mathrm{SO}(d)} g(\mathbf{S}\boldsymbol{e}) h(\mathbf{S}^{-1}\boldsymbol{x}) d\mathbf{S}. \tag{108}$$

Now, we replace $\mathbf{S}\boldsymbol{e}$ by $\boldsymbol{z}$. Thus, we have

$$\mathbf{S}\boldsymbol{e} = \boldsymbol{z} \Rightarrow \boldsymbol{e} = \mathbf{S}^{-1}\boldsymbol{z} \Rightarrow (\mathbf{S}^{-1}\boldsymbol{x})^T\boldsymbol{e} = (\mathbf{S}^{-1}\boldsymbol{x})^T\mathbf{S}^{-1}\boldsymbol{z} \Rightarrow (\mathbf{S}^{-1}\boldsymbol{x})^T\boldsymbol{e} = \boldsymbol{x}^T(\mathbf{S}^{-1})^T\mathbf{S}^{-1}\boldsymbol{z}.$$

Because $\mathbf{S}$ is an orthonormal matrix, we have $\mathbf{S}^T = \mathbf{S}^{-1}$. Therefore, we have $(\mathbf{S}^{-1}\boldsymbol{x})^T\boldsymbol{e} = \boldsymbol{x}^T\boldsymbol{z}$. Thus, by Eq. (14), we have

$$h(\mathbf{S}^{-1}\boldsymbol{x}) = (\mathbf{S}^{-1}\boldsymbol{x})^T\boldsymbol{e}\frac{\pi - \arccos((\mathbf{S}^{-1}\boldsymbol{x})^T\boldsymbol{e})}{2\pi} = \boldsymbol{x}^T\boldsymbol{z}\frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}. \tag{109}$$

By plugging Eq. (109) into Eq. (108), we have

$$g * h(\boldsymbol{x}) = \int_{S^{d-1}} g(\boldsymbol{z})\boldsymbol{x}^T\boldsymbol{z}\frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} d\mu(\boldsymbol{z}).$$

Eq. (13) thus follows.

The following lemma shows the intrinsic symmetry of such a convolution.

Lemma 51. *Let* $\mathbf{S} \in \mathbb{R}^{d \times d}$ *denotes any rotation in* $\mathbb{R}^d$. *If* $f(\boldsymbol{x}) \in F^{\ell_2}$, *then* $f(\mathbf{S}\boldsymbol{x}) \in F^{\ell_2}$.

Proof. Because $f(\boldsymbol{x}) \in F^{\ell_2}$, we can find g such that

$$f(\boldsymbol{x}) = \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{\pi - \arccos(\boldsymbol{x}^T \boldsymbol{z})}{2\pi} g(\boldsymbol{z}) d\mu(\boldsymbol{z}).$$

Thus, we have

$$\begin{aligned}
f(\mathbf{S}\boldsymbol{x}) &= \int_{S^{d-1}} (\mathbf{S}\boldsymbol{x})^T \boldsymbol{z} \frac{\pi - \arccos((\mathbf{S}\boldsymbol{x})^T \boldsymbol{z})}{2\pi} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \\
&= \int_{S^{d-1}} \boldsymbol{x}^T (\mathbf{S}^T \boldsymbol{z}) \frac{\pi - \arccos(\boldsymbol{x}^T (\mathbf{S}^T \boldsymbol{z}))}{2\pi} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \\
&= \int_{S^{d-1}} \boldsymbol{x}^T (\mathbf{S}^T \boldsymbol{z}) \frac{\pi - \arccos(\boldsymbol{x}^T (\mathbf{S}^T \boldsymbol{z}))}{2\pi} g(\mathbf{S}\mathbf{S}^T \boldsymbol{z}) d\mu(\boldsymbol{z}) \\
&\quad \text{(because } \mathbf{S} \text{ is a rotation, we have } \mathbf{S}\mathbf{S}^T = \mathbf{I}) \\
&= \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{\pi - \arccos(\boldsymbol{x}^T \boldsymbol{z})}{2\pi} g(\mathbf{S}\boldsymbol{z}) d\mu(\mathbf{S}\boldsymbol{z}) \text{ (replace } \mathbf{S}^T \boldsymbol{z} \text{ by } \boldsymbol{z}) \\
&= \int_{S^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \frac{\pi - \arccos(\boldsymbol{x}^T \boldsymbol{z})}{2\pi} g(\mathbf{S}\boldsymbol{z}) d\mu(\boldsymbol{z}) \text{ (by Assumption 1)}
\end{aligned}$$

The result of this lemma thus follows. □

K.2. Hyper-spherical harmonics

We follow the the conventions of hyper-spherical harmonics in (Dokmanic & Petrinovic, 2009). We express $\boldsymbol{x} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_d] \in S^{d-1}$ in a set of hyper-spherical polar coordinates as follows.

$$\begin{aligned}
\boldsymbol{x}_1 &= \sin\theta_{d-1} \sin\theta_{d-2} \cdots \sin\theta_2 \sin\theta_1, \\
\boldsymbol{x}_2 &= \sin\theta_{d-1} \sin\theta_{d-2} \cdots \sin\theta_2 \cos\theta_1, \\
\boldsymbol{x}_3 &= \sin\theta_{d-1} \sin\theta_{d-2} \cdots \cos\theta_2, \\
&\vdots \\
\boldsymbol{x}_{d-1} &= \sin\theta_{d-1} \cos\theta_{d-2}, \\
\boldsymbol{x}_d &= \cos\theta_{d-1}.
\end{aligned}$$

Notice that $\theta_1 \in [0, 2\pi)$ and $\theta_2, \theta_3, \cdots, \theta_{d-1} \in [0, \pi)$. Let $\xi = [\theta_1\ \theta_2\ \cdots\ \theta_{d-1}]$. In such coordinates, hyper-spherical harmonics are given by (Dokmanic & Petrinovic, 2009)

$$Y^l_{\mathbf{K}}(\xi) = A^l_{\mathbf{K}} \prod_{i=0}^{d-3} C^{\frac{d-i-2}{2}+k_{i+1}}_{k_i - k_{i+1}}(\cos\theta_{d-i-1}) \sin^{k_{i+1}}\theta_{d-i-1} e^{\pm jk_{d-2}\theta_1}, \tag{110}$$

where the normalization factor is

$$A^l_{\mathbf{K}} = \sqrt{\frac{1}{\left(\frac{d}{2}\right)} \prod_{i=0}^{d-3} 2^{2k_{i+1}+d-i-4} \frac{(k_i - k_{i+1})!(d-i+2k_i-2)}{\pi\,(k_i + k_{i+1} + d-i-2)} \rho^{\left(\frac{d-i-2}{2} + k_{i+1}\right)}},$$

and $C^\lambda_d(t)$ are the Gegenbauer polynomials of degree d. These Gegenbauer polynomials can be defined as the coefficients of α^n in the power-series expansion of the following function,

$$(1 - 2t\alpha + \alpha^2)^{-\lambda} = \sum_{i=0}^{\infty} C^\lambda_i(t) \alpha^i.$$

Further, the Gegenbauer polynomials can be computed by a three-term recursive relation,

$$(i+2) C^\lambda_{i+2}(t) = 2(\lambda + i + 1) t C^\lambda_{i+1}(t) - (2\lambda + i) C^\lambda_i(t), \tag{111}$$

with $C^\lambda_0(t) = 1$ and $C^\lambda_1(t) = 2\lambda t$.

K.3. Calculate $\overset{l}{\mathbf{K}}(\xi)$ where $\mathbf{K} = \mathbf{0}$

Recall that $\mathbf{K} = (k_1, k_2, \quad , k_{d\ \ 2})$ and $l = k_0$. By plugging $\mathbf{K} = \mathbf{0}$ into Eq. (110), we have

$$\overset{l}{\mathbf{0}}(\xi) = A_{\mathbf{0}}^{l} \quad C_l^{\frac{d\ \ 2}{2}}(\cos\theta_{d\ \ 1}). \tag{112}$$

The following lemma gives an explicit form of Gegenbauer polynomials.

Lemma 52.

$$C_i^{\lambda}(t) = \sum_{k=0}^{\lfloor \frac{i}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda)}{(\lambda)k!(i \quad 2k)!}(2t)^{i\ \ 2k}. \tag{113}$$

Proof. We use mathematical induction. We already know that $C_0^{\lambda}(t) = 1$ and $C_1^{\lambda}(t) = 2\lambda t$, which both satisfy Eq. (113). Suppose that $C_i^{\lambda}(t)$ and $C_{i+1}^{\lambda}(t)$ satisfy Eq. (113), i.e.,

$$C_i^{\lambda}(t) = \sum_{k=0}^{\lfloor \frac{i}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda)}{(\lambda)k!(i \quad 2k)!}(2t)^{i\ \ 2k},$$
$$C_{i+1}^{\lambda}(t) = \sum_{k=0}^{\lfloor \frac{i+1}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 1)!}(2t)^{i\ \ 2k+1}.$$

It remains to show that $C_{i+2}^{\lambda}(t)$ also satisfy Eq. (113). By Eq. (111), it suffices to show that

$$\begin{aligned}
&(i + 2) \sum_{k=0}^{\lfloor \frac{i+2}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda + 2)}{(\lambda)k!(i \quad 2k + 2)!}(2t)^{i\ \ 2k+2} \\
=&2(\lambda + i + 1)t \sum_{k=0}^{\lfloor \frac{i+1}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 1)!}(2t)^{i\ \ 2k+1} \\
&(2\lambda + i) \sum_{k=0}^{\lfloor \frac{i}{2} \rfloor} (\ \ 1)^k \frac{(i \quad k + \lambda)}{(\lambda)k!(i \quad 2k)!}(2t)^{i\ \ 2k}.
\end{aligned} \tag{114}$$

To that end, it suffices to show that the coefficients of $(2t)^{i\ \ 2k+2}$ are the same for both sides of Eq. (114), for $k = 0, 1, \quad , \lfloor \frac{i+2}{2} \rfloor$. For the first step, we verify the coefficients of $(2t)^{i\ \ 2k+2}$ for $k = 1, \quad , \lfloor \frac{i+1}{2} \rfloor$. We have

$$\begin{aligned}
&\text{coefficients of } (2t)^{i\ \ 2k+2} \text{ on the right-hand-side of Eq. (114)} \\
=&(\lambda + i + 1)(\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 1)!} \quad (2\lambda + i)(\ \ 1)^{k\ \ 1} \frac{(i \quad k + \lambda + 1)}{(\lambda)(k \quad 1)!(i \quad 2k + 2)!} \\
=&(\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 2)!}((\lambda + i + 1)(i \quad 2k + 2) + (2\lambda + i)k) \\
=&(\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 2)!}((\lambda + i + 1)(i + 2) + (2\lambda + i)k \quad 2k(\lambda + i + 1)) \\
=&(\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 2)!}((\lambda + i + 1)(i + 2) \quad k(i + 2)) \\
=&(\ \ 1)^k \frac{(i \quad k + \lambda + 1)}{(\lambda)k!(i \quad 2k + 2)!}(\lambda \quad k + i + 1)(i + 2) \\
=&(i + 2)(\ \ 1)^k \frac{(i \quad k + \lambda + 2)}{(\lambda)k!(i \quad 2k + 2)!} \\
=&\text{coefficients of } (2t)^{i\ \ 2k+2} \text{ on the left-hand-side of Eq. (114).}
\end{aligned}$$

For the second step, we verify the coefficient of $(2t)^{i \quad 2k+2}$ for $k = 0$, i.e., the coefficient of $(2t)^{i+2}$. We have

$$\begin{aligned}
&\text{coefficients of } (2t)^{i+2} \text{ on the right-hand-side of Eq. (114)} \\
=&(\lambda + i + 1)\frac{(i + \lambda + 1)}{(\lambda)(i+1)!} \\
=&(i+2)\frac{(i + 2 + \lambda)}{(\lambda)(i+2)!} \\
=&\text{coefficients of } (2t)^{i+2} \text{ on the left-hand-side of Eq. (114).}
\end{aligned}$$

For the third step, we verify the coefficient of $(2t)^{i \quad 2k+2}$ for $k = b\frac{i+2}{2}c = b\frac{i}{2}c + 1$. We consider two cases: 1) i is even, and 2) i is odd. When i is even, we have $b\frac{i}{2}c + 1 = \frac{i}{2} + 1$, i.e., $i \quad 2k + 2 = 0$. Thus, we have

$$\begin{aligned}
&\text{coefficients of } (2t)^{0} \text{ on the right-hand-side of Eq. (114)} \\
=& \quad (2\lambda + i)(\quad 1)^{\frac{i}{2}} \frac{\left(\frac{i}{2} + \lambda\right)}{(\lambda)\left(\frac{i}{2}\right)!} \\
=&(i+2)(\quad 1)^{\frac{i}{2}+1} \frac{\left(\frac{i}{2} + 1 + \lambda\right)}{(\lambda)\left(\frac{i}{2} + 1\right)!} \\
=&\text{coefficients of } (2t)^{0} \text{ on the left-hand-side of Eq. (114).}
\end{aligned}$$

When i is odd, we have $k = b\frac{i}{2}c + 1 = \frac{i+1}{2} = b\frac{i+1}{2}c$ and this case has already been verified in the first step.

In conclusion, the coefficients of $(2t)^{i \quad 2k+2}$ are the same for both sides of Eq. (114), for $k = 0, 1, \quad , b\frac{i+2}{2}c$. Thus, by mathematical induction, the result of this lemma thus follows. □

Applying Lemma 52 in Eq. (112), we have

$$\begin{aligned}
{}^{l}_{\mathbf{0}}(\xi) = A^{l}_{\mathbf{0}} \sum_{k=0}^{b\frac{l}{2}c} (\quad 1)^{k} \frac{(l \quad k + \frac{d \quad 2}{2})}{(\frac{d \quad 2}{2})k!(l \quad 2k)!} (2\cos\theta_{d \quad 1})^{l \quad 2k}.
\end{aligned} \tag{115}$$

We give a few examples of ${}^{l}_{\mathbf{0}}(\xi)$ as follows.

$$\begin{aligned}
{}^{0}_{\mathbf{0}}(\xi) &= A^{0}_{\mathbf{0}}, \\
{}^{1}_{\mathbf{0}}(\xi) &= A^{1}_{\mathbf{0}}(d \quad 2)\cos\theta_{d \quad 1}, \\
{}^{2}_{\mathbf{0}}(\xi) &= A^{2}_{\mathbf{0}} \frac{d \quad 2}{2} \left(d\cos^2\theta_{d \quad 1} \quad 1\right), \\
{}^{3}_{\mathbf{0}}(\xi) &= A^{3}_{\mathbf{0}} \frac{d \quad 2}{2} \quad d \quad \left(\frac{d+2}{3}\cos^3\theta_{d \quad 1} \quad \cos\theta_{d \quad 1}\right).
\end{aligned}$$

K.4. Proof of Proposition 3

Recall that

$$h(\boldsymbol{x}) := \boldsymbol{x}^T \boldsymbol{e} \frac{\pi \quad \arccos(\boldsymbol{x}^T \boldsymbol{e})}{2\pi}, \quad \boldsymbol{e} := [0\,0 \quad 0\,1]^T \; 2 \; \mathsf{R}^d.$$

Notice that $\boldsymbol{x}^T \boldsymbol{e} = \cos\theta_{d \quad 1}$. Thus, we have

$$h(\boldsymbol{x}) = \cos\theta_{d \quad 1} \frac{\pi \quad \arccos(\cos\theta_{d \quad 1})}{2\pi}.$$

The arccos function has a Taylor Series Expansion:

$$\arccos(a) = \frac{\pi}{2} \quad \sum_{i=0}^{1} \frac{(2i)!}{2^{2i}(i!)^2} \frac{a^{2i+1}}{2i+1},$$

which converges when $1 \quad a \quad 1$. Thus, we have

$$h(\boldsymbol{x}) = \frac{1}{4}\cos\theta_{d\ 1} + \frac{1}{2\pi}\sum_{i=0}^{1}\frac{(2i)!}{2^{2i}(i!)^2}\frac{\cos^{2i+2}\theta_{d\ 1}}{2i+1}. \tag{116}$$

By comparing terms of even and odd power of $\cos\theta_{d\ 1}$ in Eq. (115) and Eq. (116), we immediately see that $h(\boldsymbol{x}) \ \mathcal{B} \ {}^{l}_{0}(\boldsymbol{x})$ when $l = 1$, and $h(\boldsymbol{x}) \ ? \ {}^{l}_{0}(\boldsymbol{x})$ when $l = 3, 5, 7, \quad$. It remains to examine whether $h(\boldsymbol{x}) \ ? \ {}^{l}_{0}(\boldsymbol{x})$ or $h(\boldsymbol{x}) \ \mathcal{B} \ {}^{l}_{0}(\boldsymbol{x})$ for $l \ 2 \ f0, 1, 2, 4, 6, \quad g$. We first introduce the following lemma.

Lemma 53. *Let a and b be two non-negative integers. Define the function*

$$Q(a,b) := \int_{S^{d\ 1}} \cos^{a}(\theta_{d\ 1}) \ {}^{b}_{0}(\xi)d\mu(\boldsymbol{x}).$$

We must have

$$Q(2k, 2m)\begin{cases} > 0, \ \text{if } m \quad k, \\ = 0, \ \text{if } m > k. \end{cases} \tag{117}$$

Proof. We have

$$Q(2k,0) = \int_{S^{d\ 1}} \cos^{2k}(\theta_{d\ 1}) \ {}^{0}_{0}(\xi)d\mu(\boldsymbol{x}) = A^{0}_{0}\int_{S^{d\ 1}} \cos^{2k}(\theta_{d\ 1})d\mu(\boldsymbol{x}) > 0.$$

Thus, to finish the proof, we only need to consider the case of $m \quad 1$ in Eq. (117). We then prove by mathematical induction on the first parameter of $Q(\ ,\ )$, i.e., k in Eq. (117). When $m > 0$, we have

$$Q(0,2m) = \int_{S^{d\ 1}} {}^{2m}_{0}(\xi)d\mu(\boldsymbol{x}) = \frac{1}{A^{0}_{0}}\int_{S^{d\ 1}} {}^{0}_{0}(\xi) \ {}^{2m}_{0}(\xi)d\mu(\boldsymbol{x}) = 0$$
(by the orthogonality of the basis).

Thus, Eq. (117) holds for all m when $k = 0$. Suppose that Eq. (117) holds when $k = i$. To complete the mathematical induction, it only remains to show that Eq. (117) also holds for all m when $k = i + 1$. By Eq. (111) and Eq. (112), for any l, we have

$$\cos(\theta_{d\ 1}) \ {}^{l+1}_{0}(\xi) = \frac{(l+2)A^{l+1}_{0}}{(d+2l)A^{l+2}_{0}} \ {}^{l+2}_{0}(\xi) + \frac{(d \quad 2+l)A^{l+1}_{0}}{(d+2l)A^{l}_{0}} \ {}^{l}_{0}(\xi).$$

Thus, we have

$$Q(a+1, l+1) = q_{l,1} \ Q(a, l+2) + q_{l,2} \ Q(a,l), \tag{118}$$

where

$$q_{l,1} := \frac{(l+2)A^{l+1}_{0}}{(d+2l)A^{l+2}_{0}}, \quad q_{l,2} := \frac{(d \quad 2+l)A^{l+1}_{0}}{(d+2l)A^{l}_{0}}.$$

It is obvious that $q_{l,1} > 0$ and $q_{l,2} > 0$. Applying Eq. (118) multiple times, we have

$$Q(2i+2, 2m) = q_{2m\ 1,1} \ Q(2i+1, 2m+1) + q_{2m\ 1,2} \ Q(2i+1, 2m \quad 1), \tag{119}$$
$$Q(2i+1, 2m+1) = q_{2m,1} \ Q(2i, 2m+2) + q_{2m,2} \ Q(2i, 2m), \tag{120}$$
$$Q(2i+1, 2m \quad 1) = q_{2m\ 2,1} \ Q(2i, 2m) + q_{2m\ 2,2} \ Q(2i, 2m \quad 2). \tag{121}$$

(Notice that we have already let $m \quad 1$, so all $q_{\ ,1}, q_{\ ,2}, Q(\ ,\ )$ in those equations are well-defined.) By plugging Eq. (120) and Eq. (121) into Eq. (119), we have

$$\begin{aligned} Q(2i+2, 2m) =& q_{2m,1}q_{2m\ 1,1}Q(2i, 2m+2) + (q_{2m\ 1,1}q_{2m,2} + q_{2m\ 1,2}q_{2m\ 2,1})Q(2i, 2m) \\ &+ q_{2m\ 1,2}q_{2m\ 2,2}Q(2i, 2m \quad 2). \end{aligned} \tag{122}$$

To prove that Eq. (117) holds when $k = i + 1$ for all m, we consider two cases, Case 1: $m \quad i + 1$, and Case 2: $m > i + 1$. Notice that by the induction hypothesis, we already know that Eq. (117) holds when $k = i$ for all m.

Case 1. When $m \quad i + 1$, we have $m \quad 1 \quad i$. Thus, by the induction hypothesis for $k = i$, we have $Q(2i, 2m \quad 2) > 0$ (by $m \quad 1 \quad i$), which implies that the third term of the right-hand-side of Eq. (122) is positive. Further, by the induction hypothesis for $k = i$, we also know that $Q(2i, 2m + 2) \quad 0$ and $Q(2i, 2m) \quad 0$ (regardless of the value of m), which means that the first and the second term of Eq. (122) is non-negative. Thus, by considering all three terms in Eq. (122) together, we have $Q(2i + 2, 2m) > 0$ when $m \quad i + 1$.

Case 2. When $m > i + 1$, we have $m + 1 > i$, $m > i$, and $m \quad 1 > i$. Thus, by the induction hypothesis for $k = i$, we have $Q(2i, 2m + 2) = Q(2i, 2m) = Q(2i, 2m \quad 2) = 0$. Therefore, by Eq. (122), we have $Q(2i + 2, 2m) = 0$.

In summary, Eq. (117) holds when $k = i + 1$ for all m. The mathematical induction is completed and the result of this lemma follows. □

By Lemma 53, for all $k \quad 0$, we have

$$\begin{aligned}
&\int_{S^{d \ 1}} \frac{1}{2\pi} \sum_{i=0}^{1} \frac{(2i)!}{2^{2i}(i!)^2} \frac{\cos^{2i+2} \theta_{d \ 1} \quad {}^{2k}_{0}(\xi) d\mu(\boldsymbol{x}) \\
=& \frac{1}{2\pi} \sum_{i=0}^{1} \frac{(2i)!}{2^{2i}(i!)^2} \frac{1}{2i + 1} \int_{S^{d \ 1}} \cos^{2i+2} \theta_{d \ 1} \quad {}^{2k}_{0}(\xi) d\mu(\boldsymbol{x}) \\
>& 0.
\end{aligned}$$

Thus, by Eq. (116), we know that $h(\boldsymbol{x}) \ \boldsymbol{\theta} \quad {}^{l}_{0}(\boldsymbol{x})$ for all $l \ 2 \ f0, 2, 4, \quad g$.

K.5. A special case: when $d = 2$

When $d = 2$, $S^{d \ 1}$ denotes a unit circle. Therefore, every $\boldsymbol{x}$ corresponds to an angle $\varphi \ 2 \ [\ \pi, \ \pi]$ such that $\boldsymbol{x} = [\cos\varphi \ \sin\varphi]^T$. In this situation, the hyper-spherical harmonics are the well-known Fourier series, i.e., $1, \cos(\theta), \sin(\theta), \cos(2\theta), \sin(2\theta), \quad$. Thus, we can explicitly calculate all Fourier coefficients of h more easily.

Similarly to Appendix K.1, we first write down the convolution for $d = 2$, which is also in a simpler form. For any function $f_g \ 2 \ F^{\ell_2}$, we have

$$\begin{aligned}
f_g(\varphi) &= \frac{1}{2\pi} \int_{\varphi \ \pi}^{\varphi+\pi} \frac{\pi \quad j\theta \quad \varphi j}{2\pi} \cos(\theta \quad \varphi) g(\theta) d\theta \\
&= \frac{1}{2\pi} \int_{\pi}^{\pi} \frac{\pi \quad j\theta j}{2\pi} \cos\theta \ g(\theta + \varphi) \ d\theta \text{ (replace } \theta \text{ by } \theta \quad \varphi) \\
&= \frac{1}{2\pi} \int_{\pi}^{\pi} \frac{\pi \quad j\theta j}{2\pi} \cos\theta \ g(\varphi \quad \theta) \ d\theta \text{ (replace } \theta \text{ by } \quad \theta).
\end{aligned}$$

Define $h(\theta) := \frac{\pi \quad j\theta j}{2\pi} \cos\theta$. We then have

$$f_g(\varphi) = \frac{1}{2\pi} h(\varphi) \sim g(\varphi),$$

where $\sim$ denotes (continuous) circular convolution. Let $c_{f_g}(k), c_h(k)$ and $c_g(k)$ (where $k = \quad , \quad 1, 0, 1, \quad$) denote the (complex) Fourier series coefficients for $f_g(\varphi)$, $h(\varphi)$, and $g(\varphi)$, correspondingly. Specifically, we have

$$f_g(\varphi) = \sum_{k= \quad 1}^{1} c_{f_g}(k) e^{ik\varphi}, \quad h(\varphi) = \sum_{k= \quad 1}^{1} c_h(k) e^{ik\varphi}, \quad g(\varphi) = \sum_{k= \quad 1}^{1} c_g(k) e^{ik\varphi}.$$

Thus, we have

$$c_{f_g}(k) = c_h(k) c_g(k). \tag{123}$$

Now we calculate $c_h(k)$, i.e., the Fourier decomposition of $h(\ )$. We have

$$\begin{aligned} c_h(k) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\pi - \frac{|\theta|}{2\pi}\right)\cos\theta\, e^{-ik\theta}d\theta \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(1 - \frac{|\theta|}{\pi}\right)\frac{e^{i(k+1)\theta} + e^{i(k-1)\theta}}{2}d\theta \\ &= \frac{1}{8\pi^2}\int_{-\pi}^{\pi}|\theta|\left(e^{i(k+1)\theta} + e^{i(k-1)\theta}\right)d\theta + \frac{1}{8\pi}\int_{-\pi}^{\pi}\left(e^{i(k+1)\theta} + e^{i(k-1)\theta}\right)d\theta. \end{aligned}$$

It is easy to verify that

$$\int xe^{cx}dx = e^{cx}\left(\frac{cx-1}{c^2}\right), \quad \forall c \neq 0.$$

Thus, we have

$$\begin{aligned} c_h(1) &= \frac{1}{8\pi^2}\int_{-\pi}^{\pi}|\theta|\left(e^{i2\theta} + 1\right)d\theta + \frac{1}{4} \\ &= \frac{1}{8\pi^2}\left(\pi^2 - \int_{-\pi}^{0}\theta e^{-i2\theta}d\theta + \int_{0}^{\pi}\theta e^{-i2\theta}d\theta\right) + \frac{1}{4} \\ &= \frac{1}{8\pi^2}\left(\pi^2 + \frac{i2\pi}{4} + \frac{i2\pi}{4}\right) + \frac{1}{4} \\ &= \frac{1}{8} + \frac{1}{4} \\ &= \frac{1}{8}. \end{aligned}$$

Similarly, we have

$$c_h(-1) = \frac{1}{8}.$$

Now we consider the situation of $n \neq -1$. We have

$$\begin{aligned} \int_{-\pi}^{0}|\theta|e^{-i(k+1)\theta}d\theta &= -e^{-i(k+1)\theta}\left.\frac{i(k+1)\theta - 1}{(k+1)^2}\right|_{-\pi}^{0} = -\frac{1}{(k+1)^2} + \frac{1 - i(k+1)\pi}{(k+1)^2}e^{i(k+1)\pi}, \\ \int_{0}^{\pi}|\theta|e^{-i(k+1)\theta}d\theta &= -e^{-i(k+1)\theta}\left.\frac{i(k+1)\theta - 1}{(k+1)^2}\right|_{0}^{\pi} = -\frac{1}{(k+1)^2} + \frac{1 + i(k+1)\pi}{(k+1)^2}e^{-i(k+1)\pi}. \end{aligned}$$

Notice that $e^{-i(k+1)\pi} = e^{-i(k+1)2\pi}e^{i(k+1)\pi} = e^{i(k+1)\pi}$. Therefore, we have

$$\int_{-\pi}^{\pi}|\theta|e^{-i(k+1)\theta}d\theta = \frac{2}{(k+1)^2}\left(e^{i(k+1)\pi} - 1\right).$$

Similarly, we have

$$\int_{-\pi}^{\pi}|\theta|e^{-i(k-1)\theta}d\theta = \frac{2}{(k-1)^2}\left(e^{i(k-1)\pi} - 1\right).$$

In summary, we have

$$\begin{aligned} c_h(k) &= \begin{cases} \frac{1}{8}, & k = -1 \\ \frac{1}{4\pi^2}\left(\frac{1}{(k+1)^2} + \frac{1}{(k-1)^2}\right)\left(e^{i(k+1)\pi} - 1\right), & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{8}, & k = -1 \\ \frac{1}{2\pi^2}\left(\frac{1}{(k+1)^2} + \frac{1}{(k-1)^2}\right), & k = 0, -2, -4, \cdots \\ 0, & k = -3, -5, \cdots \end{cases} \end{aligned}$$

By Eq. (123), we thus have

$$c_{f_g}(k) = \begin{cases} \frac{1}{8}c_g(k), & k = 1 \\ \frac{1}{2\pi^2}\left(\frac{1}{(k+1)^2} + \frac{1}{(k-1)^2}\right)c_g(k), & k = 0, 2, 4, \\ 0, & k = 3, 5, \end{cases} .$$

In other words, when $d = 2$, functions in F^{ℓ_2} can only contain frequencies $0, \theta, 2\theta, 4\theta, 6\theta,$, and cannot contain other frequencies $3\theta, 5\theta, 7\theta,$.

K.6. Details of Remark 2

As we discussed in Remark 2, a ReLU activation function with bias that operates on $\boldsymbol{x} \in \mathbb{R}^{d-1}, \|\boldsymbol{x}\|_2^2 = \frac{d-1}{d}$ can be equivalently viewed as one without bias that operates on $\boldsymbol{x} \in S^{d-1}$, but with the last element of $\boldsymbol{x}$ fixed at $1/\sqrt{d}$. Note that by fixing the last element of $\boldsymbol{x} \in S^{d-1}$ at a constant $\frac{1}{\sqrt{d}}$, we essentially consider ground-truth functions with a much smaller domain $D := \left\{\boldsymbol{x} = \begin{bmatrix} \tilde{\boldsymbol{x}} \\ 1/\sqrt{d} \end{bmatrix} \mid \boldsymbol{x} \in \mathbb{R}^{d-1}, \|\boldsymbol{x}\|_2^2 = \frac{d-1}{d}\right\} \subset S^{d-1}$. Correspondingly, define a vector $\boldsymbol{a} \in \mathbb{R}^{d-1}$ and $a_0 \in \mathbb{R}$ such that $\boldsymbol{a} = \begin{bmatrix} \tilde{\boldsymbol{a}} \\ a_0 \end{bmatrix} \in \mathbb{R}^d$. We claim that for any $\boldsymbol{a} \in \mathbb{R}^d$ and for all non-negative integer l, a ground-truth function $f(\boldsymbol{x}) = (\boldsymbol{x}^T\boldsymbol{a})^l, \boldsymbol{x} \in D$ must be learnable. In other words, all polynomials can be learned in the constrained domain D. Towards this end, recall that we have already shown that polynomials (of $\boldsymbol{x} \in S^{d-1}$) to the power of $l = 0, 1, 2, 4, 6,$ are learnable. Thus, it suffices to prove that polynomials of $\boldsymbol{x} \in D$ to the power of $l = 3, 5, 7,$ can be represented by a finite sum of those to the power of $l = 0, 1, 2, 4, 6,$. The idea is to utilize the fact that the binomial expansion of $(\boldsymbol{x}^T\boldsymbol{a} + \frac{a_0}{\sqrt{d}})^l$ contains $(\boldsymbol{x}^T\boldsymbol{a})^k$ for all $k = 0, 1, 2, 3, , l$. Here we give an example for writing $(\boldsymbol{x}^T\boldsymbol{a})^3$ as a linear combination of learnable components. Other values of $l = 5, 7, 9,$ can be proved in a similar way. Notice that

$$\begin{aligned} (\boldsymbol{x}^T\boldsymbol{a})^3 =& \frac{1}{4}\left((\boldsymbol{x}^T\boldsymbol{a}+1)^4 \quad (\boldsymbol{x}^T\boldsymbol{a})^4 \quad 6(\boldsymbol{x}^T\boldsymbol{a})^2 \quad 4(\boldsymbol{x}^T\boldsymbol{a})^2 \quad 1\right) \text{ (by the binomial expansion of } (\boldsymbol{x}^T\boldsymbol{a}+1)^4) \\ =& \frac{1}{4}\left(\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ \frac{a_0}{\sqrt{d}}\end{bmatrix}\right)^4 \quad \left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right)^4 \quad 6\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right)^2 \quad 4\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right) \quad 1\right). \end{aligned} \tag{124}$$

Thus, for all $\boldsymbol{x} = \begin{bmatrix} \tilde{\boldsymbol{x}} \\ 1/\sqrt{d} \end{bmatrix}$ and $\boldsymbol{a} = \begin{bmatrix} \tilde{\boldsymbol{a}} \\ a_0 \end{bmatrix}$, we have

$$\begin{aligned} (\boldsymbol{x}^T\boldsymbol{a})^3 =& \left(\boldsymbol{x}^T\boldsymbol{a} + \frac{a_0}{\sqrt{d}}\right)^3 \\ =& (\boldsymbol{x}^T\boldsymbol{a})^3 + 3\left(\frac{a_0}{\sqrt{d}}\right)(\boldsymbol{x}^T\boldsymbol{a})^2 + 3\left(\frac{a_0}{\sqrt{d}}\right)^2(\boldsymbol{x}^T\boldsymbol{a}) + \left(\frac{a_0}{\sqrt{d}}\right)^3 \\ =& (\boldsymbol{x}^T\boldsymbol{a})^3 + 3\left(\frac{a_0}{\sqrt{d}}\right)\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right)^2 + 3\left(\frac{a_0}{\sqrt{d}}\right)^2\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right) + \left(\frac{a_0}{\sqrt{d}}\right)^3 \\ =& \frac{1}{4}\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ \frac{a_0}{\sqrt{d}}\end{bmatrix}\right)^4 \quad \frac{1}{4}\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right)^4 + \left(3\left(\frac{a_0}{\sqrt{d}}\right) \quad \frac{3}{2}\right)\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right)^2 \\ &+ \left(3\left(\frac{a_0}{\sqrt{d}}\right)^2 \quad 1\right)\left(\boldsymbol{x}^T\begin{bmatrix}\boldsymbol{a} \\ 0\end{bmatrix}\right) + \left(\left(\frac{a_0}{\sqrt{d}}\right)^3 \quad \frac{1}{4}\right) \text{ (by Eq. (124)),} \end{aligned}$$

which is a sum of 5 learnable components (corresponding to the polynomials with power of 4, 4, 2, 1, and 0, respectively).

L. Discussion when g is a δ-function ($\|g\|_1 = 1$)

We now discuss what happens to the conclusion of Theorem 1 if g contains a δ-function, in which case $\|g\|_1 = 1$. In Eq. (10) of Theorem 1, only Term 1 and Term 4 (come from Proposition 5) will be affected when $\|g\|_1 = 1$. That is because only Proposition 5 requires $\|g\|_1 < 1$ during the proof of Theorem 1. To accommodate the situation when g contains a δ-function ($\|g\|_1 = 1$), we need a new version of Proposition 5. In other words, we need to know the performance of the overfitted NTK solution in learning the pseudo ground-truth when $\|g\|_1 = 1$.

Without loss of generality, we consider the situation that $g = \delta_{\boldsymbol{z}_0}$. We have the following proposition.

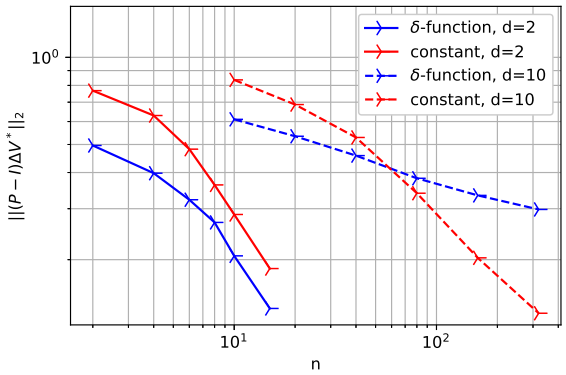

Figure 7. The curves of the model error $k(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} \; k_2$ for learning the pseudo ground-truth $f^g_{\mathbf{V}_0}$ with respect to n for different g and different d, where $p = 20000$, and $\epsilon = 0$. Every curve is the average of 10 random simulation runs.

Proposition 54. *If the ground-truth function is $f = f^g_{\mathbf{V}_0}$ in Definition 2 with $g = \delta_{z_0}$ and $\epsilon = 0$, for any $x \; 2 \; S^{d \;\; 1}$ and $q \; 2 \; (1, \;\; 1)$, we have*

$$\Pr_{\mathbf{X},\mathbf{V}_0}\left\{ j\hat{f}^{\ell_2}(x) \quad f(x)j \quad \left(\sqrt{\frac{3}{4}+\frac{\pi^2}{2}}\right)\left((d \quad 1)B(\frac{d}{2}\;\; 1,\frac{1}{2})\right)^{\frac{1}{2(d\;\; 1)}} n \;\; {}^{\frac{1}{2(d\;\; 1)}(1 \;\; \frac{1}{q})} \right\}$$
$$1 \quad \exp\left(\;\; n^{\frac{1}{q}}\right) \quad 2\exp\left(\;\; \frac{p}{24}\left((d \quad 1)B(\frac{d}{2}\;\; 1,\frac{1}{2})\right)^{\frac{1}{d\;\; 1}} n \;\; {}^{\frac{1}{d\;\; 1}(1 \;\; \frac{1}{q})}\right),$$

when

$$n \quad \left((d \quad 1)B(\frac{d}{2}\;\; 1,\frac{1}{2})\right)^{\frac{q}{q\;\; 1}}, \text{ i.e., } \left((d \quad 1)B(\frac{d}{2}\;\; 1,\frac{1}{2})\right) n \;\; {}^{(1 \;\; \frac{1}{q})} \quad 1. \tag{125}$$

(Estimates of $B(\frac{d\;\; 1}{2}, \frac{1}{2})$ can be found in Lemma 32.)

Proposition 54 implies that when n is large and p is much larger than $n \;\; {}^{\frac{1}{2(d\;\; 1)}(1 \;\; \frac{1}{q})}$, the test error between the pseudo ground-truth and learned result decreases with n at the speed $O(n \;\; {}^{\frac{1}{2(d\;\; 1)}(1 \;\; \frac{1}{q})})$. Further, if we let q be large, then the decreasing speed with n is almost $O(n \;\; {}^{\frac{1}{2(d\;\; 1)}})$. When $d \quad 3$, this speed is slower than $O(n \;\; {}^{\frac{1}{2}})$ described in Proposition 5 (i.e., Term 1 in Eq. (10) of Theorem 1). When $d = 2$, the decreasing speed with respect to n is $O(n \;\; {}^{\frac{1}{2}})$ for both Proposition 5 and Proposition 54. Nonetheless, Proposition 54 implies that the ground-truth functions $f_g \; 2 \; F^{\ell_2}$ is still learnable even when g is a δ-function (i.e., $kgk_1 \;\; = \;\; 1$), but the test error potentially suffers a slower convergence speed with respect to n when d is large.

In Fig. 7, we plot the curves of the model error $k(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} \; k_2$ for learning the pseudo ground-truth $f^g_{\mathbf{V}_0}$ with respect to n when $g = \delta_{z_0}$ (two blue curves) and when g is constant (two red curves). We plot both the case when $d = 2$ (two solid curves) and the case when $d = 10$ (two dashed curves). By Lemma 44, the model error $k(\mathbf{P} \quad \mathbf{I}) \quad \mathbf{V} \; k_2$ can represent the generalization performance for learning the pseudo ground-truth $f^g_{\mathbf{V}_0}$ when there is no noise. In Fig. 7, we can see that those two curves corresponding to $d = 10$ have different slopes and the other two curves corresponding to $d = 2$ have a similar slope, which confirms our prediction in the earlier paragraph (i.e., when $d = 2$ the test error will decay at the same speed regardless of whether g contains a δ-function or not, but when $d > 2$ the test error will decay more slowly when g contains a δ-function).

L.1. Proof of Proposition 54

We first show two useful lemmas.

Lemma 55. *For any $q \; 2 \; (1, \;\; 1)$, if $b \; 2 \; [n \;\; {}^{(1 \;\; 1/q)}, 1]$, then*

$$(1 \quad b)^n \quad \exp\left(\;\; n^{\frac{1}{q}}\right).$$

Proof. By Lemma 29, we have

$$\begin{aligned} & e^{\ b} \quad 1 \quad b \\ =) \ & e^{\ 1} \quad (1 \quad b)^{\frac{1}{b}} \\ =) \ & \exp\left(\quad n^{\frac{1}{q}} \right) \quad (1 \quad b)^{n^{\frac{1}{q}}/b} \\ =) \ & \exp\left(\quad n^{\frac{1}{q}} \right) \quad (1 \quad b)^n \text{ because } b \ 2 \ [n^{\ (1 \ \ 1/q)}, 1]. \end{aligned}$$

□

Lemma 56. *Consider* $\boldsymbol{x}_1 \ 2 \ S^{d \ 1}$ *where* $\varphi = \arccos(\boldsymbol{x}_1^T \boldsymbol{z}_0)$. *For any* $\theta \ 2 \ [\varphi, \pi]$, *there must exist* $\boldsymbol{x}_2 \ 2 \ S^{d \ 1}$ *such that* $\arccos(\boldsymbol{x}_2^T \boldsymbol{z}_0) = \theta$ *and*

$$C^{\mathbf{V}_0}_{\boldsymbol{x}_1, \boldsymbol{z}_0} \quad C^{\mathbf{V}_0}_{\boldsymbol{x}_2, \boldsymbol{z}_0}, \quad C^{\mathbf{V}_0}_{\boldsymbol{x}_1, \ \boldsymbol{z}_0} \quad C^{\mathbf{V}_0}_{\boldsymbol{x}_2, \ \boldsymbol{z}_0}. \tag{126}$$

We will explain the intuition of Lemma 56 in Remark 8 right after we use the lemma. We put the proof of Lemma 56 in Section L.2.

Now we are ready to prove Proposition 54. Recall $\mathbf{V}$ defined in Eq. (84). By Eq. (1) and $g = \delta_{\boldsymbol{z}_0}$, we have

$$\mathbf{V} \ = \frac{(\boldsymbol{h}_{\mathbf{V}_0, \boldsymbol{z}_0})^T}{p}.$$

Define

$$\begin{aligned} i \ &= \ \underset{i 2 f 1,2, \ \ ,n g}{\arg\min} \ k\mathbf{X}_i \quad \boldsymbol{z}_0 k_2, \\ \theta \ &= \arccos(\mathbf{X}_i^T \boldsymbol{z}_0). \end{aligned}$$

Thus, we have

$$\begin{aligned} k\mathbf{X}_i \quad \boldsymbol{z}_0 k_2 \ &= \sqrt{2 \quad 2\cos\theta} \text{ (by the law of cosines)} \\ &= 2\sin\frac{\theta}{2} \text{ (by the half angle identity)} \\ &\ \theta \text{ (by Lemma 41).} \end{aligned} \tag{127}$$

(Graphically, Eq. (127) means that a chord is not longer than the corresponding arc.)

As we discussed in the proof sketch of Proposition 5, we now construct the vector $\boldsymbol{a}$ such that $\mathbf{H}^T \boldsymbol{a}$ is close to $\mathbf{V}$. Define $\boldsymbol{a} \ 2 \ \mathsf{R}^n$ whose i-th element is

$$\boldsymbol{a}_i = \begin{cases} 1/p, & \text{if } i = i \\ 0, & \text{if } i \ 2 \ f1, 2, \ \ , n g \ n \ f i \ g \end{cases}.$$

Thus, we have $\mathbf{H}^T \boldsymbol{a} = (\boldsymbol{h}_{\mathbf{V}_0, \mathbf{X}_i \ })^T / p$. Therefore, we have

$$\begin{aligned} k\mathbf{H}^T \boldsymbol{a} \quad \mathbf{V} \ k_2^2 \ &= \sum_{j=1}^p k(\mathbf{H}^T \boldsymbol{a})[j] \quad \mathbf{V} \ [j] k_2^2 \\ &= \frac{1}{p^2} \sum_{j=1}^p \left(\mathbf{1}_{f \mathbf{X}_i^T \mathbf{V}_0[j] > 0, \boldsymbol{z}_0^T \mathbf{V}_0[j] > 0 g} k\mathbf{X}_i \quad \boldsymbol{z}_0 k_2^2 + \mathbf{1}_{f (\mathbf{X}_i^T \mathbf{V}_0[j]) (\boldsymbol{z}_0^T \mathbf{V}_0[j]) < 0 g} \right) \\ &\ \frac{1}{p^2} \left(p k\mathbf{X}_i \quad \boldsymbol{z}_0 k_2^2 + j C^{\mathbf{V}_0}_{\mathbf{X}_i \ , \boldsymbol{z}_0} j + j C^{\mathbf{V}_0}_{\mathbf{X}_i^{\ }, \ \boldsymbol{z}_0} j \right) \text{ (by Eq. (6))} \\ &\ \frac{1}{p^2} \left(p \quad (\theta \ )^2 + j C^{\mathbf{V}_0}_{\mathbf{X}_i \ , \boldsymbol{z}_0} j + j C^{\mathbf{V}_0}_{\mathbf{X}_i^{\ }, \ \boldsymbol{z}_0} j \right) \text{ (by Eq. (127)).} \end{aligned}$$

Thus, we have

$$\begin{aligned} p_{\bar{p}k\mathbf{H}a} \quad \mathbf{V}\ k_2 \quad & \sqrt{(\theta\ )^2 + \frac{jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j}{p}} \\ & \sqrt{\pi\theta\ + \frac{jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j}{p}} \text{ (because } \theta \quad \pi). \end{aligned} \tag{128}$$

Remark 7. We give a geometric interpretation of Eq. (128) when $d = 2$ by Fig. 4, where $\overrightarrow{OA}$ denotes z_0, $\overrightarrow{OB}$ denotes $\mathbf{X}_i$. Then, $jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j$ corresponds to the number of $\mathbf{V}_0[j]$'s whose direction is in the arc $\overset{\frown}{CE}$ or the arc $\overset{\frown}{FD}$, and θ corresponds to the angle $\angle AOB$. Intuitively, when n increases, $\mathbf{X}_i$ and z_0 get closer, so θ becomes smaller. At the same time, both the arc $\overset{\frown}{CE}$ and the arc $\overset{\frown}{FD}$ become shorter. Consequently, the value of Eq. (128) decreases as n increases. In the rest of the proof, we will quantitatively estimate the above relationship.

Recall C_d in Eq. (60). Define

$$\theta := \frac{\pi}{2}\left(\frac{2p\bar{2}(d\quad 1)}{C_d}\right)^{\frac{1}{d\ 1}} n^{\ \frac{1}{d\ 1}(1\ \ \frac{1}{q})}\ 2\left[0, \frac{\pi}{2}\right] \quad \text{(by Eq. (125))}. \tag{129}$$

For any $q\ 2\ (1,\ 1\ )$, we define two events:

$$\begin{aligned} J_1 &:= \left\{\frac{jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j}{p} \quad \frac{3\theta}{2\pi}\right\}, \\ J_2 &:= f\theta \quad \theta g. \end{aligned}$$

If both J_1 and J_2 happen, by Eq. (128), we must then have

$$\begin{aligned} p_{\bar{p}k\mathbf{H}a} \quad \mathbf{V}\ k_2 \quad & \left(\sqrt{\frac{3}{2\pi} + \pi}\right)\ p_{\bar{\theta}} \\ & = \left(\sqrt{\frac{3}{4} + \frac{\pi^2}{2}}\right)\left(\frac{2p\bar{2}(d\quad 1)}{C_d}\right)^{\frac{1}{2(d\ 1)}} n^{\ \frac{1}{2(d\ 1)}(1\ \ \frac{1}{q})}. \end{aligned}$$

Thus, by Lemma 44 and Lemma 45, if $f = f^g_{\mathbf{V}_0}$ and both J_1 and J_2 happen, then for any $x\ 2\ S^{d\ 1}$, we must have

$$j\hat{f}^{\nu_2}(x)\quad f(x)j \quad \left(\sqrt{\frac{3}{4} + \frac{\pi^2}{2}}\right)\left(\frac{2p\bar{2}(d\quad 1)}{C_d}\right)^{\frac{1}{2(d\ 1)}} n^{\ \frac{1}{2(d\ 1)}(1\ \ \frac{1}{q})}. \tag{130}$$

It then only remains to estimate the probability of $J_1\ \setminus\ J_2$.

Step 1: Estimate the probability of J_1 conditional on J_2.

When J_2 happens, we have $\theta\ < \theta$. By Lemma 56, we can find $x\ 2\ S^{d\ 1}$ such that the angle between x and z_0 is exactly θ and

$$\frac{jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j}{p} \quad \frac{jC^{\mathbf{V}_0}_{x,z_0}j + jC^{\mathbf{V}_0}_{x,\ z_0}j}{p}. \tag{131}$$

Remark 8. We give a geometric interpretation of Eq. (131) (i.e., Lemma 56) when $d = 2$ by Fig. 4. Recall in Remark 7 that, if we take $\overrightarrow{OA}$ as z_0 and $\overrightarrow{OB}$ as $\mathbf{X}_i$, then $jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}j + jC^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}j$ corresponds to the number of $\mathbf{V}_0[j]$'s whose direction is in the arc $\overset{\frown}{CE}$ or the arc $\overset{\frown}{FD}$. If we fix $\overrightarrow{OA}$ (i.e., z_0) and increase the angle $\angle AOB$ (corresponding to θ), then both the arc $\overset{\frown}{CE}$ and the arc $\overset{\frown}{FD}$ will become longer. In other words, if we replace $\mathbf{X}_i$ by x such that the angle θ (between z_0 and $\mathbf{X}_i$) increases to the angle θ (between z_0 and x), then $C^{\mathbf{V}_0}_{\mathbf{X}_i\ ,z_0}$ $C^{\mathbf{V}_0}_{x,z_0}$ and $C^{\mathbf{V}_0}_{\mathbf{X}_i\ ,\ z_0}$ $C^{\mathbf{V}_0}_{x,\ z_0}$, and thus Eq. (131) follows.

We next estimate the probability that the right-hand-side of Eq. (131) is greater than $\frac{3\theta}{2\pi}$. By Eq. (6), we have

$$\frac{|C^{\mathbf{V}_0}_{\boldsymbol{x},\boldsymbol{z}_0}| + |C^{\mathbf{V}_0}_{\boldsymbol{x},\ \boldsymbol{z}_0}|}{p} = \frac{1}{p}\sum_{j=1}^{p} \underbrace{\mathbf{1}_{\{\ \boldsymbol{x}^T\mathbf{V}_0[j]>0, \boldsymbol{z}_0^T\mathbf{V}_0[j]>0 \text{ OR } \boldsymbol{x}^T\mathbf{V}_0[j]>0,\ \ \boldsymbol{z}_0^T\mathbf{V}_0[j]<0\}}}_{\text{Term A}}. \tag{132}$$

Notice that the angle between $\boldsymbol{x}$ and $\boldsymbol{z}_0$ is $\pi - \theta$, and the angle between $\boldsymbol{x}$ and $-\boldsymbol{z}_0$ is also $\pi - \theta$. By Lemma 17 and Assumption 1, we know that the Term A in Eq. (132) follows Bernoulli distribution with the probability $2\ \frac{\pi - (\pi - \theta)}{2\pi} = \frac{\theta}{\pi}$. By letting $\delta = 1/2$, $a = p$, $b = \frac{\theta}{\pi}$ in Lemma 14, we have

$$\Pr_{\mathbf{V}_0}\left\{\left||C^{\mathbf{V}_0}_{\boldsymbol{x},\boldsymbol{z}_0}| + |C^{\mathbf{V}_0}_{\boldsymbol{x},\ \boldsymbol{z}_0}| - \frac{p\theta}{\pi}\right| > \frac{p\theta}{2\pi}\right\} \quad 2\exp\left(-\frac{p\theta}{12\pi}\right).$$

By Eq. (131), we then have

$$\Pr_{\mathbf{V}_0}[J_1^c \cup J_2] \quad \Pr_{\mathbf{V}_0}\left\{\frac{|C^{\mathbf{V}_0}_{\boldsymbol{x},\boldsymbol{z}_0}| + |C^{\mathbf{V}_0}_{\boldsymbol{x},\ \boldsymbol{z}_0}|}{p} > \frac{3\theta}{2\pi}\right\} \quad 2\exp\left(-\frac{p\theta}{12\pi}\right).$$

Step 2: Estimate the probability of J_2.

By Lemma 8 and Assumption 1, for any $i \in \{1, 2, \ldots, n\}$ and because $\theta \in [0, \pi/2]$, we have

$$\begin{aligned}\Pr_{\mathbf{X}}\left\{\arccos(\mathbf{X}_i^T\boldsymbol{z}_0) > \theta\right\} &= 1 - \frac{1}{2}I_{\sin^2\theta}\left(\frac{d-1}{2}, \frac{1}{2}\right)\\ &\le 1 - \frac{C_d}{2\sqrt{\pi}(d-1)}\sin^{d-1}\theta \text{ (by Lemma 35).}\end{aligned}$$

Note that since $\Pr_{\mathbf{X}}\left\{\arccos(\mathbf{X}_i^T\boldsymbol{z}_0) > \theta\right\} \ge 0$, we must have

$$\frac{C_d}{2\sqrt{\pi}(d-1)}\sin^{d-1}\theta \le 1. \tag{133}$$

Further, because all $\mathbf{X}_i$'s are *i.i.d.* for $i \in \{1, 2, \ldots, n\}$, we have

$$\Pr_{\mathbf{X}}\{\theta_{\max} > \theta\} = \Pr_{\mathbf{X}}\left\{\min_{i\in\{1,2,\ldots,n\}}\arccos(\mathbf{X}_i^T\boldsymbol{z}_0) > \theta\right\} \le \left(1 - \frac{C_d}{2\sqrt{\pi}(d-1)}\sin^{d-1}\theta\right)^n. \tag{134}$$

By Eq. (129) and Lemma 41, we then have

$$\sin\theta \le \left(\frac{2\sqrt{\pi}(d-1)}{C_d}\right)^{\frac{1}{d-1}} n^{-\frac{1}{d-1}(1-\frac{1}{q})},$$

i.e.,

$$\frac{C_d}{2\sqrt{\pi}(d-1)}\sin^{d-1}\theta \le n^{-(1-1/q)}.$$

Thus, by Eq. (133), Eq. (134), and Lemma 55, we have

$$\Pr_{\mathbf{X}}[J_2^c] = \Pr_{\mathbf{X}}\{\theta_{\max} > \theta\} \le \exp\left(-n^{\frac{1}{q}}\right).$$

Combining the results of Step 1 and Step 2, we thus have

$$\begin{aligned}
\Pr_{\mathbf{X},\mathbf{V}_0}[J_1 \setminus J_2] &= \Pr_{\mathbf{X},\mathbf{V}_0}[J_1 \mid J_2] \Pr_{\mathbf{X},\mathbf{V}_0}[J_2] \\
&= \Pr_{\mathbf{V}_0}[J_1 \mid J_2] \Pr_{\mathbf{X}}[J_2] \text{ (because of } \mathbf{V}_0 \text{ and } \mathbf{X} \text{ are independent)} \\
&\quad \left(1 - 2\exp\left(-\frac{p\theta}{12\pi}\right)\right)\left(1 - \exp\left(-n^{\frac{1}{q}}\right)\right) \\
&\quad 1 - \exp\left(-n^{\frac{1}{q}}\right) - 2\exp\left(-\frac{p\theta}{12\pi}\right) \\
&= 1 - \exp\left(-n^{\frac{1}{q}}\right) - 2\exp\left(-\frac{p}{24}\left(\frac{2^{\frac{p}{2}}(d-1)}{C_d}\right)^{\frac{1}{d-1}} n^{\frac{1}{d-1}(1-\frac{1}{q})}\right) \text{ (by Eq. (130)).}
\end{aligned}$$

By Eq. (60), the conclusion of Proposition 54 thus follows.

L.2. Proof of Lemma 56

Proof. When $\boldsymbol{x}_1 = \boldsymbol{z}_0$, the conclusion of this lemma trivially holds because $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0} = C^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0} = ?$ (because $\boldsymbol{x}^T\mathbf{V}_0[j]$ and $\boldsymbol{z}_0^T\mathbf{V}_0[j]$ cannot be both positive or negative at the same time when $\boldsymbol{x}_1 = \boldsymbol{z}_0$.). It remains to consider $\boldsymbol{x}_1 \neq \boldsymbol{z}_0$. Define

$$\boldsymbol{z}_{0,?} := \frac{\boldsymbol{x}_1 - (\boldsymbol{x}_1^T\boldsymbol{z}_0)\boldsymbol{z}_0}{\|\boldsymbol{x}_1 - (\boldsymbol{x}_1^T\boldsymbol{z}_0)\boldsymbol{z}_0\|_2}.$$

Thus, we have $\boldsymbol{z}_{0,?}^T\boldsymbol{z}_0 = 0$ and $\|\boldsymbol{z}_{0,?}\|_2 = 1$, i.e., $\boldsymbol{z}_0$ and $\boldsymbol{z}_{0,?}$ are orthonormal basis vectors on the 2D plane L spanned by $\boldsymbol{x}_1$ and $\boldsymbol{z}_0$. Thus, we can represent $\boldsymbol{x}_1$ as

$$\boldsymbol{x}_1 = \cos\varphi \cdot \boldsymbol{z}_0 + \sin\varphi \cdot \boldsymbol{z}_{0,?} \in L.$$

For any $\theta \in [\varphi, \pi]$, we construct $\boldsymbol{x}_2$ as

$$\boldsymbol{x}_2 := \cos\theta \cdot \boldsymbol{z}_0 + \sin\theta \cdot \boldsymbol{z}_{0,?} \in L.$$

In order to show $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0} \subseteq C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$, we only need to prove any $j \in C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0}$ must in $C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$. For any $\mathbf{V}_0[j]$, $j = 1, 2, \ldots, p$, define the angle $\theta_j \in [0, 2\pi]$ as the angle between $\boldsymbol{z}_0$ and $\mathbf{V}_0[j]$'s projected component $\boldsymbol{v}_j$ on L[10], i.e.,

$$\boldsymbol{v}_j = \cos\theta_j \cdot \boldsymbol{z}_0 + \sin\theta_j \cdot \boldsymbol{z}_{0,?} \in L.$$

By the proof of Lemma 17, we know that $j \in C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0}$ if and only if $\theta_j \in (\frac{\pi}{2}, \frac{\pi}{2}) \setminus (\pi + \varphi - \frac{\pi}{2}, \pi + \varphi + \frac{\pi}{2}) \pmod{2\pi}$. Similarly, $j \in C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$ if and only if $\theta_j \in (\frac{\pi}{2}, \frac{\pi}{2}) \setminus (\pi + \theta - \frac{\pi}{2}, \pi + \theta + \frac{\pi}{2}) \pmod{2\pi}$. Because $\varphi \in [0, \pi]$ and $\theta \in [\varphi, \pi]$, we have

$$(-\frac{\pi}{2}, \frac{\pi}{2}) \setminus (\pi + \varphi - \frac{\pi}{2}, \pi + \varphi + \frac{\pi}{2}) \subseteq (-\frac{\pi}{2}, \frac{\pi}{2}) \setminus (\pi + \theta - \frac{\pi}{2}, \pi + \theta + \frac{\pi}{2}) \pmod{2\pi}.$$

Thus, whenever $j \in C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0}$, we must have $j \in C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$. Therefore, we conclude that $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0} \subseteq C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$. Using a similar method, we can also show that $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0} \subseteq C^{\mathbf{V}_0}_{\boldsymbol{x}_2,-\boldsymbol{z}_0}$. The result of this lemma thus follows. □

[10] Note that such an angle θ_j is well defined as long as $\mathbf{V}_0[j]$ is not perpendicular to L. The reason that we do not need to worry about those j's such that $\mathbf{V}_0[j] \perp L$ is as follows. When $\mathbf{V}_0[j] \perp L$, we then have $\boldsymbol{x}_1^T\mathbf{V}_0[j] = \boldsymbol{x}_2^T\mathbf{V}_0[j] = \boldsymbol{z}_0^T\mathbf{V}_0[j] = 0$. Thus, those j's do not belong to any set $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,\boldsymbol{z}_0}$, $C^{\mathbf{V}_0}_{\boldsymbol{x}_2,\boldsymbol{z}_0}$, $C^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0}$, or $C^{\mathbf{V}_0}_{\boldsymbol{x}_2,-\boldsymbol{z}_0}$ in Eq. (126).