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 $\lambda(\cdot)$ on $\mathbb{R}^d$, and the directions of the initial weights (i.e., the normalized initial weights $\frac{\mathbf{V}_0[j]}{\|\mathbf{V}_0[j]\|_2}$) follows the probability density $\tilde{\lambda}(\cdot)$ on $\mathcal{S}^{d-1}$. Let $\lambda_a(\cdot)$ be the Lebesgue measure on $\mathbb{R}^a$ 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 $I_{\cdot}(\cdot,\cdot)$ denote the regularized incomplete beta function (Dutka, 1981). Let $B(\cdot,\cdot)$ 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 $\mathcal{S}^{d-1}$ as the intersection of $\mathcal{S}^{d-1}$ with an open ball in $\mathbb{R}^d$ centered at $\boldsymbol{v}_*$ with radius r, i.e.,

$$\mathcal{B}^r_{\boldsymbol{v}_*} := \left\{\boldsymbol{v}\in\mathcal{S}^{d-1} \mid \|\boldsymbol{v}-\boldsymbol{v}_*\|_2 < r\right\}. \tag{21}$$

Remark 4. For ease of exposition, we will sometimes neglect the subscript $\boldsymbol{v}_*$ of $\mathcal{B}^r_{\boldsymbol{v}_*}$ and use $\mathcal{B}^r$ 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 $\mathcal{B}^r_{\boldsymbol{v}_*}$, it only depends on r but not $\boldsymbol{v}_*$. Thus, we write $\lambda_{d-1}(\mathcal{B}^r)$ instead.

For any $\boldsymbol{x}\in\mathbb{R}^d$ such that $\boldsymbol{x}^T\boldsymbol{v}_* = 0$, define two halves of the cap $\mathcal{B}^r_{\boldsymbol{v}_*}$ as

$$\mathcal{B}^{r,\boldsymbol{x}}_{\boldsymbol{v}_*,+} := \left\{\boldsymbol{v}\in\mathcal{B}^r_{\boldsymbol{v}_*} \mid \boldsymbol{x}^T\boldsymbol{v} > 0\right\}, \quad \mathcal{B}^{r,\boldsymbol{x}}_{\boldsymbol{v}_*,-} := \left\{\boldsymbol{v}\in\mathcal{B}^r_{\boldsymbol{v}_*} \mid \boldsymbol{x}^T\boldsymbol{v} < 0\right\}. \tag{22}$$

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

$$\mathcal{A}_{\mathbf{V}_0} := \left\{\frac{\mathbf{V}_0[j]}{\|\mathbf{V}_0[j]\|_2}\ \middle|\ j\in\{1,2,\cdots,p\}\right\}. \tag{23}$$

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

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

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

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

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

C. Assumptions and Justifications

Because $\hat{f}_{\Delta\mathbf{V},\mathbf{V}_0}(a\boldsymbol{x}) = a\cdot\hat{f}_{\Delta\mathbf{V},\mathbf{V}_0}(\boldsymbol{x})$ for any $a\in\mathbb{R}$, 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 $\mathcal{S}^{d-1}$. The initial weights $\mathbf{V}_0[j]$'s are uniform in all directions. In other words, $\mu(\cdot)$ and $\tilde{\lambda}(\cdot)$ are both $\mathsf{unif}(\mathcal{S}^{d-1})$.*

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 $\mathsf{Pr}\{\mathbf{X}_i = \mathbf{X}_j\} = 0$ and $\mathsf{Pr}\{\mathbf{X}_i = -\mathbf{X}_j\} = 0$ (because the probability that a continuous random variable equals to a given value is zero). Thus, $\mathsf{Pr}\{\mathbf{X}_i \parallel \mathbf{X}_j\} = 0$, and $\mathsf{Pr}\{\mathbf{X}_i \nparallel \mathbf{X}_j\} = 1$. Similarly, we can show that $\mathsf{Pr}\{\mathbf{V}_0[k] \nparallel \mathbf{V}_0[l]\} = 1$. We thus make Assumption 3.

Assumption 3. *$\mathbf{X}_i \nparallel \mathbf{X}_j$ for any $i \neq j$, and $\mathbf{V}_0[k] \nparallel \mathbf{V}_0[l]$ 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 $\Delta \mathbf{V}^{\ell_2}$ exists).

Lemma 7. $\lim_{p\to\infty} \underset{\mathbf{V}_0}{\mathsf{Pr}} \{\mathsf{rank}(\mathbf{H}) = n \mid \mathbf{X}\} = 1.$

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 $\mathcal{S}^{d-1}$, then the area (in the measure of λ_{d-1}) of this hyper-spherical cap is*

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

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

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

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

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* $\boldsymbol{b} \in \mathcal{S}^{d-1}$ *follows uniform distribution in all directions. Given any* $\boldsymbol{a} \in \mathcal{S}^{d-1}$ *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

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

Further, by symmetry, we have

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

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

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

□

D.2. Estimation of certain norms

In this subsection, we will show $\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \leq \sqrt{p}$ 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. $\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2 \leq \sqrt{p}$ *for any* $\boldsymbol{x} \in \mathcal{S}^{d-1}$.

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

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

□

Lemma 12. *If* $\mathbf{C} = \mathbf{AB}$*, then* $\|\mathbf{C}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2$*. 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*

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

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

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

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 $\|\cdot\|_2^2$ is convex. Thus, we have

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

□

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*

$$\mathsf{Pr}\{|X - ab| > \delta ab\} \leq 2e^{-ab\delta^2/3}.$$

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

Lemma 15 (Azuma–Hoeffding inequality for random vectors)**.** *Let $X_1, X_2, \cdots, X_k$ be* i.i.d. *random vectors with zero mean (of the same dimension) in a real Euclidean space such that $\|X_i\|_2 \leq 1$ for all $i = 1, 2, \cdots, k$. Then, for every $a > 0$,*

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

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 $X_1, X_2, \cdots, X_k$ be* i.i.d. *random vectors (of the same dimension) in a real Euclidean space such that $\|X_i\|_2 \leq U$ for all $i = 1, 2, \cdots, k$. Then, for any $q \in [1, \infty)$,*

$$\mathsf{Pr}\left\{\left\|\left(\frac{1}{k}\sum_{i=1}^{k} X_i\right) - \mathsf{E}\,X_1\right\|_2 \geq k^{\frac{1}{2q}-\frac{1}{2}}\right\} < 2e^2 \exp\left(-\frac{\sqrt[q]{k}}{8U^2}\right).$$

Proof. Because $\|X_i\|_2 \leq U$, we have $\mathsf{E}\,\|X_i\|_2 \leq U$. By triangle inequality, we have $\|X_i - \mathsf{E}\,X_i\|_2 \leq \|X_i\|_2 + \mathsf{E}\,\|X_i\|_2 \leq 2U$, i.e.,

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

We also have

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

We then have

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

□

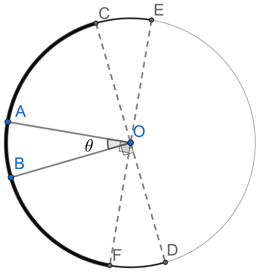

Figure 4. The arc $\overset{\frown}{\mathrm{CBF}}$ is $\frac{\pi-\theta}{2\pi}$ of the perimeter of the circle O.

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.

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

(Recall that $\tilde{\lambda}(\cdot)$ denotes the distribution of the normalized version of $\mathbf{V}_0[j]$ on $\mathcal{S}^{d-1}$ 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 $\tilde{\lambda}$ is uniform on $\mathcal{S}^{d-1}$, 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 $\{\boldsymbol{v}\in\mathcal{S}^{d-1}\mid \boldsymbol{z}^T\boldsymbol{v}>0\}$ and $\{\boldsymbol{v}\in\mathcal{S}^{d-1}\mid \boldsymbol{x}^T\boldsymbol{v}>0\}$, 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, $\overrightarrow{\mathrm{OA}}$ denotes $\boldsymbol{z}$, $\overrightarrow{\mathrm{OB}}$ denotes $\boldsymbol{x}$. Thus, the arc $\overset{\frown}{\mathrm{EAF}}$ denotes $\{\boldsymbol{v}\mid \boldsymbol{z}^T\boldsymbol{v}>0\}$, and the arc $\overset{\frown}{\mathrm{CBD}}$ denotes $\{\boldsymbol{v}\mid \boldsymbol{x}^T\boldsymbol{v}>0\}$. The intersection of $\overset{\frown}{\mathrm{EAF}}$ and $\overset{\frown}{\mathrm{CBD}}$, i.e., the arc $\overset{\frown}{\mathrm{CBF}}$, represents $\{\boldsymbol{v}\mid \boldsymbol{z}^T\boldsymbol{v}>0, \boldsymbol{x}^T\boldsymbol{v}>0\}$. Notice that the angle of $\overset{\frown}{\mathrm{CBF}}$ equals $\pi-\theta$, where θ denotes the angle between $\boldsymbol{z}$ and $\boldsymbol{x}$. Therefore, ratio of the length of $\overset{\frown}{\mathrm{CBF}}$ 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 $\overrightarrow{\mathrm{OA}}$ and the semi-sphere in the direction of $\overrightarrow{\mathrm{OB}}$. We can see that the area of the spherical lune ICHF is still proportional to the angle $\angle$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

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

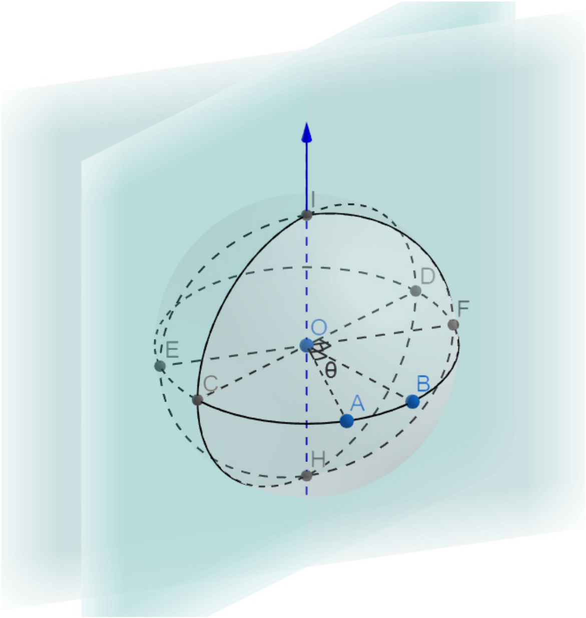

Figure 5. The area of the spherical lune ICHF is $\frac{\pi-\theta}{2\pi}$ of the area of the whole sphere.

We compute the spherical coordinates $\boldsymbol{\varphi}_{\boldsymbol{x}} = [\varphi_1^{\boldsymbol{x}}\ \varphi_2^{\boldsymbol{x}}\ \cdots\ \varphi_{d-1}^{\boldsymbol{x}}]^T$ 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 $\boldsymbol{\varphi}_{\boldsymbol{x}} = [\pi/2\ \pi/2\ \cdots\ \pi/2\ 0]^T$. Similarly, the spherical coordinates for $\boldsymbol{z}$ is $\boldsymbol{\varphi}_{\boldsymbol{z}} = [\pi/2\ \pi/2\ \cdots \pi/2\ \theta]^T$. Let the spherical coordinates for $\boldsymbol{v}$ be $\boldsymbol{\varphi}_{\boldsymbol{v}} = [\varphi_1^{\boldsymbol{v}}\ \varphi_2^{\boldsymbol{v}}\ \cdots\ \varphi_{d-1}^{\boldsymbol{v}}]^T$. 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., $\varphi_{d-1}^{\boldsymbol{v}} \in (-\pi/2,\ \pi/2) \cap (\theta - \pi/2,\ \theta + \pi/2) \pmod{2\pi}$. We have

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

The result of this lemma thus follows. □

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

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

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

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

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

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

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

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.

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

In the language of Glivenko-Cantelli class, the function class $\mathscr{F}_*$ consists of functions $\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}>0,\boldsymbol{z}^T\boldsymbol{v}>0\}}$ that map $\boldsymbol{v}\in\mathcal{S}^{d-1}$ to 0 or 1, where every $\boldsymbol{x}\in\mathcal{S}^{d-1}$ and $\boldsymbol{z}\in\mathcal{S}^{d-1}$ corresponds to a distinct function in $\mathscr{F}_*$. According to Lemma 18, we need to calculate the order of the polynomial discrimination for this $\mathscr{F}_*$. 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* $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_k \in \mathcal{S}^{d-1}$*, the number of different values (i.e., the cardinality) of the set* $\left\{\left(\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_1>0\}}, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_2>0\}}, \cdots, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_k>0\}}\right) \;\middle|\; \boldsymbol{x}\in\mathcal{S}^{d-1}\right\}$ *is at most* $Q_{d,k}$*, where*

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

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 $\mathcal{S}^{d-1}$ into, because all $\boldsymbol{x}$ in one region corresponds to the same value of the tuple $\left(\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_1>0\}}, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_2>0\}}, \cdots, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_k>0\}}\right)$. 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* $k \geq 1$ *and* $d \geq 2$*, we must have* $Q_{d,k} \leq (k+1)^{d+1}$.

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

We can now calculate the order of the polynomial discrimination for the function class $\mathscr{F}_*$. Because

$$\begin{aligned}
&\text{card}\left(\left\{\left(\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_1>0,\boldsymbol{z}^T\boldsymbol{v}_1>0\}}, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_2>0,\boldsymbol{z}^T\boldsymbol{v}_2>0\}}, \cdots, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_k>0,\boldsymbol{z}^T\boldsymbol{v}_k0\}}\right) \;\middle|\; \boldsymbol{x}\in\mathcal{S}^{d-1}, \boldsymbol{z}\in\mathcal{S}^{d-1}\right\}\right)\\
\leq&\text{card}\left(\left\{\left(\mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_1>0\}}, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_2>0\}}, \cdots, \mathbf{1}_{\{\boldsymbol{x}^T\boldsymbol{v}_k>0\}}\right) \;\middle|\; \boldsymbol{x}\in\mathcal{S}^{d-1}\right\}\right)\\
&\cdot\text{card}\left(\left\{\left(\mathbf{1}_{\{\boldsymbol{z}^T\boldsymbol{v}_1>0\}}, \mathbf{1}_{\{\boldsymbol{z}^T\boldsymbol{v}_2>0\}}, \cdots, \mathbf{1}_{\{\boldsymbol{z}^T\boldsymbol{v}_k>0\}}\right) \;\middle|\; \boldsymbol{z}\in\mathcal{S}^{d-1}\right\}\right),
\end{aligned}$$

by Lemma 19 and Lemma 20, we know that

$$\text{card}(\mathscr{F}_*(X_1^k)) \leq \left(Q_{d,k}\right)^2 \leq (k+1)^{2(d+1)}.$$

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

Thus, $\mathscr{F}_*$ has polynomial discrimination with order at most $2(d+1)$. Notice that all functions in $\mathscr{F}_*$ 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 $\mathscr{F}_*$ is Glivenko-Cantelli. In other words, we have shown the following lemma.

Lemma 21.

$$\sup_{\boldsymbol{x},\boldsymbol{z}\in\mathcal{S}^{d-1}}\left|\frac{|\mathcal{C}_{\boldsymbol{z},\boldsymbol{x}}^{\mathbf{V}_0}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi}\right| \xrightarrow{P} 0, \text{ as } p\to\infty. \tag{34}$$

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

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

The following lemma states that, given $\mathbf{X}$ (that satisfies Assumption 3) and $k \in \{1, 2, \cdots, n\}$, there always exists a vector $\boldsymbol{v}\in\mathcal{S}^{d-1}$ that is only orthogonal to one training input $\mathbf{X}_k$ but not orthogonal to other training inputs $\mathbf{X}_i$ for all $i \neq k$. 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* $k \in \{1, 2, \cdots, n\}$ *we have*

$$\mathcal{T}_k := \left\{\boldsymbol{v}\in\mathcal{S}^{d-1} \mid \boldsymbol{v}^T\mathbf{X}_k = 0, \boldsymbol{v}^T\mathbf{X}_i \neq 0, \text{for all } i \in \{1, 2, \cdots, n\}\setminus\{k\}\right\} \neq \varnothing.$$

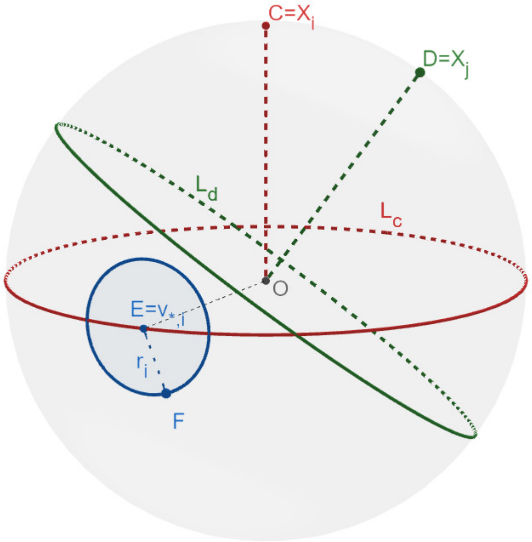

Figure 6. Geometric interpretation of $\mathcal{B}^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{*,i},+}$ and $\mathcal{B}^{r_i,\mathbf{X}_i}_{\boldsymbol{v}_{*,i},-}$ on a sphere (i.e., $\mathcal{S}^2$).

Proof. We have

$$\begin{aligned}
\mathcal{T}_k &= \mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\setminus\left(\bigcup_{i\in\{1,2,\cdots,n\}\setminus\{k\}}\ker(\mathbf{X}_i)\right)\\
&= \mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\setminus\left(\bigcup_{i\in\{1,2,\cdots,n\}\setminus\{k\}}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\cap\ker(\mathbf{X}_i)\right)\right).
\end{aligned}$$

Because

$$\begin{aligned}
&\dim(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)) = d-2,\\
&\dim(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\cap\ker(\mathbf{X}_i)) = d-3 \text{ for all } i\in\{1,2,\cdots,n\}\setminus\{k\} \text{ (because } \mathbf{X}_i\nparallel\mathbf{X}_k\text{)},
\end{aligned}\tag{35}$$

we have

$$\begin{aligned}
&\lambda_{d-2}(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)) = \lambda_{d-2}(\mathcal{S}^{d-2}) > 0,\\
&\lambda_{d-2}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\cap\ker(\mathbf{X}_i)\right) = 0 \text{ for all } i\in\{1,2,\cdots,n\}\setminus\{k\}.
\end{aligned}\tag{36}$$

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

$$\begin{aligned}
\lambda_{d-2}(\mathcal{T}_k) &= \lambda_{d-2}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\right) - \lambda_{d-2}\left(\bigcup_{i\in\{1,2,\cdots,n\}\setminus\{k\}}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\cap\ker(\mathbf{X}_i)\right)\right)\\
&\geq \lambda_{d-2}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\right) - \sum_{i\in\{1,2,\cdots,n\}\setminus\{k\}}\lambda_{d-2}\left(\mathcal{S}^{d-1}\cap\ker(\mathbf{X}_k)\cap\ker(\mathbf{X}_i)\right)\\
&= \lambda_{d-2}(\mathcal{S}^{d-2})\\
&> 0.
\end{aligned}$$

Therefore, $\mathcal{T}_k\neq\varnothing$. □

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 $\min\operatorname{eig}(\mathbf{H}\mathbf{H}^T)$ later in Appendix F.

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

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

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

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

then we must have

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

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

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

(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 $\mathcal{S}^{d-1}$, the vector $\overrightarrow{\mathrm{OC}}$ denotes $\mathbf{X}_i$, the vector $\overrightarrow{\mathrm{OD}}$ denotes one of other $\mathbf{X}_j$'s, the vector $\overrightarrow{\mathrm{OE}}$ denotes $\boldsymbol{v}_{*,i}$, which is perpendicular to $\mathbf{X}_i$ (i.e., $\mathbf{X}_i^T \boldsymbol{v}_{*,i} = 0$). The upper half of the cap E denotes $\mathcal{B}^{r_i, \mathbf{X}_i}_{\boldsymbol{v}_{*,i},+}$, the lower half of the cap E denotes $\mathcal{B}^{r_i, \mathbf{X}_i}_{\boldsymbol{v}_{*,i},-}$. The great circle $\mathrm{L_c}$ cuts the sphere into two semi-spheres. The semi-sphere in the direction of $\overrightarrow{\mathrm{OC}}$ corresponds to all vectors $\boldsymbol{v}$ on the sphere that have positive inner product with $\mathbf{X}_i$ (i.e., $\boldsymbol{v}_{*,i}^T \mathbf{X}_i > 0$), and the semi-sphere in the opposite direction of $\overrightarrow{\mathrm{OC}}$ 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 $\mathrm{L_d}$ is similar to the great circle $\mathrm{L_c}$, but is perpendicular to the direction $\overrightarrow{\mathrm{OD}}$ (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 $\frac{\mathbf{V}_0[k]}{\|\mathbf{V}_0[k]\|_2} \in \mathcal{B}^{r_i, \mathbf{X}_i}_{\boldsymbol{v}_{*,i},+}$ 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 $\frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2} \in \mathcal{B}^{r_i, \mathbf{X}_i}_{\boldsymbol{v}_{*,i},-}$ 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 $\min_{\boldsymbol{a} \in \mathcal{S}^{n-1}} \|\mathbf{H}^T \boldsymbol{a}\|_2^2$. To see this, let a_j denotes the j-th element of $\boldsymbol{a}$. By Eq. (39), we have $\sum_{j \in \{1,2,\cdots,n\} \setminus \{i\}} ((\mathbf{H}^T a_j)[k] - (\mathbf{H}^T a_j)[l]) = \mathbf{0}$. 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 $\min \mathsf{eig}(\mathbf{H}\mathbf{H}^T)/p$ in Appendix F.

Proof. By the definition of r_i, we have

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

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

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

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

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

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

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

Thus, by Eq. (1), we have

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

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 $p \to \infty$. Note that when the design matrix is not full row-rank, there exists a set of indices $\mathcal{I} \subseteq \{1, \cdots, n\}$ such that

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

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

Step 1:

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

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

Define

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

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

$$\begin{aligned}
\mathcal{J}_i &:= \left\{ \mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},+}^{r_i, \mathbf{X}_i} \neq \varnothing, \ \mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},-}^{r_i, \mathbf{X}_i} \neq \varnothing \right\}, \\
\mathcal{J}_{i,+} &= \left\{ \mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},+}^{r_i, \mathbf{X}_i} \neq \varnothing \right\}, \\
\mathcal{J}_{i,-} &= \left\{ \mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},-}^{r_i, \mathbf{X}_i} \neq \varnothing \right\}, \\
\mathcal{J} &:= \bigcap_{i=1}^{n} \mathcal{J}_i.
\end{aligned}$$

(Recall the geometric interpretation in Remark 6. The events $\mathcal{J}_{i,+}$ and $\mathcal{J}_{i,-}$ 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 $\mathcal{J}_i = \mathcal{J}_{i,+} \cap \mathcal{J}_{i,-}$ means that there exist $\mathbf{V}_0[j]/\|\mathbf{V}_0[j]\|_2$

in both halves of the cap E. Finally, the event $\mathcal{J}$ occurs when $\mathcal{J}_i$ occurs for all i, although the vector $\mathbf{V}_0[j]/\|\mathbf{V}_0[j]\|_2$ that falls into the two halves may differ across i. As we will show later, whenever the event $\mathcal{J}$ occurs, the matrix $\mathbf{H}$ will have the full row-rank, which is why we are interesting in the probability of the event $\mathcal{J}$.)

Those definitions implies that

$$\mathcal{J}_i^c = \mathcal{J}_{i,+}^c \cup \mathcal{J}_{i,-}^c \text{ for all } i = 1, 2, \cdots, n, \tag{47}$$

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

Thus, we have

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

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

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

Therefore, we have

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

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

$$\lim_{p\to\infty} \Pr_{\mathbf{V}_0}[\mathcal{J}_i^c] = 0. \tag{51}$$

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

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

Step 2:

To complete the proof, it remains to show that the event $\mathcal{J}$ contradicts Eq. (44). Towards this end, we assume the event $\mathcal{J}$ happens. By Eq. (44), we can pick one $i \in \mathcal{I}$. Further, by the definition of $\mathcal{J}$, there exists r_i such that $\mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},+}^{r_i,\mathbf{X}_i} \neq \varnothing$ and $\mathcal{A}_{\mathbf{V}_0} \cap \mathcal{B}_{\boldsymbol{v}_{*,i},-}^{r_i,\mathbf{X}_i} \neq \varnothing$. In other words, there must exist $k, l \in \{1, \cdots, p\}$ such that

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

By Lemma 23, we have

$$\mathbf{H}_j[k] = \mathbf{H}_j[l], \text{ for all } j \in \{1, 2, \cdots, n\} \setminus \{i\}, \tag{52}$$

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

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

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

This contradicts the assumption Eq. (44) that

$$\sum_{j\in\mathcal{I}} b_j\mathbf{H}_j[k] = \sum_{j\in\mathcal{I}} b_j\mathbf{H}_j[l] = \mathbf{0}.$$

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 $\min \mathsf{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$.

Lemma 24.

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

Proof. We have

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

Thus, we have

$$\begin{aligned}&|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}|\\ =&\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}\|_2 \text{ (}\ell_2\text{-norm of a number equals to its absolute value)}\\ \leq&\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2\cdot\|\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\boldsymbol{\epsilon}\|_2 \text{ (by Lemma 12)}\\ \leq&\frac{\sqrt{p}\|\boldsymbol{\epsilon}\|_2}{\sqrt{\min \mathsf{eig}(\mathbf{H}\mathbf{H}^T)}} \text{ (by Lemma 11 and Eq. (54))}.\end{aligned}$$

□

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

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

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

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 $\min \mathsf{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$, it is equivalent to estimate $\min_{\boldsymbol{a}\in\mathcal{S}^{n-1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$.

[8] We can see that the key part during the proof of Proposition 4 is to estimate $\min \mathsf{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$. Lemma 25 shows a lower bound of $\min \mathsf{eig}\left(\mathbf{H}\mathbf{H}^T\right)/p$ which is almost $\Omega(n^{1-2d})$ 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.

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

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

$$\mathbf{\Sigma}\in\mathbb{R}^{(dp)\times n}=\mathsf{diag}(\Sigma_1,\Sigma_2,\cdots,\Sigma_n).$$

By properties of singular values, we have

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

We also have

$$\begin{aligned}\mathbf{H}\mathbf{H}^T&=\mathbf{W}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{W}^T\\&=\mathbf{W}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{W}^T\text{ (because }\mathbf{U}^T\mathbf{U}=\mathbf{I})\\&=\mathbf{W}\mathsf{diag}(\Sigma_1^2,\Sigma_2^2,\cdots,\Sigma_n^2)\mathbf{W}^T.\end{aligned}$$

This equation is indeed the eigenvalue decomposition of $\mathbf{H}\mathbf{H}^T$, which implies that its eigenvalues are $\Sigma_1^2,\Sigma_2^2,\cdots,\Sigma_n^2$. Thus, we have

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

□

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

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. $\min_{\boldsymbol{a}\in\mathcal{S}^{n-1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2>0$) when $p\to\infty$. 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 $i\in\{1,2,\cdots,n\}$, we choose one

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

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

$$c_{r_0}^i:=\min\left\{|\mathcal{A}_{\mathbf{V}_0}\cap\mathcal{B}_{\boldsymbol{v}_{*,i},+}^{r_0,\mathbf{X}_i}|,\ |\mathcal{A}_{\mathbf{V}_0}\cap\mathcal{B}_{\boldsymbol{v}_{*,i},-}^{r_0,\mathbf{X}_i}|\right\}.\tag{56}$$

Then, we define

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

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

(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

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

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 E, and $\frac{\mathbf{V}_0[l]}{\|\mathbf{V}_0[l]\|_2}$ is in the lower half of the cap E. The quantities r_0, r_i, and $\hat{r}$ can all be used as the radius of the cap E. The ratio $D_{\mathbf{X}}$ is proportional to the area of the cap E with radius $\hat{r}$ (or equivalently, the probability that the normalized $\mathbf{V}_0[j]$ falls in the cap 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*

$$\Pr_{\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD_{\mathbf{X}},\ \textit{for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\} \geq 1-4ne^{-npD_{\mathbf{X}}/6}.$$

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

Define

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

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

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

$$\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2\geq pD(n,d,\delta),\ \textit{for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\}\geq 1-4ne^{-npD(n,d,\delta)/6}-\delta.$$

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.

$$\mathcal{J} := \left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2\geq pD_{\mathbf{X}},\ \text{for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\},$$

$$\mathcal{J}_i := \left\{\text{there exists } \boldsymbol{a}\in\mathcal{S}^{n-1} \text{ that } i\in\operatorname*{arg\,max}_{j\in\{1,2,\cdots,n\}}|a_j|,\ \text{and } \|\mathbf{H}^T\boldsymbol{a}\|_2^2\leq pD_{\mathbf{X}}\right\},$$

$$\mathcal{K}_i := \left\{c_{r_i}^i\leq 2npD_{\mathbf{X}}\right\},\ \text{for } i=1,2,\cdots,n.$$

Those definitions directly imply that

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

Step 1: prove $\mathcal{J}_i\subseteq\mathcal{K}_i$

To show $\mathcal{J}_i\subseteq\mathcal{K}_i$, we only need to prove that $\mathcal{J}_i$ implies $\mathcal{K}_i$. To that end, it suffices to show $\|\mathbf{H}^T\boldsymbol{a}\|_2^2\geq\frac{c_{r_i}^i}{2n}$ for the vector $\boldsymbol{a}$ defined in $\mathcal{J}_i$. Because $i\in\arg\max_{j=1}^n|a_j|$ and $\|\boldsymbol{a}\|_2=1$, we have

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

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

$$\frac{\mathbf{V}_0[k_j]}{\|\mathbf{V}_0[k_j]\|_2}\in\mathcal{B}_{\boldsymbol{v}_{*,i},+}^{r_i,\mathbf{X}_i},\quad\frac{\mathbf{V}_0[l_j]}{\|\mathbf{V}_0[l_j]\|_2}\in\mathcal{B}_{\boldsymbol{v}_{*,i},-}^{r_i,\mathbf{X}_i}.$$

Thus, we have

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

We then have

$$\begin{aligned}\|(\mathbf{H}^T\boldsymbol{a})[k_j]\|_2^2+\|(\mathbf{H}^T\boldsymbol{a})[l_j]\|_2^2\geq&\frac{1}{2}\|a_i\mathbf{X}_i\|_2^2 \text{ (by Lemma 13)}\\ \geq&\frac{1}{2n}\text{ (by Eq. (63)).}\end{aligned}$$

Further, we have

$$\|\mathbf{H}^T\boldsymbol{a}\|_2^2=\sum_{j=1}^{p}\|(\mathbf{H}^T\boldsymbol{a})[j]\|_2^2\geq\sum_{j=1}^{c_{r_i}^i}\|(\mathbf{H}^T\boldsymbol{a})[k_j]\|_2^2+\|(\mathbf{H}^T\boldsymbol{a})[l_j]\|_2^2=\frac{c_{r_i}^i}{2n}. \tag{64}$$

Clearly, if the event $\mathcal{J}_i$ occurs, then $\|\mathbf{H}\boldsymbol{a}\|_2^2\leq pD_{\mathbf{X}}$. Combining with Eq. (64), we then have $c_{r_i}^i\leq 2npD_{\mathbf{X}}$. In other words, the event $\mathcal{K}_i$ must occur. Hence, we have shown that $\mathcal{J}_i\subseteq\mathcal{K}_i$.

Step 2: estimate the probability of $\mathcal{K}_i$

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

$$\Pr_{\mathbf{V}_0}\left\{\frac{\mathbf{V}_0[j]}{\|\mathbf{V}_0[j]\|_2}\in\mathcal{B}_{\boldsymbol{v}_*,i,+}^{r_0,\mathbf{X}_i}\,\middle|\, i\right\}=\frac{\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{2\lambda_{d-1}(\mathcal{S}^{d-1})}.$$

Thus, $|\mathcal{A}_{\mathbf{V}_0}\cap\mathcal{B}_{\boldsymbol{v}_*,i,+}^{r_0,\mathbf{X}_i}|$ follows the distribution $\mathsf{Bino}\left(p,\ \frac{\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{2\lambda_{d-1}(\mathcal{S}^{d-1})}\right)$ given i and $\mathbf{X}$. By Lemma 14 (with $\delta=\frac{1}{2}$), we have

$$\Pr_{\mathbf{V}_0}\left\{|\mathcal{A}_{\mathbf{V}_0}\cap\mathcal{B}_{\boldsymbol{v}_*,i,+}^{r_0,\mathbf{X}_i}|<\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{4\lambda_{d-1}(\mathcal{S}^{d-1})}\,\middle|\, i\right\}\leq 2\exp\left(-\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{48\lambda_{d-1}(\mathcal{S}^{d-1})}\right). \tag{65}$$

Similarly, we have

$$\Pr_{\mathbf{V}_0}\left\{|\mathcal{A}_{\mathbf{V}_0}\cap\mathcal{B}_{\boldsymbol{v}_*,i,-}^{r_0,\mathbf{X}_i}|<\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{4\lambda_{d-1}(\mathcal{S}^{d-1})}\,\middle|\, i\right\}\leq 2\exp\left(-\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{48\lambda_{d-1}(\mathcal{S}^{d-1})}\right). \tag{66}$$

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

$$\Pr_{\mathbf{V}_0}\left\{c_{r_0}^i<\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{4\lambda_{d-1}(\mathcal{S}^{d-1})}\,\middle|\, i\right\}\leq 4\exp\left(-\frac{p\lambda_{d-1}(\mathcal{B}_{\boldsymbol{v}_*}^{r_0})}{48\lambda_{d-1}(\mathcal{S}^{d-1})}\right).$$

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

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

i.e.,

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

Step3: estimate the probability of $\mathcal{J}$

We have

$$\begin{aligned}\Pr_{\mathbf{V}_0}[\mathcal{J}^c]\leq&\sum_{i=1}^{n}\Pr_{\mathbf{V}_0}[\mathcal{J}_i] \text{ (by Eq. (62) and the union bound)}\\ \leq&\sum_{i=1}^{n}\Pr_{\mathbf{V}_0}[\mathcal{K}_i] \text{ (by } \mathcal{J}_i\subseteq\mathcal{K}_i \text{ proven in Step 1)}\\ \leq&4n\exp\left(-\frac{1}{6}npD_{\mathbf{X}}\right) \text{ (by Eq. (67)).}\end{aligned}$$

Thus, we have

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

The result of this lemma thus follows. □

F.2. Proof of Lemma 28

Based on Lemma 27, it remains to estimate $\hat{r}$, 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 $x \in \mathbb{R}$, 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*

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

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.,

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

Thus, we have

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

□

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 $1 - \frac{4}{e^2} \approx 0.46 \leq 0.6$, we have

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

□

Lemma 32.

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

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

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

Proof. When $d=2$, we have $B\left(\frac{d-1}{2},\,\frac{1}{2}\right)=\pi$. When $d=3$, we have $B\left(\frac{d-1}{2},\,\frac{1}{2}\right)=2$. When $d=4$, we have $B\left(\frac{d-1}{2},\,\frac{1}{2}\right)\approx 1.57$. 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

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

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

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

Then, applying Lemma 30, we have

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

Thus, by Lemma 31, we have

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

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

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

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

$$\begin{aligned}B\left(\frac{d-1}{2},\ \frac{1}{2}\right) \leq & \frac{2\sqrt{d-3}}{d-1} + \frac{2}{\sqrt{d-3}}\\ \leq & \frac{4}{\sqrt{d-3}}.\end{aligned}$$

Notice that $\frac{4}{\sqrt{5-3}} = 2\sqrt{2} < \pi$. The result of this lemma thus follows.

□

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

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

Proof. We have

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

□

Lemma 34. *For any $\delta \in \left(0,\ \frac{2}{\pi}\right]$, we must have $\frac{\delta}{n^2 C_d} \leq \frac{1}{\sqrt{2}}$.*

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

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

$$I_x\left(\frac{d-1}{2}, \frac{1}{2}\right) \geq \frac{C_d}{\sqrt{2}(d-1)} x^{\frac{d-1}{2}},$$

and

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

Proof. we have

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

Thus, we have

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

which implies

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

□

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

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

We also have

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

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

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

Thus, it remains to show that

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

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

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

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

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

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

To conclude, for all $d \in \{2, 3, \cdots\}$, 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 $\boldsymbol{v}_{*,i}$ defined in Eq. (55). For any $b \in \left(0,\ \frac{1}{\sqrt{2}}\right]$, we have, for $\boldsymbol{x}$ independent of $\boldsymbol{v}_{*,i}$ and with distribution μ,

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

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

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

Or, equivalently,

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

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

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

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

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

Thus, we have

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

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

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

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

$$\Pr_{\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\} \geq 1-4ne^{-npD(n,d,\delta)/6}.$$

In other words,

$$\Pr_{\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1} \;\middle|\; \text{any given } \mathbf{X} \text{ such that } \hat{r}\geq\frac{\delta}{n^2 C_d}\right\} \geq 1-4ne^{-npD(n,d,\delta)/6}.$$

We thus have

$$\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1} \;\middle|\; \hat{r}\geq\frac{\delta}{n^2 C_d}\right\} \geq 1-4ne^{-npD(n,d,\delta)/6}. \qquad (75)$$

Thus, we have

$$\begin{aligned}
&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\}\\
\geq & \Pr_{\mathbf{X},\mathbf{V}_0}\left\{\hat{r}\geq\frac{\delta}{n^2 C_d}, \text{ and } \|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1}\right\}\\
= & \Pr_{\mathbf{X},\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq pD(n,d,\delta), \text{ for all } \boldsymbol{a}\in\mathcal{S}^{n-1} \;\middle|\; \hat{r}\geq\frac{\delta}{n^2 C_d}\right\}\cdot\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\hat{r}\geq\frac{\delta}{n^2 C_d}\right\}\\
\geq & (1-4ne^{-npD(n,d,\delta)/6})(1-\delta) \text{ (by Eq. (73) and Eq. (75))}\\
\geq & 1-4ne^{-npD(n,d,\delta)/6}-\delta.
\end{aligned}$$

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 $\delta \leq 1$ and $d \leq n^4$, we must have*

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

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

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

Proof. We start from

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

Thus, we have

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

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

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

□

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

We have

$$\begin{aligned} D(n,d,\delta)\Big|_{\delta=\frac{1}{\sqrt[m]{n}}} \geq & \frac{1}{2^{1.5d+5.5}d^{0.5d}n^{2d-1}n^{\frac{d-1}{m}}} \text{ (by Eq. (76))} \\ = & \frac{1}{2^{1.5d+5.5}d^{0.5d}n^{\left(2+\frac{1}{m}\right)(d-1)}n} \\ = & \frac{1}{J_m(n,d)n} \text{ (by Eq. (9)).} \end{aligned}$$

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

$$1-4ne^{-npD(n,d,\delta)/6}-\delta\Big|_{\delta=\frac{1}{\sqrt[m]{n}}} \geq 1-\frac{2}{\sqrt[m]{n}}.$$

Then, we have

$$m\in\left[1,\ \frac{\ln n}{\ln\frac{\pi}{2}}\right] \implies \left(\frac{\pi}{2}\right)^m \leq n \implies n^{\frac{1}{m}} \geq \frac{\pi}{2} \implies \frac{1}{\sqrt[m]{n}} \leq \frac{2}{\pi} \implies \delta \leq \frac{2}{\pi}.$$

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 $\min_{\boldsymbol{a}\in\mathcal{S}^{n-1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$. 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 $\min_{\boldsymbol{a}\in\mathcal{S}^{n-1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$.

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 $\min_{\boldsymbol{a}\in\mathcal{S}^{n-1}}\|\mathbf{H}^T\boldsymbol{a}\|_2^2/p$ 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*

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

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., $\|\mathbf{H}_i-\mathbf{H}_k\|_2$ is not likely to be large. Thus, we can construct $\boldsymbol{a}$ such that $\mathbf{H}^T\boldsymbol{a}$ is proportional to $\mathbf{H}_i-\mathbf{H}_k$, 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 $n\geq\pi(d-1)$, there must exist two different $\mathbf{X}_i$'s such that the angle between them is at most*

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

Lemma 39 is intuitive because $\mathcal{S}^{d-1}$ has limited area. When there are many $\mathbf{X}_i$'s on $\mathcal{S}^{d-1}$, 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 $n \geq \pi(d-1)$, we have*

$$\begin{aligned}\Pr_{\mathbf{V}_0,\mathbf{X}}\Bigg\{\frac{\min \mathsf{eig}(\mathbf{H}\mathbf{H}^T)}{p} \leq \frac{3\pi^2}{8}\left((d-1)B(\frac{d-1}{2},\frac{1}{2})\right)^{\frac{2}{d-1}} n^{-\frac{2}{d-1}} \\ +\frac{3}{4}\left((d-1)B(\frac{d-1}{2},\frac{1}{2})\right)^{\frac{1}{d-1}} n^{-\frac{1}{d-1}}\Bigg\} \\ \geq 1-2\exp\left(-\frac{p}{24}\right)-2\exp\left(-\frac{p}{12}\pi\left((d-1)B(\frac{d-1}{2},\frac{1}{2})\right)^{\frac{1}{d-1}} n^{-\frac{1}{d-1}}\right).\end{aligned}$$

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 $n^{\frac{1}{d-1}}$, $\frac{\min \mathsf{eig}(\mathbf{H}\mathbf{H}^T)}{p} = O(n^{-\frac{1}{d-1}})$ with high probability.

G.1. Proof of Lemma 38

We first prove a useful lemma.

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

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

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

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

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

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

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

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

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

$$\max_{\varphi\in[0,\pi/2]}(\varphi - \frac{\pi}{2}\sin\varphi) = \max_{\varphi\in\{0,\pi/2\}}(\varphi - \frac{\pi}{2}\sin\varphi) = 0.$$

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

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

$$\begin{aligned}\|\mathbf{X}_i - \mathbf{X}_k\|_2 =& 2\sin\frac{\theta}{2} \\ \leq& 2\cdot\frac{\theta}{2} \text{ (by Lemma 41)} \\ =& \theta.\end{aligned} \tag{78}$$

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

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

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

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

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

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

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

$$\underset{\mathbf{V}_0}{\mathsf{Pr}}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}} \geq \frac{3p}{4}\right\} \leq 2\exp\left(-\frac{p}{24}\right), \tag{80}$$

$$\underset{\mathbf{V}_0}{\mathsf{Pr}}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}} \geq \frac{3p\theta}{2\pi}\right\} \leq 2\exp\left(-\frac{p\theta}{12\pi}\right). \tag{81}$$

Thus, we have

$$
\begin{aligned}
&\Pr_{\mathbf{V}_0}\left\{\|\mathbf{H}^T\boldsymbol{a}\|_2^2 \geq \frac{3p\theta^2}{8}+\frac{3p\theta}{4\pi}\right\}\\
\leq &\Pr_{\mathbf{V}_0}\left\{\frac{\theta^2}{2}\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}}+\frac{1}{2}\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}\geq \frac{3p\theta^2}{8}+\frac{3p\theta}{4\pi}\right\}\\
&\text{(by Eq. (79))}\\
\leq &\Pr_{\mathbf{V}_0}\left\{\left\{\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}}>\frac{3p}{4}\right\}\cup\left\{\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}\geq\frac{3p\theta}{2\pi}\right\}\right\}\\
\leq &\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{\mathbf{X}_i^T\mathbf{V}_0[j]>0\}}>\frac{3p}{4}\right\}+\Pr_{\mathbf{V}_0}\left\{\sum_{j=1}^{p}\mathbf{1}_{\{(\mathbf{X}_i^T\mathbf{V}_0[j])(\mathbf{X}_k^T\mathbf{V}_0[j])<0\}}\geq\frac{3p\theta}{2\pi}\right\}\\
&\text{(by the union bound)}\\
\leq &2\exp\left(-\frac{p}{24}\right)+2\exp\left(-\frac{p\theta}{12}\right)\ \text{(by Eq. (80) and Eq. (81))}.
\end{aligned}
$$

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*

$$\frac{2\sqrt{2}(d-1)}{nC_d}\in\left[\frac{d-1}{n\sqrt{d}},\ \frac{\pi(d-1)}{n}\right].$$

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

$$C_d\in\left[\frac{2\sqrt{2}}{\pi},\ 2\sqrt{2d}\right].$$

Thus, we have

$$\frac{2\sqrt{2}(d-1)}{nC_d}\in\left[\frac{d-1}{n\sqrt{d}},\ \frac{\pi(d-1)}{n}\right].$$

□

Now we are ready to proof Lemma 39.

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

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

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

$$\sin\varphi\geq\frac{2\varphi}{\pi}=\left(\frac{2\sqrt{2}(d-1)}{nC_d}\right)^{\frac{1}{d-1}}. \tag{83}$$

By Lemma 8, the area of each cap is

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

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

$$A \geq \frac{1}{2}\lambda_{d-1}(\mathcal{S}^{d-1})\frac{C_d}{\sqrt{2}(d-1)}(\sin^2\varphi)^{\frac{d-1}{2}} = \frac{1}{n}\lambda_{d-1}(\mathcal{S}^{d-1}).$$

In other words, we have

$$\frac{\lambda_{d-1}(\mathcal{S}^{d-1})}{A} \leq n.$$

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 $\Delta\mathbf{V}^* \in \mathbb{R}^{dp}$ that

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

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

Lemma 43. $\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\Delta\mathbf{V}^* = f^g_{\mathbf{V}_0}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \mathcal{S}^{d-1}$.

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

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

□

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* $f(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\Delta\mathbf{V}^*$ *for all* $\boldsymbol{x}$*, then we have*

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

Proof. Because $f(\boldsymbol{x}) = \boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\Delta\mathbf{V}^*$, we have $\boldsymbol{y} = \mathbf{H}\Delta\mathbf{V}^* + \boldsymbol{\epsilon}$. Thus, we have

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

Further, we have

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

Finally, using Eq. (4), we have

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

□

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

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

$$|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*|\leq\sqrt{p}\|\Delta\mathbf{V}^*-\mathbf{H}\boldsymbol{a}\|_2.$$

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

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

Therefore, we have

$$\begin{aligned}|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*|=&\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*\|_2\\ \leq&\|\boldsymbol{h}_{\mathbf{V}_0,\boldsymbol{x}}\|_2\cdot\|(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*\|_2 \text{ (by Lemma 12)}\\ \leq&\sqrt{p}\|\Delta\mathbf{V}^*-\mathbf{H}\boldsymbol{a}\|_2 \text{ (by Lemma 11).}\end{aligned}$$

□

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

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

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

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

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

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

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

$$\mathop{\mathsf{E}}_{\mathbf{X}_i} [\mathbf{K}_i[j]] = \Delta \mathbf{V}^*[j]. \tag{88}$$

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

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

Thus, we have

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

i.e.,

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

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

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

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 $X_i = \sqrt{p}\mathbf{K}_i$, $U = \|g\|_\infty$, and $k = n$), we have

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

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

$$\Pr_{\mathbf{X}} \left\{ \sqrt{p} \|\mathbf{H}^T \boldsymbol{a} - \Delta \mathbf{V}^*\|_2 \geq n^{-\frac{1}{2}\left(1 - \frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{n}}{8\|g\|_\infty^2} \right). \tag{91}$$

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

$$\Pr_{\mathbf{X}} \left\{ \left| \hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x}) \right| \geq n^{-\frac{1}{2}\left(1 - \frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{n}}{8\|g\|_\infty^2} \right).$$

□

I. Proof of Theorem 1

We first prove a useful lemma.

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

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

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

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

Thus, we have

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

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

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

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

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

□

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

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

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

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

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

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

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

Thus, we have

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

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

$$\begin{aligned}
\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\text{term } A \geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\} &= \int_{\mathbf{V}_0\in\mathbb{R}^{dp}} \Pr_{\mathbf{X}}\left\{\text{term } A \geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\ \middle|\ \mathbf{V}_0\right\} d\lambda(\mathbf{V}_0)\\
&\leq \int_{\mathbf{V}_0\in\mathbb{R}^{dp}} 2e^2\exp\left(-\frac{\sqrt[q]{n}}{8\|g\|_\infty^2}\right) d\lambda(\mathbf{V}_0) \text{ (by Proposition 5)}\\
&= 2e^2\exp\left(-\frac{\sqrt[q]{n}}{8\|g\|_\infty^2}\right).
\end{aligned} \tag{94}$$

Step 3: estimate term C

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

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

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

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

From Definition 2, we have

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

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

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

Further, for any $j \in \{1, 2, \cdots, p\}$, we have

$$\begin{aligned}
\|K_j^{\boldsymbol{x}}\|_2 =& |K_j^{\boldsymbol{x}}| \text{ (because } K_j^{\boldsymbol{x}} \text{ is a scalar)} \\
=& \left| \int_{\mathcal{S}^{d-1}} \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{\{\boldsymbol{z}^T \mathbf{V}_0[j]>0,\ \boldsymbol{x}^T\mathbf{V}_0[j]>0\}} g(\boldsymbol{z}) d\mu(\boldsymbol{z}) \right| \\
\leq& \int_{\mathcal{S}^{d-1}} \left| \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{\{\boldsymbol{z}^T \mathbf{V}_0[j]>0,\ \boldsymbol{x}^T\mathbf{V}_0[j]>0\}} g(\boldsymbol{z}) \right| d\mu(\boldsymbol{z}) \\
\leq& \int_{\mathcal{S}^{d-1}} \left| \boldsymbol{x}^T \boldsymbol{z} \mathbf{1}_{\{\boldsymbol{z}^T \mathbf{V}_0[j]>0,\ \boldsymbol{x}^T\mathbf{V}_0[j]>0\}} \right| \cdot |g(\boldsymbol{z})| d\mu(\boldsymbol{z}) \\
\leq& \int_{\mathcal{S}^{d-1}} |g(\boldsymbol{z})| d\mu(\boldsymbol{z}) \\
=& \|g\|_1.
\end{aligned} \tag{98}$$

Thus, by Lemma 16, we have

$$\mathop{\mathsf{Pr}}_{\mathbf{V}_0} \left\{ \left\| \left(\frac{1}{p}\sum_{j=1}^{p} K_j^{\boldsymbol{x}} \right) - \mathop{\mathsf{E}}_{\mathbf{V}_0} K_1 \right\|_2 \geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{p}}{8\|g\|_1^2} \right).$$

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

$$\mathop{\mathsf{Pr}}_{\mathbf{V}_0} \left\{ \left| f_{\mathbf{V}_0}^g(\boldsymbol{x}) - f_g(\boldsymbol{x}) \right| \geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{p}}{8\|g\|_1^2} \right).$$

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

$$\mathop{\mathsf{Pr}}_{\mathbf{V}_0} \left\{ \left| f_{\mathbf{V}_0}^g(\boldsymbol{x}) - f(\boldsymbol{x}) \right| \geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{p}}{8\|g\|_1^2} \right).$$

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

$$\mathop{\mathsf{Pr}}_{\mathbf{V}_0, \mathbf{X}} \left\{ \text{term } C \geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)} \right\} \leq 2e^2 \exp\left(-\frac{\sqrt[q]{p}}{8\|g\|_1^2} \right). \tag{99}$$

Step 4: estimate term B

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

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

Then, we have

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

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

$$\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\left\|\mathbf{F}^g_{\mathbf{V}_0}(\mathbf{X})-\mathbf{F}(\mathbf{X})\right\|_2\geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\leq 2e^2\exp\left(-\frac{\sqrt[q]{p}}{8n\|g\|_1^2}\right). \tag{100}$$

Thus, we have

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

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

We have

$$\begin{aligned}
&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{|\hat{f}^{\ell_2}(\boldsymbol{x})-f(\boldsymbol{x})|\geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}+\frac{1+\sqrt{J_m(n,d)n}}{\sqrt[4]{p}}\right\}\\
\leq&\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term }A+\text{term }B+\text{term }C\geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}+\frac{1+\sqrt{J_m(n,d)n}}{\sqrt[4]{p}}\right\}\text{ (by Eq. (93))}\\
\leq&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\left\{\text{term }A\geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\cup\left\{\text{ term }B\geq\sqrt{J_m(n,d)n}p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\right.\\
&\left.\cup\left\{\text{ term }C\geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\right\}\\
\leq&\Pr_{\mathbf{X},\mathbf{V}_0}\left\{\text{term }A\geq n^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}+\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term }B\geq\sqrt{J_m(n,d)n}p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\\
&+\Pr_{\mathbf{V}_0,\mathbf{X}}\left\{\text{term }C\geq p^{-\frac{1}{2}\left(1-\frac{1}{q}\right)}\right\}\text{(by the union bound)}\\
\leq&2e^2\left(\exp\left(-\frac{\sqrt[q]{n}}{8\|g\|_\infty^2}\right)+\exp\left(-\frac{\sqrt[q]{p}}{8\|g\|_1^2}\right)+\exp\left(-\frac{\sqrt[q]{p}}{8n\|g\|_1^2}\right)\right)+\frac{2}{\sqrt[m]{n}}\\
&\text{(by Eqs. (94)(99)(101)).}
\end{aligned}$$

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 $\overline{\mathcal{F}^{\ell_2}}$)

We first show what $\hat{f}^{\ell_2}_\infty$ looks like. Define $\mathbf{H}^\infty\in\mathbb{R}^{n\times n}$ where its (i,j)-th element is

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

Notice that

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

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,

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

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

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

Further, define the number

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

and

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

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

By Eq. (102), we have

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

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

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

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

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

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

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

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

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

By Eq. (6), we thus have

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

By the definition of the Dirac delta function, we have

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

□

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

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

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

$$\mathcal{J}_1 := \left\{ \sup_{\boldsymbol{x},\boldsymbol{z}\in\mathcal{S}^{d-1}} \left| \frac{|\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{z},\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\boldsymbol{z})}{2\pi} \right| < \zeta \right\},$$

$$\mathcal{J}_2 := \left\{ \max_{i\in\{1,2,\cdots,n\}} |g_{i,p} - g_{i,\infty}| < \zeta \right\}.$$

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

$$\mathsf{Pr}[\mathcal{J}_1] > 1 - \zeta.$$

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

$$\mathsf{Pr}[\mathcal{J}_2] > 1 - \zeta.$$

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

$$\mathsf{Pr}[\mathcal{J}_1 \cap \mathcal{J}_2] > 1 - 2\zeta. \tag{107}$$

When $\mathcal{J}_1 \cap \mathcal{J}_2$ happens, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{S}^{d-1}} |\hat{f}^{\ell_2}(\boldsymbol{x}) - \hat{f}^{\ell_2}_{\infty}(\boldsymbol{x})| \\
=&\sup_{\boldsymbol{x}\in\mathcal{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^T\mathbf{X}_i \left(\frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} g_{i,\infty} \right) \right| \\
&\text{(by Lemma 48 and Eq. (105))} \\
\leq& \sup_{\boldsymbol{x}\in\mathcal{S}^{d-1}, i\in\{1,2,\cdots,n\}} \left| \left(\frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} g_{i,p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} g_{i,\infty} \right) \right| \text{ (because } |\boldsymbol{x}^T\mathbf{X}_i| \leq 1) \\
=& \sup_{\boldsymbol{x}\in\mathcal{S}^{d-1}, i\in\{1,2,\cdots,n\}} \left| \left(\frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \right) g_{i,\infty} + (g_{i,p} - g_{i,\infty}) \frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \right| \\
\leq& \sup_{\boldsymbol{x}\in\mathcal{S}^{d-1}, i\in\{1,2,\cdots,n\}} \left| \left(\frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} - \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \right) g_{i,\infty} \right| + \left| (g_{i,p} - g_{i,\infty}) \frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \right| \\
\leq& \zeta \cdot \left(\max_i |g_{i,\infty}| + 1 \right) \text{ (because } \mathcal{J}_1 \cap \mathcal{J}_2 \text{ happens, } \frac{|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_i,\boldsymbol{x}}|}{p} \in [0,1] \text{, and } \frac{\pi - \arccos(\boldsymbol{x}^T\mathbf{X}_i)}{2\pi} \in [0, 0.5]).
\end{aligned}$$

Because $\max_i |g_{i,\infty}|$ is fixed when $\mathbf{X}$ is given, $\zeta \cdot (\max_i |g_{i,\infty}| + 1)$ 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 $f \notin \overline{\mathcal{F}^{\ell_2}}$ (or equivalently, $D(f, \mathcal{F}^{\ell_2}) > 0$), then the MSE of $\hat{f}_{\infty}^{\ell_2}$ (with respect to the ground-truth function f) is at least $D(f, \mathcal{F}^{\ell_2})$ (because $\hat{f}_{\infty}^{\ell_2} \in \mathcal{F}^{\ell_2}$). 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 $\hat{f}_{\infty}^{\ell_2}$, 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 $p > p_0$, $\Pr\{\sqrt{MSE} \geq D(f, \mathcal{F}^{\ell_2}) - \zeta\} > 1 - \zeta$.*

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

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

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

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

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

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

Notice that 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 $\mathcal{S}^{d-1}$)

K.1. Convolution on $\mathcal{S}^{d-1}$

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

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

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

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

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

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

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

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

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* $f(\boldsymbol{x}) \in \mathcal{F}^{\ell_2}$*, then* $f(\mathbf{S}\boldsymbol{x}) \in \mathcal{F}^{\ell_2}$*.*

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

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

Thus, we have

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

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 $\boldsymbol{x} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_d] \in \mathcal{S}^{d-1}$ 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)

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

where the normalization factor is

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

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,

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

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

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

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

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

The following lemma gives an explicit form of Gegenbauer polynomials.

Lemma 52.

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

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.,

$$\begin{aligned}C_i^{\lambda}(t)&=\sum_{k=0}^{\lfloor\frac{i}{2}\rfloor}(-1)^k\frac{\Gamma(i-k+\lambda)}{\Gamma(\lambda)k!(i-2k)!}(2t)^{i-2k},\\ C_{i+1}^{\lambda}(t)&=\sum_{k=0}^{\lfloor\frac{i+1}{2}\rfloor}(-1)^k\frac{\Gamma(i-k+\lambda+1)}{\Gamma(\lambda)k!(i-2k+1)!}(2t)^{i-2k+1}.\end{aligned}$$

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

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

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 $k=0,1,\cdots,\lfloor\frac{i+2}{2}\rfloor$. For the first step, we verify the coefficients of $(2t)^{i-2k+2}$ for $k=1,\cdots,\lfloor\frac{i+1}{2}\rfloor$. We have

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

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

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

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

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

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

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

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

$$\Xi_{\mathbf{0}}^l(\xi)=A_{\mathbf{0}}^l\sum_{k=0}^{\lfloor\frac{l}{2}\rfloor}(-1)^k\frac{\Gamma(l-k+\frac{d-2}{2})}{\Gamma(\frac{d-2}{2})k!(l-2k)!}(2\cos\theta_{d-1})^{l-2k}. \tag{115}$$

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

$$\begin{aligned}
\Xi_{\mathbf{0}}^0(\xi)&=A_{\mathbf{0}}^0,\\
\Xi_{\mathbf{0}}^1(\xi)&=A_{\mathbf{0}}^1(d-2)\cos\theta_{d-1},\\
\Xi_{\mathbf{0}}^2(\xi)&=A_{\mathbf{0}}^2\frac{d-2}{2}\left(d\cos^2\theta_{d-1}-1\right),\\
\Xi_{\mathbf{0}}^3(\xi)&=A_{\mathbf{0}}^3\frac{d-2}{2}\cdot d\cdot\left(\frac{d+2}{3}\cos^3\theta_{d-1}-\cos\theta_{d-1}\right).
\end{aligned}$$

K.4. Proof of Proposition 3

Recall that

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

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

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

The $\arccos$ function has a Taylor Series Expansion:

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

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

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

By comparing terms of even and odd power of $\cos\theta_{d-1}$ in Eq. (115) and Eq. (116), we immediately see that $h(\boldsymbol{x}) \not\perp \Xi_{\mathbf{0}}^{l}(\boldsymbol{x})$ when $l = 1$, and $h(\boldsymbol{x}) \perp \Xi_{\mathbf{0}}^{l}(\boldsymbol{x})$ when $l = 3, 5, 7, \cdots$. It remains to examine whether $h(\boldsymbol{x}) \perp \Xi_{\mathbf{0}}^{l}(\boldsymbol{x})$ or $h(\boldsymbol{x}) \not\perp \Xi_{\mathbf{0}}^{l}(\boldsymbol{x})$ for $l \in \{0, 1, 2, 4, 6, \cdots\}$. We first introduce the following lemma.

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

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

We must have

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

Proof. We have

$$Q(2k, 0) = \int_{\mathcal{S}^{d-1}} \cos^{2k}(\theta_{d-1})\Xi_{\mathbf{0}}^{0}(\xi)d\mu(\boldsymbol{x}) = A_{\mathbf{0}}^{0}\int_{\mathcal{S}^{d-1}} \cos^{2k}(\theta_{d-1})d\mu(\boldsymbol{x}) > 0.$$

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

$$Q(0, 2m) = \int_{\mathcal{S}^{d-1}} \Xi_{\mathbf{0}}^{2m}(\xi)d\mu(\boldsymbol{x}) = \frac{1}{A_{\mathbf{0}}^{0}}\int_{\mathcal{S}^{d-1}} \Xi_{\mathbf{0}}^{0}(\xi)\Xi_{\mathbf{0}}^{2m}(\xi)d\mu(\boldsymbol{x}) = 0$$
$$\text{(by the orthogonality of the basis).}$$

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

$$\cos(\theta_{d-1})\Xi_{\mathbf{0}}^{l+1}(\xi) = \frac{(l+2)A_{\mathbf{0}}^{l+1}}{(d+2l)A_{\mathbf{0}}^{l+2}}\Xi_{\mathbf{0}}^{l+2}(\xi) + \frac{(d-2+l)A_{\mathbf{0}}^{l+1}}{(d+2l)A_{\mathbf{0}}^{l}}\Xi_{\mathbf{0}}^{l}(\xi).$$

Thus, we have

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

where

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

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

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

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

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

To prove that Eq. (117) holds when $k = i + 1$ for all m, we consider two cases, Case 1: $m \leq i + 1$, 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 $m \leq i + 1$, we have $m - 1 \leq i$. Thus, by the induction hypothesis for $k = i$, we have $Q(2i, 2m - 2) > 0$ (by $m - 1 \leq i$), 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 $Q(2i, 2m + 2) \geq 0$ and $Q(2i, 2m) \geq 0$ (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 $m \leq i + 1$.

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

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

Thus, by Eq. (116), we know that $h(\boldsymbol{x}) \not\perp \Xi_{\mathbf{0}}^{l}(\boldsymbol{x})$ for all $l \in \{0, 2, 4, \cdots\}$.

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

When $d = 2$, $\mathcal{S}^{d-1}$ denotes a unit circle. Therefore, every $\boldsymbol{x}$ corresponds to an angle $\varphi \in [-\pi, \pi]$ such that $\boldsymbol{x} = [\cos\varphi \ \ \sin\varphi]^T$. In this situation, the hyper-spherical harmonics are the well-known Fourier series, i.e., $1, \cos(\theta), \sin(\theta), \cos(2\theta), \sin(2\theta), \cdots$. 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 $f_g \in \mathcal{F}^{\ell_2}$, we have

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

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

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

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

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

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 $h(\cdot)$. We have

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

It is easy to verify that

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

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 $n \neq \pm 1$. We have

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

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

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

By Eq. (123), we thus have

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

In other words, when $d = 2$, functions in $\mathcal{F}^{\ell_2}$ can only contain frequencies $0, \theta, 2\theta, 4\theta, 6\theta, \cdots$, and cannot contain other frequencies $3\theta, 5\theta, 7\theta, \cdots$.

K.6. Details of Remark 2

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

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

Thus, for all $\boldsymbol{x} = \left[\begin{smallmatrix} \tilde{\boldsymbol{x}} \\ 1/\sqrt{d} \end{smallmatrix}\right]$ and $\boldsymbol{a} = \left[\begin{smallmatrix} \tilde{\boldsymbol{a}} \\ a_0 \end{smallmatrix}\right]$, we have

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

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 ($\|g\|_\infty = \infty$)

We now discuss what happens to the conclusion of Theorem 1 if g contains a δ-function, in which case $\|g\|_\infty = \infty$. In Eq. (10) of Theorem 1, only Term 1 and Term 4 (come from Proposition 5) will be affected when $\|g\|_\infty = \infty$. That is because only Proposition 5 requires $\|g\|_\infty < \infty$ during the proof of Theorem 1. To accommodate the situation when g contains a δ-function ($\|g\|_\infty = \infty$), 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 $\|g\|_\infty = \infty$.

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

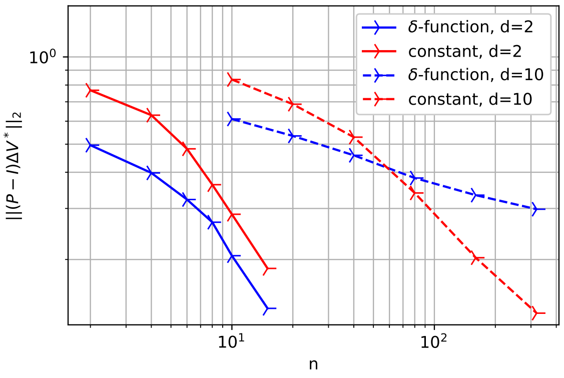

Figure 7. The curves of the model error $\|(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*\|_2$ for learning the pseudo ground-truth $f_{\mathbf{V}_0}^g$ with respect to n for different g and different d, where $p=20000$, and $\boldsymbol{\epsilon}=\mathbf{0}$. Every curve is the average of 10 random simulation runs.

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

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

when

$$n\geq\left((d-1)B(\frac{d-1}{2},\frac{1}{2})\right)^{\frac{q}{q-1}},\ i.e.,\ \left((d-1)B(\frac{d-1}{2},\frac{1}{2})\right)n^{-(1-\frac{1}{q})}\leq 1. \tag{125}$$

(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 $d\geq 3$, 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 $f_g\in\mathcal{F}^{\ell_2}$ is still learnable even when g is a δ-function (i.e., $\|g\|_\infty=\infty$), 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 $\|(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*\|_2$ for learning the pseudo ground-truth $f_{\mathbf{V}_0}^g$ with respect to n when $g=\delta_{\boldsymbol{z}_0}$ (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 $\|(\mathbf{P}-\mathbf{I})\Delta\mathbf{V}^*\|_2$ can represent the generalization performance for learning the pseudo ground-truth $f_{\mathbf{V}_0}^g$ 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 $q\in(1,\infty)$, if $b\in[n^{-(1-1/q)},1]$, then*

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

Proof. By Lemma 29, we have

$$\begin{aligned}
&e^{-b} \geq 1-b\\
\implies &e^{-1} \geq (1-b)^{\frac{1}{b}}\\
\implies &\exp\left(-n^{\frac{1}{q}}\right) \geq (1-b)^{n^{\frac{1}{q}}/b}\\
\implies &\exp\left(-n^{\frac{1}{q}}\right) \geq (1-b)^{n} \text{ because } b\in[n^{-(1-1/q)}, 1].
\end{aligned}$$

□

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

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

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 $\Delta\mathbf{V}^*$ defined in Eq. (84). By Eq. (1) and $g = \delta_{\boldsymbol{z}_0}$, we have

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

Define

$$\begin{aligned}
i^* &= \underset{i\in\{1,2,\cdots,n\}}{\arg\min} \|\mathbf{X}_i - \boldsymbol{z}_0\|_2,\\
\theta^* &= \arccos(\mathbf{X}_{i^*}^T \boldsymbol{z}_0).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\|\mathbf{X}_{i^*} - \boldsymbol{z}_0\|_2 =&\sqrt{2-2\cos\theta^*} \text{ (by the law of cosines)}\\
=&2\sin\frac{\theta^*}{2} \text{ (by the half angle identity)}\\
\leq&\theta^* \text{ (by Lemma 41).}
\end{aligned} \tag{127}$$

(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 $\Delta\mathbf{V}^*$. Define $\boldsymbol{a} \in \mathbb{R}^n$ whose i-th element is

$$\boldsymbol{a}_i = \begin{cases} 1/p, & \text{if } i = i^*\\ 0, & \text{if } i \in \{1,2,\cdots,n\}\setminus\{i^*\}\end{cases}.$$

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

$$\begin{aligned}
\|\mathbf{H}^T\boldsymbol{a} - \Delta\mathbf{V}^*\|_2^2 =&\sum_{j=1}^{p} \|(\mathbf{H}^T\boldsymbol{a})[j] - \Delta\mathbf{V}^*[j]\|_2^2\\
=&\frac{1}{p^2}\sum_{j=1}^{p}\left(\mathbf{1}_{\{\mathbf{X}_{i^*}^T\mathbf{V}_0[j]>0, \boldsymbol{z}_0^T\mathbf{V}_0[j]>0\}}\|\mathbf{X}_{i^*} - \boldsymbol{z}_0\|_2^2 + \mathbf{1}_{\{(\mathbf{X}_{i^*}^T\mathbf{V}_0[j])(\boldsymbol{z}_0^T\mathbf{V}_0[j])<0\}}\right)\\
\leq&\frac{1}{p^2}\left(p\|\mathbf{X}_{i^*} - \boldsymbol{z}_0\|_2^2 + |\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*}, \boldsymbol{z}_0}| + |\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*}, -\boldsymbol{z}_0}|\right) \text{ (by Eq. (6))}\\
\leq&\frac{1}{p^2}\left(p\cdot(\theta^*)^2 + |\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*}, \boldsymbol{z}_0}| + |\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*}, -\boldsymbol{z}_0}|\right) \text{ (by Eq. (127)).}
\end{aligned}$$

Thus, we have

$$\begin{aligned}\sqrt{p}\|\mathbf{H}\boldsymbol{a}-\Delta\mathbf{V}^*\|_2 \leq& \sqrt{(\theta^*)^2+\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|}{p}}\\ \leq& \sqrt{\pi\theta^*+\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|}{p}}\ \text{(because } \theta^*\leq\pi\text{)}.\end{aligned} \tag{128}$$

Remark 7. We give a geometric interpretation of Eq. (128) when $d=2$ by Fig. 4, where $\overrightarrow{\mathrm{OA}}$ denotes $\boldsymbol{z}_0$, $\overrightarrow{\mathrm{OB}}$ denotes $\mathbf{X}_{i^*}$. Then, $|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|$ corresponds to the number of $\mathbf{V}_0[j]$'s whose direction is in the arc $\overset{\frown}{\mathrm{CE}}$ or the arc $\overset{\frown}{\mathrm{FD}}$, and θ^* corresponds to the angle $\angle\mathrm{AOB}$. Intuitively, when n increases, $\mathbf{X}_{i^*}$ and $\boldsymbol{z}_0$ get closer, so θ^* becomes smaller. At the same time, both the arc $\overset{\frown}{\mathrm{CE}}$ and the arc $\overset{\frown}{\mathrm{FD}}$ 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

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

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

$$\begin{aligned}\mathcal{J}_1 &:= \left\{\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|}{p}\leq\frac{3\theta}{2\pi}\right\},\\ \mathcal{J}_2 &:= \{\theta^*\leq\theta\}.\end{aligned}$$

If both $\mathcal{J}_1$ and $\mathcal{J}_2$ happen, by Eq. (128), we must then have

$$\begin{aligned}\sqrt{p}\|\mathbf{H}\boldsymbol{a}-\Delta\mathbf{V}^*\|_2 \leq& \left(\sqrt{\frac{3}{2\pi}+\pi}\right)\cdot\sqrt{\theta}\\ =& \left(\sqrt{\frac{3}{4}+\frac{\pi^2}{2}}\right)\left(\frac{2\sqrt{2}(d-1)}{C_d}\right)^{\frac{1}{2(d-1)}} n^{-\frac{1}{2(d-1)}(1-\frac{1}{q})}.\end{aligned}$$

Thus, by Lemma 44 and Lemma 45, if $f=f^g_{\mathbf{V}_0}$ and both $\mathcal{J}_1$ and $\mathcal{J}_2$ happen, then for any $\boldsymbol{x}\in\mathcal{S}^{d-1}$, we must have

$$|\hat{f}^{\ell_2}(\boldsymbol{x})-f(\boldsymbol{x})|\leq\left(\sqrt{\frac{3}{4}+\frac{\pi^2}{2}}\right)\left(\frac{2\sqrt{2}(d-1)}{C_d}\right)^{\frac{1}{2(d-1)}} n^{-\frac{1}{2(d-1)}(1-\frac{1}{q})}. \tag{130}$$

It then only remains to estimate the probability of $\mathcal{J}_1\cap\mathcal{J}_2$.

Step 1: Estimate the probability of $\mathcal{J}_1$ conditional on $\mathcal{J}_2$.

When $\mathcal{J}_2$ happens, we have $\theta^*<\theta$. By Lemma 56, we can find $\boldsymbol{x}\in\mathcal{S}^{d-1}$ such that the angle between $\boldsymbol{x}$ and $\boldsymbol{z}_0$ is exactly θ and

$$\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|}{p}\leq\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x},-\boldsymbol{z}_0}|}{p}. \tag{131}$$

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 $\overrightarrow{\mathrm{OA}}$ as $\boldsymbol{z}_0$ and $\overrightarrow{\mathrm{OB}}$ as $\mathbf{X}_{i^*}$, then $|\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}|$ corresponds to the number of $\mathbf{V}_0[j]$'s whose direction is in the arc $\overset{\frown}{\mathrm{CE}}$ or the arc $\overset{\frown}{\mathrm{FD}}$. If we fix $\overrightarrow{\mathrm{OA}}$ (i.e., $\boldsymbol{z}_0$) and increase the angle $\angle\mathrm{AOB}$ (corresponding to θ^*), then both the arc $\overset{\frown}{\mathrm{CE}}$ and the arc $\overset{\frown}{\mathrm{FD}}$ will become longer. In other words, if we replace $\mathbf{X}_i$ by $\boldsymbol{x}$ such that the angle θ^* (between $\boldsymbol{z}_0$ and $\mathbf{X}_i$) increases to the angle θ (between $\boldsymbol{z}_0$ and $\boldsymbol{x}$), then $\mathcal{C}^{\mathbf{V}_0}_{-\mathbf{X}_{i^*},\boldsymbol{z}_0}\subseteq\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x},\boldsymbol{z}_0}$ and $\mathcal{C}^{\mathbf{V}_0}_{\mathbf{X}_{i^*},-\boldsymbol{z}_0}\subseteq\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x},-\boldsymbol{z}_0}$, 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

$$\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x},-\boldsymbol{z}_0}|}{p}=\frac{1}{p}\sum_{j=1}^{p}\underbrace{\mathbf{1}_{\{-\boldsymbol{x}^T\mathbf{V}_0[j]>0,\boldsymbol{z}_0^T\mathbf{V}_0[j]>0\text{ OR }\boldsymbol{x}^T\mathbf{V}_0[j]>0,-\boldsymbol{z}_0^T\mathbf{V}_0[j]<0\}}}_{\text{Term A}}. \tag{132}$$

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 $2\cdot\frac{\pi-(\pi-\theta)}{2\pi}=\frac{\theta}{\pi}$. By letting $\delta=1/2$, $a=p$, $b=\frac{\theta}{\pi}$ in Lemma 14, we have

$$\Pr_{\mathbf{V}_0}\left\{\left||\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x},-\boldsymbol{z}_0}|-\frac{p\theta}{\pi}\right|>\frac{p\theta}{2\pi}\right\}\leq 2\exp\left(-\frac{p\theta}{12\pi}\right).$$

By Eq. (131), we then have

$$\Pr_{\mathbf{V}_0}[\mathcal{J}_1^c\mid\mathcal{J}_2]\leq\Pr_{\mathbf{V}_0}\left\{\frac{|\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x},\boldsymbol{z}_0}|+|\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x},-\boldsymbol{z}_0}|}{p}>\frac{3\theta}{2\pi}\right\}\leq 2\exp\left(-\frac{p\theta}{12\pi}\right).$$

Step 2: Estimate the probability of $\mathcal{J}_2$.

By Lemma 8 and Assumption 1, for any $i\in\{1,2,\cdots,n\}$ and because $\theta\in[0,\pi/2]$, we have

$$\begin{aligned}\Pr_{\mathbf{X}}\left\{\arccos(\mathbf{X}_i^T\boldsymbol{z}_0)>\theta\right\}=&1-\frac{1}{2}I_{\sin^2\theta}\left(\frac{d-1}{2},\frac{1}{2}\right)\\ \leq&1-\frac{C_d}{2\sqrt{2}(d-1)}\sin^{d-1}\theta\text{ (by Lemma 35).}\end{aligned}$$

Note that since $\Pr_{\mathbf{X}}\left\{\arccos(\mathbf{X}_i^T\boldsymbol{z}_0)>\theta\right\}\geq 0$, we must have

$$\frac{C_d}{2\sqrt{2}(d-1)}\sin^{d-1}\theta\leq 1. \tag{133}$$

Further, because all $\mathbf{X}_i$'s are *i.i.d.* for $i\in\{1,2,\cdots,n\}$, we have

$$\Pr_{\mathbf{X}}\{\theta^*>\theta\}=\Pr_{\mathbf{X}}\left\{\min_{i\in\{1,2,\cdots,n\}}\arccos(\mathbf{X}_i^T\boldsymbol{z}_0)>\theta\right\}\leq\left(1-\frac{C_d}{2\sqrt{2}(d-1)}\sin^{d-1}\theta\right)^n. \tag{134}$$

By Eq. (129) and Lemma 41, we then have

$$\sin\theta\geq\left(\frac{2\sqrt{2}(d-1)}{C_d}\right)^{\frac{1}{d-1}}n^{-\frac{1}{d-1}(1-\frac{1}{q})},$$

i.e.,

$$\frac{C_d}{2\sqrt{2}(d-1)}\sin^{d-1}\theta\geq n^{-(1-1/q)}.$$

Thus, by Eq. (133), Eq. (134), and Lemma 55, we have

$$\Pr_{\mathbf{X}}[\mathcal{J}_2^c]=\Pr_{\mathbf{X}}\{\theta^*>\theta\}\leq\exp\left(-n^{\frac{1}{q}}\right).$$

Combining the results of Step 1 and Step 2, we thus have

$$\begin{aligned}\Pr_{\mathbf{X},\mathbf{V}_0}[\mathcal{J}_1\cap\mathcal{J}_2] &= \Pr_{\mathbf{X},\mathbf{V}_0}[\mathcal{J}_1\mid\mathcal{J}_2]\cdot\Pr_{\mathbf{X},\mathbf{V}_0}[\mathcal{J}_2]\\ &=\Pr_{\mathbf{V}_0}[\mathcal{J}_1\mid\mathcal{J}_2]\cdot\Pr_{\mathbf{X}}[\mathcal{J}_2] \text{ (because of } \mathbf{V}_0 \text{ and } \mathbf{X} \text{ are independent)}\\ &\geq\left(1-2\exp\left(-\frac{p\theta}{12\pi}\right)\right)\left(1-\exp\left(-n^{\frac{1}{q}}\right)\right)\\ &\geq 1-\exp\left(-n^{\frac{1}{q}}\right)-2\exp\left(-\frac{p\theta}{12\pi}\right)\\ &=1-\exp\left(-n^{\frac{1}{q}}\right)-2\exp\left(-\frac{p}{24}\left(\frac{2\sqrt{2}(d-1)}{C_d}\right)^{\frac{1}{d-1}}n^{-\frac{1}{d-1}(1-\frac{1}{q})}\right)\text{ (by Eq. (130)).}\end{aligned}$$

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 $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}=\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0}=\varnothing$ (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

$$\boldsymbol{z}_{0,\perp}:=\frac{\boldsymbol{x}_1-(\boldsymbol{x}_1^T\boldsymbol{z}_0)\boldsymbol{z}_0}{\|\boldsymbol{x}_1-(\boldsymbol{x}_1^T\boldsymbol{z}_0)\boldsymbol{z}_0\|_2}.$$

Thus, we have $\boldsymbol{z}_{0,\perp}^T\boldsymbol{z}_0=0$ and $\|\boldsymbol{z}_{0,\perp}\|_2=1$, i.e., $\boldsymbol{z}_0$ and $\boldsymbol{z}_{0,\perp}$ are orthonormal basis vectors on the 2D plane $\mathcal{L}$ spanned by $\boldsymbol{x}_1$ and $\boldsymbol{z}_0$. Thus, we can represent $\boldsymbol{x}_1$ as

$$\boldsymbol{x}_1=\cos\varphi\cdot\boldsymbol{z}_0+\sin\varphi\cdot\boldsymbol{z}_{0,\perp}\in\mathcal{L}.$$

For any $\theta\in[\varphi,\pi]$, we construct $\boldsymbol{x}_2$ as

$$\boldsymbol{x}_2:=\cos\theta\cdot\boldsymbol{z}_0+\sin\theta\cdot\boldsymbol{z}_{0,\perp}\in\mathcal{L}.$$

In order to show $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}\subseteq\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$, we only need to prove any $j\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}$ must in $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$. For any $\mathbf{V}_0[j]$, $j=1,2,\cdots,p$, 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 $\mathcal{L}$[10], i.e.,

$$\boldsymbol{v}_j=\cos\theta_j\cdot\boldsymbol{z}_0+\sin\theta_j\cdot\boldsymbol{z}_{0,\perp}\in\mathcal{L}.$$

By the proof of Lemma 17, we know that $j\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}$ if and only if $\theta_j\in(-\frac{\pi}{2},\frac{\pi}{2})\cap(\pi+\varphi-\frac{\pi}{2},\pi+\varphi+\frac{\pi}{2})\ (\mathrm{mod}\ 2\pi)$. Similarly, $j\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$ if and only if $\theta_j\in(-\frac{\pi}{2},\frac{\pi}{2})\cap(\pi+\theta-\frac{\pi}{2},\pi+\theta+\frac{\pi}{2})\ (\mathrm{mod}\ 2\pi)$. Because $\varphi\in[0,\pi]$ and $\theta\in[\varphi,\pi]$, we have

$$(-\frac{\pi}{2},\frac{\pi}{2})\cap(\pi+\varphi-\frac{\pi}{2},\pi+\varphi+\frac{\pi}{2})\subseteq(-\frac{\pi}{2},\frac{\pi}{2})\cap(\pi+\theta-\frac{\pi}{2},\pi+\theta+\frac{\pi}{2})\ (\mathrm{mod}\ 2\pi).$$

Thus, whenever $j\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}$, we must have $j\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$. Therefore, we conclude that $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}\in\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$. Using a similar method, we can also show that $\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0}\subseteq\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x}_2,-\boldsymbol{z}_0}$. 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 $\mathcal{L}$. The reason that we do not need to worry about those j's such that $\mathbf{V}_0[j]\perp\mathcal{L}$ is as follows. When $\mathbf{V}_0[j]\perp\mathcal{L}$, 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 $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_1,\boldsymbol{z}_0}$, $\mathcal{C}^{\mathbf{V}_0}_{-\boldsymbol{x}_2,\boldsymbol{z}_0}$, $\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x}_1,-\boldsymbol{z}_0}$, or $\mathcal{C}^{\mathbf{V}_0}_{\boldsymbol{x}_2,-\boldsymbol{z}_0}$ in Eq. (126).