Proceedings of Machine Learning Research vol 134:1–68, 2021 34th Annual Conference on Learning Theory

Learning with invariances in random features and kernel models

Song Mei SONGMEI@BERKELEY.EDU
Department of Statistics, University of California, Berkeley

Theodor Misiakiewicz MISIAKIE@STANFORD.EDU
Department of Statistics, Stanford University

Andrea Montanari MONTANARI@STANFORD.EDU
Department of Statistics and Department of Electrical Engineering, Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

A number of machine learning tasks entail a high degree of invariance: the data distribution does not change if we act on the data with a certain group of transformations. For instance, labels of images are invariant under translations of the images. Certain neural network architectures —for instance, convolutional networks—are believed to owe their success to the fact that they exploit such invariance properties. With the objective of quantifying the gain achieved by invariant architectures, we introduce two classes of models: invariant random features and invariant kernel methods. The latter includes, as a special case, the neural tangent kernel for convolutional networks with global average pooling. We consider uniform covariates distributions on the sphere and hypercube and a general invariant target function. We characterize the test error of invariant methods in a high-dimensional regime in which the sample size and number of hidden units scale as polynomials in the dimension, for a class of groups that we call 'degeneracy α', with $\alpha \le 1$. We show that exploiting invariance in the architecture saves a d^{α} factor (d stands for the dimension) in sample size and number of hidden units to achieve the same test error as for unstructured architectures. Finally, we show that output symmetrization of an unstructured kernel estimator does not give a significant statistical improvement; on the other hand, data augmentation with an unstructured kernel estimator is equivalent to an invariant kernel estimator and enjoys the same improvement in statistical efficiency.

Keywords: Invariant function estimation, Random features, Kernel methods, convolutional neural tangent kernel, high dimensional limit

1. Introduction

Consider the following image classification problem. We are given data $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ is an image, and $y_i \in \mathbb{R}$ is its label. We would like to learn a function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ to predict labels of new unseen images. Throughout this paper we will measure prediction error in terms of the square loss $R(\hat{f}) := \mathbb{E}\{(y_{\text{new}} - \hat{f}(\boldsymbol{x}_{\text{new}}))^2\}$.

We can think of $\boldsymbol{x} \in \mathbb{R}^d$ as a pixel representation of an image. For instance if this is a grayscale (one channel) two-dimensional image, $\boldsymbol{x}$ can represent the pixel values on a $d_1 \times d_2$ grid with $d = d_1 d_2$. For mathematical convenience, we here work with the cartoon example of one-dimensional 'images' (or 'signals') with d pixels arranged on a line. Most of our results cover two-dimensional images as well.

We assume a model whereby the labels are $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, with noise ε_i independent of $\boldsymbol{x}_i$ with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2$. In many applications, the target function f_* is invariant under translations of the image: if $\boldsymbol{x}'$ is obtained by translating image $\boldsymbol{x}$, then $f_*(\boldsymbol{x}') = f_*(\boldsymbol{x})$. We

will consider here periodic shifts (in the case of one-dimensional images): for $\boldsymbol{x} \in \mathbb{R}^d$, $g_\ell \cdot \boldsymbol{x} := (x_{\ell+1}, \dots, x_d, x_1, \dots, x_\ell)$ denotes its ℓ-shift. Invariance implies $f_*(\boldsymbol{x}) = f_*(g_\ell \cdot \boldsymbol{x})$ for all ℓ and $\boldsymbol{x}$.

Convolutional neural networks are the state-of-the-art architecture for image classification and related computer vision tasks, and they are believed to exploit the translation invariance in a crucial way (Krizhevsky et al., 2012). Consider the simple example of two-layer convolutional networks with global average pooling. The network computes a nonlinear convolution of N filters $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N$ with the image $\boldsymbol{x}$. The results are then combined linearly with coefficients $a_1, \dots, a_N$:

$$f_{\mathsf{CNN}}(\boldsymbol{x}) = \frac{1}{d} \sum_{i=1}^{N} a_i \sum_{\ell=1}^{d} \sigma(\langle \boldsymbol{w}_i, g_\ell \cdot \boldsymbol{x} \rangle) \,. \tag{1}$$

This simple convolutional network can be compared with a standard fully-connected two-layer network with the same number of parameters: $f_{\mathsf{NN}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$. It is clear that —when the target function f_* is translation invariant— the convolutional model $f_{\mathsf{CNN}}(\boldsymbol{x})$ is at least as powerful as $f_{\mathsf{NN}}(\boldsymbol{x})$ in terms of approximation, since it is invariant by construction (see Appendix A.1 for a simple formal argument).

The main objective of this paper is to quantify the advantage of architectures —such as convolutional ones— that enforce invariance. We are interested in characterizing the gain both in approximation error and in generalization error. We consider a general type of invariance, defined by a group $\mathcal{G}_d$ that is represented as a subgroup of $\mathcal{O}(d)$, the orthogonal group in d dimensions. This means that each element $g \in \mathcal{G}_d$ is identified with an orthogonal matrix (which we will also denote by g), and group composition corresponds to matrix multiplication. The group element $g \in \mathcal{G}_d$ acts on $\mathbb{R}^d$ via $\boldsymbol{x} \mapsto g \cdot \boldsymbol{x}$. We will consider two simple distributions for the the signals $\boldsymbol{x}$: $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ (the uniform distribution over the sphere in d dimensions with radius $\sqrt{d}$) and $\boldsymbol{x} \sim \mathrm{Unif}(\mathcal{Q}^d)$ (with $\mathcal{Q}^d = \{+1, -1\}^d$ the discrete hypercube in d dimensions). We will write $(\mathcal{A}_d, \tau_d) \in \{(S^{d-1}(\sqrt{d}), \mathrm{Unif}), (\mathcal{Q}^d, \mathrm{Unif})\}$ for either of these two probability spaces. In the case of $\mathcal{A}_d = \mathcal{Q}^d$, we will further require the action of $\mathcal{G}_d$ to preserve $\mathcal{Q}^d$.

In order to gain some insights on the behavior of actual neural networks, we consider two classes of linear 'overparametrized' models: invariant random features models and invariant kernel machines. We next describe these two approaches.

Invariant random feature models. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and a group $\mathcal{G}_d$ endowed with invariant (Haar) measure π_d, we define the invariant random features (RF) function class

$$\mathcal{F}^N_{\mathrm{RF,inv}}(\boldsymbol{W}, \mathcal{G}_d) = \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{w}_i, g \cdot \boldsymbol{x} \rangle) \, \pi_d(\mathrm{d}g) : a_i \in \mathbb{R}, i \in [N] \Big\} \,. \tag{2}$$

Here $\boldsymbol{W} := (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$ is the set of first layer weights which are fixed and not optimized over. We draw them randomly with $(\sqrt{d} \cdot \boldsymbol{w}_i)_{i \le N} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ or $\mathrm{Unif}(\mathcal{Q}^d)$ depending on whether the feature vectors are $\boldsymbol{x}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ or $\mathrm{Unif}(\mathcal{Q}^d)$. If we let $\mathcal{G}_d$ be the cyclic group $\mathrm{Cyc}_d := \{g_0, g_1, \dots, g_{d-1}\}$ (here g_ℓ is the shift by ℓ positions), we obtain a random features version of the convolutional network of Eq. (1). Other examples will be presented in Section 2.

Given data $\{(\boldsymbol{x}_i, y_i)\}_{i \le n}$, we consider to fit the second-layer coefficients $(a_i)_{i \le N}$ in Eq. (2) using the random features ridge regression (RFRR). Notice that the estimated function $\hat{f}$ is invariant by construction, $\hat{f}(\boldsymbol{x}) = \hat{f}(g \cdot \boldsymbol{x})$. We will denote the space of square integrable $\mathcal{G}_d$-invariant functions on $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{Q}^d\}$ by $L^2(\mathcal{A}_d, \mathcal{G}_d)$.

Invariant kernel machines. We then consider kernel ridge regression (KRR) in the reproducing kernel Hilbert space (RKHS) defined by a $\mathcal{G}_d$-invariant kernel. By this we mean a kernel $H \in L^2(\mathcal{A}_d \times \mathcal{A}_d)$ such that, for all $g, g' \in \mathcal{G}_d$, the following folds for every $\boldsymbol{x}_1, \boldsymbol{x}_2$:

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = H(g \cdot \boldsymbol{x}_1, g' \cdot \boldsymbol{x}_2)\,. \tag{3}$$

Note that, as a consequence of this property, any function that is not in $L^2(\mathcal{A}_d, \mathcal{G}_d)$ (i.e. any function that is not invariant) has infinite RKHS norm: indeed this provides an alternate characterization of invariant kernel methods. Among $\mathcal{G}_d$-invariant kernels, we focus on the subclass that is obtained by averaging an inner product kernel over the group $\mathcal{G}_d$

$$H_{\rm inv}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2\rangle/d)\, \pi_d(\mathrm{d}g). \tag{4}$$

Invariant kernel machines can be regarded as large-width ($N \to \infty$) limits of invariant random features methods. Vice versa, the latter can be regarded as randomized approximations of invariant kernel methods. Moreover, invariant kernel methods also capture the large-width limits of other models, for instance, neural tangent models associated to convolutional networks (c.f. Section A.3).

We focus on a type of groups $\mathcal{G}_d$ that we call *groups of degeneracy* α.

Definition 1 (Groups of degeneracy α) *Let $V_{d,k}$ be the subspace of degree-k polynomials that are orthogonal to polynomials of degree at most $(k-1)$ in $L^2(\mathcal{A}_d)$, and denote by $V_{d,k}(\mathcal{G}_d)$ the subspace of $V_{d,k}$ formed by polynomials that are $\mathcal{G}_d$-invariant. We say that $\mathcal{G}_d$ has* degeneracy α *if for any integer $k \geq \alpha$ we have $\dim(V_{d,k})/\dim(V_{d,k}(\mathcal{G}_d)) \asymp d^\alpha$ (i.e., there exists $0 < c_k \leq C_k < \infty$ such that $c_k \leq \dim(V_{d,k})/\dim(V_{d,k}(\mathcal{G}_d))/d^\alpha \leq C_k$ for any $d \geq 2$).*

This definition includes as special cases the cyclic group for one and two-dimensional signals (see Section 2), which have both degeneracy 1. Note that we can define an equivalence relation between degree-k polynomials: for $p_k, p_k' \in V_{d,k}$, we have $p_k \sim p_k'$ if and only if there exists $g \in \mathcal{G}_d$ such that $p_k(g \cdot \boldsymbol{x}) = p_k'(\boldsymbol{x})$. The dimension of the quotient space $\dim(V_{d,k}/V_{d,k}(\mathcal{G}_d))$ is then exactly equal to the ratio $\dim(V_{d,k})/\dim(V_{d,k}(\mathcal{G}_d))$. For a group $\mathcal{G}_d$ with degeneracy α, we can think about d^α as the 'effective dimension' of the group seen through its action on polynomials. The effective dimension of the group is not necessary equal to the size of the group (e.g., see Example 3 which is an infinite group with degeneracy 1). We will see below that this effective dimension is exactly equal to the factor that we save in sample size and number of hidden units by using invariant architectures.

We compare invariant methods to standard (non-invariant) random features models with inner product activation, defined as

$$\mathcal{F}^N_{\rm RF}(\boldsymbol{W}) = \left\{ f(\boldsymbol{x}) = \sum_{i=1}^N a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : a_i \in \mathbb{R}, i \in [N] \right\}, \tag{5}$$

and standard inner product kernels $H(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$. For groups with degeneracy $\alpha \leq 1$, we obtain a fairly complete characterization of the gain achieved by using invariant models, when the target function is an arbitrary invariant function $f_* \in L^2(\mathcal{A}_d; \mathcal{G}_d)$.

Invariance gain: underparametrized case. Consider the invariant RF class (2) in the underparametrized regime $N \ll n$. We prove that the test error is dominated by the approximation error. Namely,

if $d^{\ell-\alpha} \ll N \ll d^{\ell+1-\alpha}$, then the test error (c.f. Eq. (7)) gives $R(f_*;\lambda) \approx \|\overline{\mathsf{P}}_{>\ell} f_*\|_{L^2}^2$, where $\overline{\mathsf{P}}_{>\ell}$ is the projection orthogonal to the subspace of degree ℓ polynomials. In order to achieve the same risk, standard (non-invariant) RF models would require $d^\ell \ll N \ll d^{\ell+1}$: invariance saves a d^α factor in the network width to achieve the same risk.

Invariance gain: overparametrized case. Consider next the overparametrized regime $n \ll N$. In this case the test error is dominated by the statistical error. Namely, if $d^{\ell-\alpha} \ll n \ll d^{\ell+1-\alpha}$, then the test error gives $R(f_*;\lambda) \approx \|\overline{\mathsf{P}}_{>\ell} f_*\|_{L^2}^2$. In order to achieve the same risk, standard (non-invariant) RF models would require $d^\ell \ll n \ll d^{\ell+1}$: invariance saves a d^α factor in the sample size to achieve the same risk.

These results are precisely presented in Theorem 2 and summarized in Table 1. We establish the same gain for invariant kernel methods in Theorem 5. While we focused in this paper on groups with degeneracy $\alpha \le 1$ (which include our primary motivating examples, cyclic group in one or two dimensions), we expect similar results to hold for groups with $\alpha > 1$ (indeed our current proof techniques can handle the case $\alpha > 1$ at the price of adding the condition $N, n \ge d^{O(\alpha)}$ in our theorems). We defer this to future work.

Output symmetrization and data augmentation. Output symmetrization and data augmentation are two alternative approaches to incorporate invariances in machine learning models. We show that the performance of output symmetrization of standard KRR does not improve over standard KRR, and hence is sub-optimal compared to invariant KRR. On the other hand, it was shown that (c.f. Li et al. (2019)) data augmentation is mathematically equivalent to invariant KRR for discrete groups. As a consequence, our theoretical results characterize the statistical gain by performing data augmentation.

It is important to mention that our treatment omits an important characteristic of convolutional architectures: the fact that the filters $\boldsymbol{w}_i$ of Eq. (1) have a short window size $q \ll d$. Namely, they have only q non-zero entries, for instance the first q entries. Using short-window filters has some interesting consequences, which can be investigated using the same approach developed here. We will report on these in a forthcoming article, and instead focus here on the impact of invariance.

Our analysis is enabled by a simple yet important observation, which might generalize to other settings. The subspaces $V_{d,k}$ of degree-k polynomials (see Definition 1) are eigenspaces for inner product kernels. At the same time, they are preserved under the symmetry group $\mathcal{G}_d$. Namely, define $f^{(g)}(\boldsymbol{x}) = f(g \cdot \boldsymbol{x})$, we have $f^{(g)} \in V_{d,k}$ for any $f \in V_{d,k}$, $g \in \mathcal{G}_d$. This observation is crucial in determining the eigendecomposition of the relevant kernels.

Let us finally emphasize, that the factor-d gain in sample size for degeneracy-one groups is not correctly predicted by a naive 'data augmentation heuristics'. The latter would suggest a gain of the order of $|\mathcal{G}_d|$ or of the size of orbits of $\mathcal{G}_d$. As shown by the example of band limited functions (see below) $|\mathcal{G}_d|$ can be ∞ but the degeneracy can still be one (and hence the gain is d).

1.1. Related literature

Invariant function estimation

A number of mathematical works emphasized the role of invariance in neural network architectures. Among others, Mallat (2012); Bruna and Mallat (2013); Mallat (2016) propose architectures

To fit a degree ℓ polynomial	Inner product random features	Invariant random features
Underparameterized regime ($N \ll n$)	$N \gg d^{\ell}$	$N \gg d^{\ell-\alpha}$
Overparameterized regime ($n \ll N$)	$n \gg d^{\ell}$	$n \gg d^{\ell-\alpha}$

Table 1: Sample size n and number of features N required to fit a $\mathcal{G}_d$-invariant polynomial of degree ℓ using ridge regression with the standard random features model (Eq. (5)) and the invariant random features model (Eq. (2)), for group $\mathcal{G}_d$ of degeneracy $\alpha \leq 1$.

('deep scattering networks') that explicitly achieve invariance to a rich group of transformations. However, these papers do not characterize the statistical error of these approaches.

The recent paper Li et al. (2020) constructs a simple data distribution on which a gap is proven between the sample complexity for convolutional architectures, and the one for standard (fully connected) architectures. This result differs from ours in several aspects. Most importantly, we study the risk for estimating general invariant functions using invariant kernels and random features, while Li et al. (2020) obtain results for a specific distribution using CNNs. Also, the weight sharing structure in Li et al. (2020) is different from the one in Eq. (1).

Another work Chen et al. (2020) studied the statistical benefits of data augmentation in the parametric setting via a group theory framework. Our result is different in the sense that we consider the non-parametric setting to estimate an invariant function using kernel methods.

To the best of the our knowledge, our paper is the first that characterizes the precise statistical benefit of using invariant random features and kernel models.

Convolutional neural networks and convolutional kernels

A recent line of work (Jacot et al., 2018; Li and Liang, 2018; Du et al., 2019b,a; Allen-Zhu et al., 2019b,a; Arora et al., 2019a; Zou et al., 2020; Oymak and Soltanolkotabi, 2020) studied the training dynamics of overparametrized neural networks under certain random initialization, and showed that it converges to a kernel estimator, which corresponds to the "neural tangent kernel". The convolutional neural tangent kernel, which corresponds to the tangent kernel of convolutional neural networks, was studied in Arora et al. (2019b); Li et al. (2019); Bietti and Mairal (2019). The connection between convolutional kernel ridge regression and data augmentation was pointed out in Li et al. (2019).

The network in Eq. (1) corresponds to a two-layer convolutional neural network with global average pooling, which is a special case of the convolutional network that was defined as in Arora et al. (2019b).

Random features and kernel methods

A number of authors have studied the generalization error of kernel machines (Caponnetto and De Vito, 2007; Jacot et al., 2020; Liang et al., 2020b,a) (Wainwright, 2019, Theorem 13.17) and random features models (Rahimi and Recht, 2009; Rudi and Rosasco, 2017; Ma et al., 2020; Bach, 2017). However, these results are not fine-grained enough to characterize the separation between invariant kernels (or random feature models) and standard inner product kernels, for several reasons. First, some of these results concern restricted target functions with bounded RKHS norm. Second, we establish a gap that holds pointwise, i.e. for any given target function f_*, while most of earlier

work only obtain minimax lower bounds. Finally, we need the upper and lower bounds match up to a $1 + o_d(1)$ factor, while earlier results only match up to unspecified constants.

The recent paper Jacot et al. (2020) provides sharp predictions for kernel machines, but it assumes that a certain random kernel matrix behaves like a random matrix with Gaussian components: proving an equivalence of this type is the central mathematical challenge we face here.

Our analysis builds on the general results of Ghorbani et al. (2021); Mei et al. (2021). In particular, Mei et al. (2021) provides general conditions under which the risk of random features and kernel methods can be characterized precisely. Checking these conditions for invariant methods requires to prove certain concentration properties for the entries of the relevant kernels. We achieve this goal for the cyclic group with general activations, and for degeneracy-α groups (for $\alpha \le 1$) with polynomial activations. Generalizing these results to other groups, data distributions, and activations is a promising direction.

2. Examples

In this section, we provide three examples of our general setting. We show in Appendix D that all these groups have degeneracy 1 and therefore satisfy the assumptions of our general theorems.

Example 1 (One-dimensional images) *The cyclic group has elements* $\mathrm{Cyc}_d = \{g_0, g_1, \dots, g_{d-1}\}$ *where* g_i *is a shift by* i *pixels. For any* $\boldsymbol{x} = (x_1, \dots, x_d)^{\mathsf{T}} \in \mathcal{A}_d$*, the action of group element* g_i *on* $\boldsymbol{x}$ *is defined by* $g_i \cdot \boldsymbol{x} = (x_{i+1}, x_{i+2}, \dots x_d, x_1, x_2, \dots, x_i)^{\mathsf{T}} \in \mathcal{A}_d$*. (In particular,* g_i *is identified with an orthogonal transfromation in* $\mathbb{R}^d$*.) The measure* π_d *is the uniform probability measure on* Cyc_d*, i.e.,*

$$\int_{\mathrm{Cyc}_d} f(g)\pi_d(\mathrm{d}g) = \frac{1}{d}\sum_{i=0}^{d-1} f(g_i).$$

We will refer to the invariant functions $L^2(\mathcal{A}_d, \mathrm{Cyc}_d)$ *as the 'cyclic functions'.*

Example 2 (Two-dimensional images) *Let* $d = d_1 \times d_2$*. We identify* $\mathcal{X}_{d_1\times d_2} = \{\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2} : \|\boldsymbol{X}\|_F^2 = d\}$ *with* $\mathbb{S}^{d-1}(\sqrt{d})$ *(simply by 'vectorizing' the matrix). The two-direction cyclic group has elements* $\mathrm{Cyc2D}_{d_1,d_2} = \{g_{ij} : 0 \le i < d_1, 0 \le j < d_2, \}$*. For any* $\boldsymbol{X} = (X_{ij})_{i\in[d_1],j\in[d_2]} \in \mathcal{X}_{d_1\times d_2}$*, the action of group element* $g_{ij} \in \mathrm{Cyc2D}_{d_1,d_2}$ *on* $\boldsymbol{X}$ *is defined by*

$$g_{ij} \cdot \boldsymbol{X} = \begin{bmatrix} X_{i+1,j+1} & \dots & X_{i+1,d_2} & X_{i+1,1} & \dots & X_{i+1,j} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{d_1,j+1} & \dots & X_{d_1,d_2} & X_{d_1,1} & \dots & X_{d_1,j} \\ X_{1,j+1} & \dots & X_{1,d_2} & X_{1,1} & \dots & X_{1,j} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{i,j+1} & \dots & X_{i,d_2} & X_{i,1} & \dots & X_{i,j} \end{bmatrix}.$$

Again, this is an orthogonal transformation in $\mathcal{X}_{d_1\times d_2} \cong \mathbb{S}^{d-1}(\sqrt{d})$*, and* $\mathrm{Cyc2D}_{d_1,d_2}$ *is isomorphic to a subgroup of* $\mathcal{O}(d)$*. The measure* π_d *is the uniform probability measure on* $\mathrm{Cyc2D}_{d_1,d_2}$*. We will refer to the invariant functions* $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc2D}_d)$ *as the 'two-direction cyclic functions'.*

Example 3 (The translation invariant function class on band-limited signals) *Suppose we have one-dimensional signals with very high resolution, but the signals are band-limited: their Fourier*

transforms have only d non-zero coefficients. We assume that the labels of the band-limited signals are invariant under translations. The following model captures this setting.

Let $\{\varphi_j\}_{j\in[d]} \subseteq \mathcal{F}([0,1])$ be the real Fourier basis functions in $L^2([0,1],\mathrm{Unif})$. That is, we define $\varphi_1(t) = 1$, and for $p = 1, 2, \ldots, \lfloor d/2 \rfloor$ (we assume d is odd), $\varphi_{2p}(t) = \sqrt{2}\cos(2\pi pt)$, $\varphi_{2p+1}(t) = \sqrt{2}\sin(2\pi pt)$. We define the band-limited covariate subspace $\mathbb{W}_d \subseteq L^2([0,1],\mathrm{Unif})$ to be ($\mathbb{W}$ stands for waves)

$$\mathbb{W}_d = \Big\{ x \in L^2([0,1]) : x(t) = \sum_{j=1}^{d} \hat{x}_j \varphi_j(t), \;\; \hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_d) \in \mathbb{S}^{d-1}(\sqrt{d}) \Big\}.$$

Then the space $\mathbb{W}_d$ can be identified with the space $\mathbb{S}^{d-1}(\sqrt{d})$.

Let $\mathrm{Sft}_d = \{g_u, u \in [0,1]\} \simeq \mathrm{SO}(2)$ be the translation group that can act on $\mathbb{W}_d$. For any $x \in \mathbb{W}_d$, the action of group element $g_u \in \mathrm{Sft}_d$ on x is defined by

$$[g_u \cdot x](t) = x(t-u).$$

Equivalently, the action of group element $g_u \in \mathrm{Sft}_d$ on $\hat{\boldsymbol{x}} \in \mathbb{S}^{d-1}(\sqrt{d})$ is defined by

$$g_u \cdot \hat{\boldsymbol{x}} = (\hat{x}_1, \cos(2\pi u)\hat{x}_2 + \sin(2\pi u)\hat{x}_3, -\sin(2\pi u)\hat{x}_2 + \cos(2\pi u)\hat{x}_3, \ldots).$$

That means, Sft_d can be interpreted as a subgroup of $\mathcal{O}(d)$. The measure π_d is the uniform distribution on Sft_d, i.e.,

$$\int_{\mathrm{Sft}_d} f(g)\pi_d(\mathrm{d}g) = \int_{[0,1]} f(g_s)\mathrm{d}s.$$

The function class $L^2(\mathbb{W}_d, \mathrm{Sft}_d)$, or equivalently $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{SO}(2))$, can be regarded as the translation invariant function class on band-limited signals.

3. Invariant random feature models

Let $\mathcal{G}_d$ be a group of degeneracy α with $\alpha \le 1$ as defined in Definition 1 and f_d be a function that is invariant under the action of $\mathcal{G}_d$, i.e., $f_d \in L^2(\mathcal{A}_d, \mathcal{G}_d)$. We consider fitting the data with the invariant random features model defined in Eq. (2) using ridge regression, which we call invariant RFRR. Namely, we learn a function $\hat{f}^{\mathrm{inv}}_{N,\lambda}(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda)) = \sum_{1\le j\le N} \hat{a}_j \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{w}_j, g\cdot \boldsymbol{x}\rangle)\pi_d(\mathrm{d}g)$ with

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{f}^{\mathrm{inv}}_{N,\lambda}(\boldsymbol{x}_i; \boldsymbol{a})\right)^2 + \frac{N\lambda}{d^\alpha}\|\boldsymbol{a}\|_2^2 \right\}, \tag{6}$$

where the regularization parameter λ can depend on the dimension d. (The factor d^α in the ridge penalty is introduced to compensate for the effect of averaging the random features over $\mathcal{G}_d$.) We further denote the test error of invariant RFRR by

$$R_{\mathrm{RF,inv}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\varepsilon}, \lambda) := \mathbb{E}_{\boldsymbol{x}}\left[\left(f_d(\boldsymbol{x}) - \hat{f}^{\mathrm{inv}}_{N,\lambda}(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda))\right)^2\right]. \tag{7}$$

We will make the following assumption on σ.

Assumption 1 (Conditions on σ, n, N, and $(\mathcal{A}_d, \mathcal{G}_d)$ at level $(\mathsf{s}, \mathsf{S}) \in \mathbb{N}^2$) *For $\sigma : \mathbb{R} \to \mathbb{R}$, we assume the following conditions hold.*

(a) For $(\mathcal{A}_d, \mathcal{G}_d) = (\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$, we assume σ to be $(\min(\mathsf{s}, \mathsf{S}) + 1) \vee 3$ differentiable and there exists constants $c_0 > 0$ and $c_1 < 1$ such that $|\sigma^{(k)}(u)| \le c_0 e^{c_1 u^2/2}$ for any $2 \le k \le (\min(\mathsf{s}, \mathsf{S}) + 1) \vee 3$. Moreover, there exists an integer $p > 1/\delta$ such that $n \le N^{1-\delta}$ or $N \le n^{1-\delta}$ and $|\sigma(x)|, |\sigma'(x)| \le c_0 \exp(c_1 x^2/(8p))$.

For general $(\mathcal{A}_d, \mathcal{G}_d)$, we assume that σ is a (finite degree) polynomial function.

(b) The Hermite coefficients $\mu_k(\sigma) \equiv \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)\mathrm{He}_k(G)]$ verify $\mu_k(\sigma) \neq 0$ for any $0 \le k \le \min(\mathsf{s}, \mathsf{S})$ (see Appendix H for definitions).

(c) We assume that σ is not a polynomial with degree less or equal to $\max(\mathsf{s}, \mathsf{S})$.

For $k \in \mathbb{N}$, we denote by $\overline{\mathsf{P}}_{\le k} : L^2(\mathcal{A}_d) \to L^2(\mathcal{A}_d)$ the orthogonal projection operator onto the subspace of polynomials of degree at most k, and $\overline{\mathsf{P}}_{>k} = \mathbf{I} - \overline{\mathsf{P}}_{\le k}$ (see Appendix H for details). We denote $f(d) = o_{d,\mathbb{P}}(g(d))$ if $f(d)/g(d)$ converges to 0 in probability as $d \to \infty$.

Theorem 2 (Test error of invariant RFRR) *Let $\mathcal{G}_d$ be a group of degeneracy $\alpha \le 1$ and let $\{f_d \in L^2(\mathcal{A}_d, \mathcal{G}_d)\}_{d\ge1}$ be a sequence of $\mathcal{G}_d$-invariant functions. Assume $d^{\mathsf{s}-\alpha+\delta} \le n \le d^{\mathsf{s}+1-\alpha-\delta}$ and $d^{\mathsf{S}-\alpha+\delta} \le N \le d^{\mathsf{S}+1-\alpha-\delta}$ for fixed integers s, S and some $\delta > 0$. Let σ be an activation function that satisfies Assumption 1 at level (s, S). Then the following hold for the test error of invariant RFRR (see Eq. (7)):*

(a) (Overparametrized regime) Assume $N \ge nd^\delta$ for some $\delta > 0$. Then for any regularization parameter $\lambda = O_d(1)$ (including $\lambda = 0$) and $\eta > 0$, we have

$$R_{\mathrm{RF,inv}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\varepsilon}, \lambda) = \|\overline{\mathsf{P}}_{>\mathsf{s}} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{8}$$

(b) (Underparametrized regime) Assume $n \ge Nd^\delta$ for some $\delta > 0$. Then for any regularization parameter $\lambda = O_d(n/N)$ (including $\lambda = 0$) and any $\eta > 0$, we have,

$$R_{\mathrm{RF,inv}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\varepsilon}, \lambda) = \|\overline{\mathsf{P}}_{>\mathsf{S}} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{9}$$

In particular, this theorem applies to the one-dimensional and two-dimensional cyclic groups, and band-limited functions listed in Section 2. We refer readers to Appendix A.2 for an informal intuition and Appendix B.2 for the proof of this result.

We can compare these bounds with ridge regression on the standard random features model of Eq. (5). Theorem 2 in Mei et al. (2021) (with Assumption 1) shows that the same test error holds as in Theorem 2 but with $d^{\mathsf{s}+\delta} \le n \le d^{\mathsf{s}+1-\delta}$ and $d^{\mathsf{S}+\delta} \le N \le d^{\mathsf{S}+1-\delta}$. We thus gain a factor d^α in the sample and feature complexity by using invariant features compared to non invariant ones.

Remark 3 *Assumption 1 requires the activation function to be polynomial, except for the cyclic group, for which only differentiability conditions are assumed. These conditions are sufficient for the general assumptions in Mei et al. (2021) to hold: for the sake of length, we only verify them for non-polynomial activation functions in the case of the cyclic group in one dimension. However we believe that the differentiability condition (and indeed weaker conditions) should be sufficient*

for general groups. For example, the current proofs can be modified to apply to more general subgroups of the permutation group on d elements (e.g., cyclic group in higher dimension). We defer these improvements to future work.

For the cyclic group, the current assumptions already include the interesting examples of the sigmoid $\sigma(x) = 1/(1+\exp(x-c))$ and smoothed ReLU $\sigma(x) = \mathbb{E}_{G\sim\mathsf{N}(0,\varepsilon^2)}[(x-c+G)_+]$.

Note that Assumption 1.b) is necessary for the RKHS associated to the feature map σ to include all polynomials of degree less or equal to $\min(\mathsf{s},\mathsf{S})$.

Remark 4 *Consider two-dimensional images with $d = D \times D$ (Example 2) and functions f_d that are invariant with respect to the group of cyclic translations along the horizontal direction only. It can be shown that this group has degeneracy $\alpha = 1/2$, and in fact $\dim(V_{d,k})/\dim(V_{d,k}(\mathcal{G}_d)) \asymp D = d^{1/2}$. Our theory also applies to this group.*

4. Invariant kernel machines

Note that any invariant kernel of the form (4) can be written as a kernel of the form:

$$H_{d,\mathrm{inv}}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \int_{\mathcal{G}_d} \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\big[\sigma(\langle\boldsymbol{x}_1,\boldsymbol{w}\rangle)\sigma(\langle\boldsymbol{x}_2,g\cdot\boldsymbol{w}\rangle)\big]\pi_d(\mathrm{d}g)\,. \tag{10}$$

To see this, note that any inner product kernel h can be decomposed as

$$h(\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle/d) = \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\big[\sigma(\langle\boldsymbol{x}_1,\boldsymbol{w}\rangle)\sigma(\langle\boldsymbol{x}_2,\boldsymbol{w}\rangle)\big]$$

for some activation function σ, which amounts to taking the square root of the positive semidefinite operator associated to h. Substituting in Eq. (4), we get the desired representation.

Consider Kernel ridge regression with regularization parameter λ associated to $H_{d,\mathrm{inv}}$, that we call invariant KRR. Namely, we learn a function $\hat{f}^{\mathrm{inv}}_\lambda(\boldsymbol{x};\hat{\boldsymbol{u}}(\lambda)) = \sum_{i\in[n]}\hat{u}_i H_{d,\mathrm{inv}}(\boldsymbol{x}_i,\boldsymbol{x})$ where

$$\hat{\boldsymbol{u}}(\lambda) = \arg\min_{\boldsymbol{u}}\left\{\sum_{i=1}^n\big(y_i - \hat{f}^{\mathrm{inv}}_\lambda(\boldsymbol{x}_i;\boldsymbol{u})\big)^2 + \frac{\lambda}{d^\alpha}\|\hat{f}^{\mathrm{inv}}_\lambda(\,\cdot\,;\boldsymbol{u})\|^2_{\mathcal{H}}\right\}. \tag{11}$$

with $\|\cdot\|_{\mathcal{H}}$ the RKHS norm associated to $H_{d,\mathrm{inv}}$. We further denote the test error of invariant KRR by

$$R_{\mathrm{KR,inv}}(f_d,\boldsymbol{X},\boldsymbol{\varepsilon},\lambda) := \mathbb{E}_{\boldsymbol{x}}\left[\Big(f_d(\boldsymbol{x}) - \hat{f}^{\mathrm{inv}}_\lambda(\boldsymbol{x};\hat{\boldsymbol{u}})\Big)^2\right]. \tag{12}$$

Theorem 5 (Test error of invariant KRR) *Let $\mathcal{G}_d$ be a group of degeneracy $\alpha \le 1$ and $\{f_d \in L^2(\mathcal{A}_d,\mathcal{G}_d)\}_{d\ge1}$ be a sequence of $\mathcal{G}_d$-invariant functions. Assume $d^{\mathsf{s}-\alpha+\delta} \le n \le d^{\mathsf{s}+1-\alpha-\delta}$ for some fixed integer $\mathsf{s}\ge1$ and some $\delta>0$. Let σ be an activation function that satisfies Assumption 1 at level (s,s) (and $N=\infty$) and let $H_{d,\mathrm{inv}}$ be the associated invariant kernel as defined in Eq. (10). Then, the following holds for the test error of invariant KRR (c.f. Eq. (12)): for any $\lambda = O_d(1)$ (including $\lambda=0$ identically) any $\eta>0$, we have*

$$R_{\mathrm{KR,inv}}(f_d,\boldsymbol{X},\boldsymbol{\varepsilon},\lambda) = \|\overline{\mathsf{P}}_{>\mathsf{s}}f_d\|^2_{L^2} + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|^2_{L^{2+\eta}} + \sigma_\varepsilon^2). \tag{13}$$

We can compare the performance of this kernel against a standard (inner product) kernel $H_d(\boldsymbol{x},\boldsymbol{y}) = h_d(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d)$. Then Theorem 4 in Ghorbani et al. (2021) shows that the above theorem holds but with $d^{\mathsf{s}+\delta} \le n \le d^{\mathsf{s}+1-\delta}$. We gain a factor d^{α} in sample complexity by using an invariant kernel.

Remark 6 *Recall that the neural tangent kernel (NTK) associated to a function $f(\boldsymbol{x};\boldsymbol{\Theta})$ with random initialization $\boldsymbol{\Theta}_0$ is defined as*

$$H_{\mathrm{NT}}(\boldsymbol{x},\boldsymbol{y}) := \mathbb{E}_{\boldsymbol{\Theta}_0}\Big[\langle \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{x};\boldsymbol{\Theta}_0), \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{y};\boldsymbol{\Theta}_0)\rangle\Big].$$

The neural tangent kernel associated to a multi-layers fully connected network is an inner-product kernel (as long as the weights are initialized to be isotropic Gaussian.) In contrast, the NTK associated to the CNN of Eq. (1) is an example of invariant kernel, and is covered by Theorem 5 (see Appendix A.3 for more details).

5. Comparison with alternative approaches

To provide further context, it is useful to compare invariant random features and kernel models with other approaches. Here we consider two alternatives: (i) *output symmetrization*, which uses a non-invariant method for training and then symmetrizes the estimated function over the group $\mathcal{G}_d$ to obtain an invariant function; (ii) *data augmentation*, which trains the model on a dataset augmented by samples obtained by applying group transformations to the original data. As shown in Li et al. (2019), data augmentation is mathematically equivalent to invariant kernel methods, so that it is superior to standard kernel methods (with inner-product kernels). On the other hand, we show that output symmetrization of standard kernel estimators does not significantly improve over the standard kernel estimator, and is fundamentally sub-optimal comparing to invariant kernel methods.

5.1. Output symmetrization

Given an estimater $\hat{f}$, the symmetrization operator $\mathcal{S}\hat{f}$ computes the average of $\hat{f}$ over the group:

$$(\mathcal{S}\hat{f})(\boldsymbol{x}) \equiv \int_{\mathcal{G}_d} \hat{f}(g\cdot\boldsymbol{x})\pi_d(\mathrm{d}g). \tag{14}$$

When the target function f_d is $\mathcal{G}_d$-invariant, one might naively think that the symmetrization operation will significantly improve the performance of standard kernel estimators (standard RFRR and KRR). Indeed, when $f_d \in L^2(\mathcal{A}_d,\mathcal{G}_d)$, Jensen's inequality gives $\|f_d - \mathcal{S}\hat{f}\|_{L^2}^2 = \|\mathcal{S}(f_d-\hat{f})\|_{L^2}^2 \le \|f_d-\hat{f}\|_{L^2}^2$. However, the proposition below (which is proved in Section A.4) shows that $\mathcal{S}\hat{f}$ is not significantly better when $\hat{f}$ is a standard kernel estimator.

Proposition 7 *Let $f_d \in L^2(\mathcal{A}_d,\mathcal{G}_d)$ be a sequence of target functions. For any sequence of estimators $\hat{f}_d$ satisfying $\|\hat{f}_d - \overline{\mathsf{P}}_{\le \ell} f_d\|_{L^2}^2 \le \varepsilon$, we have*

$$\begin{aligned}\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}^2 - 2\varepsilon\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2} &\le \|f_d - \mathcal{S}\hat{f}_d\|_{L^2}^2 \\ &\le \|f_d-\hat{f}_d\|_{L^2}^2 \le \|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}^2 + 2\varepsilon\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2} + \varepsilon^2.\end{aligned} \tag{15}$$

Now consider —to be definite— a setting in which $N \geq nd^{\delta}$ and $d^{\ell+\delta} \leq n \leq d^{\ell+1-\delta}$, and $\mathcal{G}_d$ is a group with degeneracy 1. For any $f_d \in L^2(\mathcal{A}_d, \mathcal{G}_d)$ with $\|f_d\|_{L^{2+\eta}}^2 = O_d(1)$, the results of Mei et al. (2021) imply that standard RFRR (c.f. Eq. (5)) with sufficiently small regularization returns a function $\hat{f}_{\rm RF}$ with $\|\overline{\mathsf{P}}_{\leq \ell} f_d - \hat{f}_{\rm RF}\|_{L^2}^2 = o_{d,\mathbb{P}}(1)$. Consequently, Proposition 7 implies that we have

$$\|f_d - \mathcal{S}\hat{f}_{\rm RF}\|_{L^2}^2 = \|f_d - \hat{f}_{\rm RF}\|_{L^2}^2 + o_{d,\mathbb{P}}(1) = \|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1),$$

while Theorem 2 implies that invariant RFRR $\hat{f}_{\rm RF}^{\rm inv}$ with sufficiently small regularization achieves a substantially smaller risk:

$$\|f_d - \hat{f}_{\rm RF}^{\rm inv}\|_{L^2}^2 = \|\overline{\mathsf{P}}_{>\ell+1} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1).$$

5.2. Data augmentation

We consider full data augmentation whereby we replace each sample $(y_i, \boldsymbol{x}_i)$ in the dataset by $|\mathcal{G}_d|$ samples $\{(y_i, g \cdot \boldsymbol{x}_i) : g \in \mathcal{G}_d\}$ (for simplicity we consider here the case of a finite group $\mathcal{G}_d$), and perform standard KRR on the augmented dataset. One might naively think that this is not as effective as enforcing invariance in the kernel structure. After all, we are only requiring invariance to hold at the sampled points. However, Li et al. (2019) showed that these two approaches are in fact equivalent.

We compare KRR using the kernel $H(\boldsymbol{x}, \boldsymbol{y}) = h(\langle \boldsymbol{x}, \boldsymbol{y}\rangle/d)$ on the augmented dataset, with invariant KRR on the original dataset using the symmetrized kernel $H_{\rm inv}(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathcal{G}_d} h(\langle \boldsymbol{x}, g \cdot \boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g)$. Denote by $\hat{f}_\lambda^{\rm data}$ and $\hat{f}_\lambda^{\rm inv}$ the KRR estimates with the standard kernel H and full data augmentation, and with the invariant kernel $H_{\rm inv}$ respectively.

Proposition 8 (Li et al. (2019)) *Let $\mathcal{G}$ be a finite group, and H, $H_{\rm inv}$ as defined above. Then we have $\hat{f}_\lambda^{\rm data} = \hat{f}_\lambda^{\rm inv}$.*

A couple of remarks are in order. First, this equivalence is general (holds for any dataset $\{(y_i, \boldsymbol{x}_i)\}_{i\leq n}$), and is in fact a consequence of the algebraic structure of ridge regressions. Second, while this result establishes that the two approaches are mathematically equivalent, there are computational advantages for invariant KRR. Indeed, full data augmentation increases the size of the kernel matrix from n to $n|\mathcal{G}_d|$ which is computationally more expensive. Finally, this equivalence shows that data augmentation with standard KRR is superior to output symmetrization of standard KRR.

6. Numerical illustration

To check our predictions, we first consider the setting of $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ with $d = 30$, and three cyclic invariant polynomials $f_{d,\rm lin}$, $f_{d,\rm quad}$, $f_{d,\rm cube} \in L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$ defined as

$$f_{d,\rm lin} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i, \quad f_{d,\rm quad} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i x_{i+1}, \quad f_{d,\rm cube} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i x_{i+1} x_{i+2}, \tag{16}$$

where the sub-index i in x_i should be understood in the modulo d sense ($d + 1 = 1 \pmod d$). We compare the performance between two kernels: a standard (inner product) kernel $H_d(\boldsymbol{x}, \boldsymbol{y}) := h_d(\langle \boldsymbol{x}, \boldsymbol{y}\rangle/d)$ that we take to be the neural tangent kernel associated to a depth-5 neural network

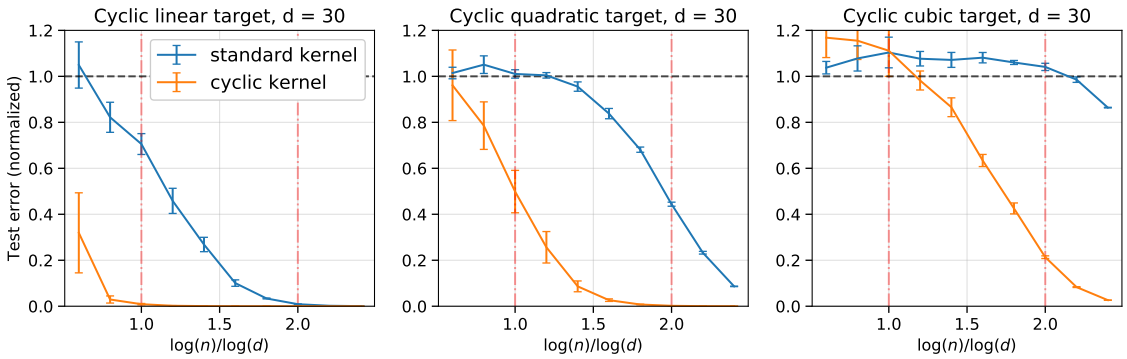

Figure 1: Learning cyclic polynomials (cf. Eq. (16)) over the d-dimensional sphere, $d = 30$, using KKR with a standard (inner-product) kernel and a cyclic invariant kernel, and regularization parameter $\lambda = 0^+$. We report the average and the standard deviation of the test error over 10 realizations, against the sample size n.

with fully connected layers and ReLu activations $\sigma(x) = \max(x, 0)$. We compare this with its cyclically invariant counterpart $H_{d,\mathrm{Cyc}}(\boldsymbol{x}, \boldsymbol{y}) = d^{-1} \sum_{0 \le i < d} h_d(\langle \boldsymbol{x}, g_i \cdot \boldsymbol{y} \rangle / d)$, where $g_i \in \mathrm{Cyc}_d$ is the shift by i positions as defined in Example 1. Note that the precise number of layers L is not important. As long as L is fixed in the large N, n limit, our predictions remain unchanged, and the simulations appear to confirm this.

In Figure 1, we report the test errors of fitting each cyclic polynomials with KRR with the two kernels, and regularization parameter $\lambda = 0^+$ (min-norm interpolation). We consider $\sigma_\varepsilon = 0$ and we report the risk averaged over 10 instances against the number of samples n. We observe that the risk in fitting $f_{d,\mathrm{lin}}$, $f_{d,\mathrm{quad}}$ and $f_{d,\mathrm{cube}}$, using KRR with the cylcic kernel $H_{d,\mathrm{Cyc}}$ drops when $n = \Theta_d(1)$, $n = \Theta_d(d)$ and $n = \Theta_d(d^2)$ respectively. In contrast, the risk of KRR with the standard kernel drops when $n = \Theta_d(d)$, $n = \Theta_d(d^2)$ and $n = \Theta_d(d^3)$ respectively. This matches well the predictions of Theorem 5.

We next investigate the relevance of our results for real data. We consider the MNIST dataset ($d = 28 \times 28 = 784$, $n_{\mathrm{train}} = 60000$, $n_{\mathrm{test}} = 10000$ and 10 classes). We encoded class labels by $y_i \in \{-4.5, -3.5, \dots, 3.5, 4.5\}$. We make these data invariant under cyclic translations in two dimensions (Example 2): for each samples in the training and test sets, we replace the image by a uniformly generated 2 dimensional (cyclic) translation of the image (see Fig. 5 in Appendix A.5.2). In this cyclic invariant MNIST data set, the labels are therefore invariant under the action of $\mathrm{Cyc2D}_{28,28}$.

Images are highly anisotropic in pixel space $\mathbb{R}^{784}$. In particular, directions corresponding to low-frequency components of the Fourier transform of $\boldsymbol{x}$ have significantly larger variance than directions corresponding to high-frequency components. Nevertheless, Ghorbani et al. (2020), showed that the analysis of random features and kernel models of Ghorbani et al. (2021); Mei et al. (2021) extends to certain anisotropic models provided the ambient dimension d is replaced by a suitably defined effective dimension d_{eff}.

In order to explore the role of data anisotropy, we pre-process images as follows. We compute the discrete Fourier transform components of the images in the training set and select the $T \in \{20, 70, 120, 200, 400, 784\}$ components with the highest average absolute value. For each T, we

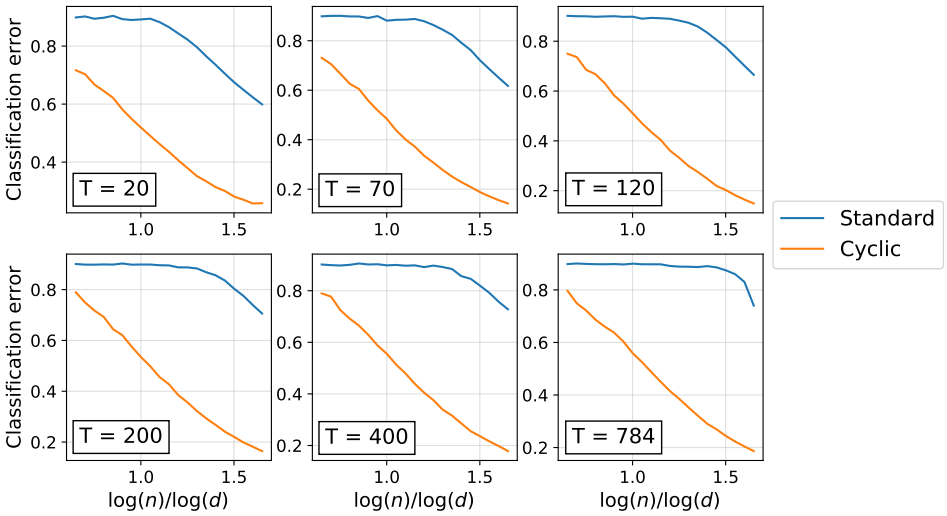

Figure 2: Classification error for the cyclic invariant MNIST dataset. For each frequencies content T, we plot the classification error averaged over 5 instances against the number of samples $\log(n)/\log(d)$, for KRR using a standard (inner-product) kernel and a cyclic invariant kernel and regularization parameter $\lambda = 0^+$.

then construct training and test sets in which we project each image onto the top T frequencies (see Fig. 3 in Appendix A.5.2). When T is small, we expect all the non-zero frequencies to have comparable variance and therefore $d_{\text{eff}} \approx T$. For larger T, we include frequencies of progressively small variance, and therefore d_{eff} should saturate.

For each frequency content T, we compare the performance of two kernels: a standard inner-product kernel $H_d(\boldsymbol{x}, \boldsymbol{y}) := h_d(\langle \boldsymbol{x}, \boldsymbol{y} \rangle / d)$ and its cyclic counterpart given by $H_{d,\text{Cyc}}(\boldsymbol{x}, \boldsymbol{y}) = 1/(28^2) \sum_{0 \le i,j < 28} h_d(\langle \boldsymbol{x}, g_{ij} \cdot \boldsymbol{y} \rangle / d)$, where $g_{ij} \in \text{Cyc2D}_{28,28}$. We choose H_d to be the neural tangent kernel associated to a two-layers neural network, and hence $H_{d,\text{Cyc}}$ is the one associated to a CNN analogous to (1) (but in two dimensions). We compute the KRR estimates with regularization parameter $\lambda = 0^+$. In Fig. 2, we report the classification error averaged over 5 instances against the number of samples $\log(n)/\log(d)$.

We observe that the cyclic invariant kernel vastly outperform the inner product kernel: the same test error is achieved at a significantly smaller sample size, in qualitative agreement with our general theory. In order to quantify this gap, for each T we fit two curves to the test error of the two kernels, which differ uniquely in an horizontal shift (see Appendix A.5.2). We estimate the sample complexity gain by the difference between these shifts, and denote this estimate by d_{eff}.

It is visually clear that d_{eff} increases with T, as expected. We plot d_{eff} as a function of T in Fig. 6 in Appendix A.5.2. We observe that the behavior of d_{eff} roughly matches our expectations: it grows linearly at small T and eventually saturates.

Acknowledgments

This work was supported by NSF through award DMS-2031883 and from the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. We also acknowledge NSF grants CCF-2006489, IIS-1741162 and the ONR grant N00014-18-1-2729.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, pages 6155–6166, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6676–6688. 2019b.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019a. URL http://proceedings.mlr.press/v97/arora19a.html.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019b.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, pages 159–182, 1975.

William Beckner. Sobolev inequalities, the Poisson semigroup, and analysis on the sphere S^n. *Proceedings of the National Academy of Sciences*, 89(11):4816–4819, 1992.

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *arXiv preprint arXiv:1905.12173*, 2019.

Aline Bonami. Etude des coefficients de Fourier des fonctions de $L^p(G)$. In *Annales de l'institut Fourier*, volume 20, pages 335–402, 1970.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8168–8177, 2018.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Zhiyuan Li, Yi Zhang, and Sanjeev Arora. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020a.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020b.

Chao Ma, Stephan Wojtowytsch, Lei Wu, et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Contents

Appendix A. Some details in the main text

A.1. Approximation power of invariant networks

In the proposition below, we show that the approximation power of two-layers $\mathcal{G}_d$-invariant neural networks are always no worse than two-layers fully-connected neural networks when the target function is $\mathcal{G}_d$-invariant.

Proposition 9 *Let $\sigma \in C(\mathbb{R})$ be an activation function. Let $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{Q}^d\}$. Let $\mathcal{G}_d$ be a subgroup of $\mathcal{O}(d)$ that preserves $\mathcal{A}_d$. Let π_d be the Haar measure of $\mathcal{G}_d$. Let $f_* \in L^2(\mathcal{A}_d; \mathcal{G}_d)$ be a $\mathcal{G}_d$-invariant function. Define the function classes of two-layers invariant neural networks and two-layers fully-connected neural networks by*

$$\mathcal{F}_{\mathsf{NN},\mathcal{G}_d,N} = \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{\theta}_i, g_\ell \cdot \boldsymbol{x} \rangle / \sqrt{d}) \pi_d(\mathrm{d}g) : \boldsymbol{\theta}_i \in \mathcal{A}_d, a_i \in \mathbb{R} \Big\}, \tag{17}$$

$$\mathcal{F}_{\mathsf{NN},N} = \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / \sqrt{d}) : \boldsymbol{\theta}_i \in \mathcal{A}_d, a_i \in \mathbb{R} \Big\}. \tag{18}$$

Then we have

$$\inf_{f \in \mathcal{F}_{\mathsf{NN},\mathcal{G}_d,N}} \|f_* - f\|_{L^2}^2 \le \inf_{f \in \mathcal{F}_{\mathsf{NN},N}} \|f_* - f\|_{L^2}^2.$$

Proof [Proof of Proposition 9]

We define the symmetrization operator $\mathcal{S} : L^2(\mathcal{A}_d) \to L^2(\mathcal{A}_d; \mathcal{G}_d)$ by

$$(\mathcal{S}f)(\boldsymbol{x}) = \int_{\mathcal{G}_d} f(g \cdot \boldsymbol{x}) \pi_d(\mathrm{d}g).$$

Since $f_* \in L^2(\mathcal{A}_d; \mathcal{G}_d)$, by Jensen's inequality, for any $f \in L^2(\mathcal{A}_d)$, we have

$$\|f_\star - \mathcal{S}f\|_{L^2}^2 = \|\mathcal{S}(f_\star - f)\|_{L^2}^2 \le \|f_\star - f\|_{L^2}^2.$$

Moreover, for any $f \in \mathcal{F}_{\mathsf{NN},N}$, we have $\mathcal{S}f \in \mathcal{F}_{\mathsf{NN},\mathcal{G}_d,N}$. This gives

$$\inf_{f \in \mathcal{F}_{\mathsf{NN},\mathcal{G}_d,N}} \|f_* - f\|_{L^2}^2 \le \inf_{f \in \mathcal{F}_{\mathsf{NN},N}} \|f_\star - \mathcal{S}f\|_{L^2}^2 \le \inf_{f \in \mathcal{F}_{\mathsf{NN},N}} \|f_\star - f\|_{L^2}^2.$$

This concludes the proof. ■

A.2. Intuition for the proofs of Theorems 2 and 5

Theorem 2 and 5 are consequences of general theorems proved in Mei et al. (2021). The d^α improvement between invariant and non-invariant models can be understood as follows: consider an inner-product activation $\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d})$ with $\boldsymbol{x}, \boldsymbol{\theta} \sim \mathrm{Unif}(\mathcal{A}_d)$ (where we denoted $\boldsymbol{\theta} = \sqrt{d} \cdot \boldsymbol{w}$), then we have the following eigendecomposition

$$\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d}) = \sum_{k=0}^{\infty} \xi_{d,k}^2 \sum_{l=1}^{B(\mathcal{A}_d;k)} Y_{kl}^{(d)}(\boldsymbol{x}) Y_{kl}^{(d)}(\boldsymbol{\theta}),$$

where $\{Y_{kl}^{(d)}\}_{l\in[B(\mathcal{A}_d;k)]}$ form an orthonormal basis of $V_{d,k}$, the subspace of degree-k polynomials on $\mathcal{A}_d$ (see Section H for background on functional spaces on the sphere and hypercube). The eigenvalues of σ are given by $\{\xi_{d,k}\}_{k\ge 0}$ with each having degeneracy $B(\mathcal{A}_d;k)$.

As mentioned in the introduction, the symmetry group $\mathcal{G}_d$ preserves $V_{d,k}$ (see Section C.2) and the invariant activation function has the following eigendecomposition

$$\overline{\sigma}(\boldsymbol{x};\boldsymbol{\theta}) := \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{x}, g\cdot\boldsymbol{\theta}\rangle/\sqrt{d})\pi_d(\mathrm{d}g) = \sum_{k=0}^{\infty} \xi_{d,k}^2 \sum_{l=1}^{D(\mathcal{A}_d;k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{x})\overline{Y}_{kl}^{(d)}(\boldsymbol{\theta})\,,$$

where the $\{\overline{Y}_{kl}^{(d)}\}_{l\in[B(\mathcal{A}_d;k)]}$ form an orthonormal basis of $V_{d,k}(\mathcal{G}_d)$, the subspace of degree-k invariant polynomials on $\mathcal{A}_d$. The eigenvalues of $\overline{\sigma}$ are given by $\{\xi_{d,k}\}_{k\ge 0}$ with each having degeneracy $D(\mathcal{A}_d;k)$.

Hence $\overline{\sigma}$ has the same eigenvalues $\xi_{d,k}$ as σ, but with degeneracy smaller by a factor

$$\frac{B(\mathcal{A}_d;k)}{D(\mathcal{A}_d;k)} = \Theta_d(d^{\alpha})\,.$$

In other words, in order to fit degree ℓ polynomials using invariant methods, one needs to fit a factor d^{α} less eigendirections, which translates to a factor d^{α} improvement in the sample and features complexity.

This intuition is verified rigorously in the proof of these theorems in Appendix B.

A.3. Convolutional neural tangent kernel

Proposition 10 *Let $\sigma \in C^1(\mathbb{R})$ be an activation function. Let $\mathcal{G}_d$ be a discrete subgroup of $\mathcal{O}(d)$ with Haar measure π_d. Let f_N be an invariant neural network*

$$f_N(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{i=1}^{N} a_i \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{w}_i, g\cdot\boldsymbol{x}\rangle)\pi_d(\mathrm{d}g).$$

Let $a_i^0 \sim_{i.i.d.} \mathcal{N}(0,1)$ and $\boldsymbol{w}_i^0 \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1})$ independently, and $\boldsymbol{\Theta}^0 = (a_1^0,\dots,a_N^0,\boldsymbol{w}_1^0,\dots,\boldsymbol{w}_N^0)$. Then there exists $h_d : [-1,1]\to\mathbb{R}$, such that for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{S}^{d-1}(\sqrt{d})$, we have almost surely

$$\lim_{N\to\infty} \langle \nabla_{\boldsymbol{\Theta}} f_N(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{\Theta}} f_N(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle/N = \int_{\mathcal{G}_d} h_d(\langle \boldsymbol{x}, g\cdot\boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g).$$

Proof [Proof of Proposition 10] For $\boldsymbol{x},\boldsymbol{y}\in\mathbb{S}^{d-1}(\sqrt{d})$, define

$$\begin{aligned} h_d^{(1)}(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d) =& \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w},\boldsymbol{y}\rangle)],\\ h_d^{(2)}(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d) =& \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma'(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w},\boldsymbol{y}\rangle)\langle \boldsymbol{x},\boldsymbol{y}\rangle]. \end{aligned}$$

By the technical backgrounds in Section H, we can see that $h_d^{(1)}$ and $h_d^{(2)}$ can be well-defined.

Calculating the derivative of the neural network with respect to $\boldsymbol{a} = (a_1,\dots,a_N)$, we have

$$\frac{1}{N}\langle \nabla_{\boldsymbol{a}} f(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{a}} f(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle = \int_{\mathcal{G}_d\times\mathcal{G}_d} \frac{1}{N}\sum_{i=1}^{N}\Big[\sigma(\langle \boldsymbol{w}_i, g\cdot\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}_i, g'\cdot\boldsymbol{y}\rangle)\Big]\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g').$$

Since $\mathcal{G}_d$ is a discrete group, by law of large numbers, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\langle \nabla_{\boldsymbol{a}} f(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{a}} f(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle \\
&= \int_{\mathcal{G}_d\times\mathcal{G}_d} \mathbb{E}_{\boldsymbol{w}}[\sigma(\langle \boldsymbol{w}, g\cdot\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, g'\cdot\boldsymbol{y}\rangle)]\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g') \\
&= \int_{\mathcal{G}_d\times\mathcal{G}_d} h_d^{(1)}(\langle g\cdot\boldsymbol{x}, g'\cdot\boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g') \\
&= \int_{\mathcal{G}_d} h_d^{(1)}(\langle \boldsymbol{x}, g\cdot\boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g).
\end{aligned}$$

Moreover, calculating the derivative of the neural network with respect to $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$, we have

$$\begin{aligned}
&\frac{1}{N}\langle \nabla_{\boldsymbol{W}} f(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{W}} f(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle \\
&= \int_{\mathcal{G}_d\times\mathcal{G}_d} \frac{1}{N}\sum_{i=1}^N \Big[(a_i^0)^2\sigma'(\langle \boldsymbol{w}_i, g\cdot\boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w}_i, g'\cdot\boldsymbol{y}\rangle)\langle g\cdot\boldsymbol{x}, g'\cdot\boldsymbol{y}\rangle\Big]\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g')
\end{aligned}$$

Since $\mathcal{G}_d$ is a discrete group, by law of large numbers, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\langle \nabla_{\boldsymbol{W}} f(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{W}} f(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle \\
&= \int_{\mathcal{G}_d\times\mathcal{G}_d} \mathbb{E}_{\boldsymbol{w}}[\sigma'(\langle \boldsymbol{w}, g\cdot\boldsymbol{x}\rangle)\sigma'(\langle \boldsymbol{w}, g'\cdot\boldsymbol{y}\rangle)\langle g\cdot\boldsymbol{x}, g'\cdot\boldsymbol{y}\rangle]\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g') \\
&= \int_{\mathcal{G}_d\times\mathcal{G}_d} h_d^{(2)}(\langle g\cdot\boldsymbol{x}, g'\cdot\boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g)\pi_d(\mathrm{d}g') \\
&= \int_{\mathcal{G}_d} h_d^{(2)}(\langle \boldsymbol{x}, g\cdot\boldsymbol{y}\rangle/d)\pi_d(\mathrm{d}g).
\end{aligned}$$

Taking $h_d = h_d^{(1)} + h_d^{(2)}$ concludes the proof. ∎

A.4. Proof of Proposition 7

Let $\hat{f}_d$ be an estimator satisfying

$$\varepsilon^2 := \|\overline{\mathsf{P}}_{\le\ell} f_d - \hat{f}_d\|_{L^2}^2 = \|\overline{\mathsf{P}}_{\le\ell} f_d - \overline{\mathsf{P}}_{\le\ell}\hat{f}_d\|_{L^2}^2 + \|\overline{\mathsf{P}}_{>\ell}\hat{f}_d\|_{L^2}^2. \tag{19}$$

By Jensen's inequality and by the equation above, we have

$$\|\mathcal{S}\overline{\mathsf{P}}_{>\ell}\hat{f}_d\|_{L^2}^2 \le \|\overline{\mathsf{P}}_{>\ell}\hat{f}_d\|_{L^2}^2 \le \varepsilon^2. \tag{20}$$

As a consequence, we have

$$\begin{aligned}
&\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}^2 - 2\varepsilon\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2} \overset{(1)}{\le} \|\overline{\mathsf{P}}_{>\ell} f_d - \mathcal{S}\overline{\mathsf{P}}_{>\ell}\hat{f}_d\|_{L^2}^2 \overset{(2)}{=} \|\overline{\mathsf{P}}_{>\ell} f_d - \overline{\mathsf{P}}_{>\ell}\mathcal{S}\hat{f}_d\|_{L^2}^2 \\
&\overset{(3)}{\le} \|f_d - \mathcal{S}\hat{f}_d\|_{L^2}^2 \overset{(4)}{=} \|\mathcal{S}(f_d - \hat{f}_d)\|_{L^2}^2 \overset{(5)}{\le} \|f_d - \hat{f}_d\|_{L^2}^2 \\
&\overset{(6)}{=} \|\overline{\mathsf{P}}_{\le\ell} f_d - \overline{\mathsf{P}}_{\le\ell}\hat{f}_d\|_{L^2}^2 + \|\overline{\mathsf{P}}_{>\ell} f_d - \overline{\mathsf{P}}_{>\ell}\hat{f}_d\|_{L^2}^2 \overset{(7)}{=} \|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}^2 + \varepsilon^2 + 2\varepsilon\|\overline{\mathsf{P}}_{>\ell} f_d\|_{L^2}.
\end{aligned} \tag{21}$$

Here, (1) is by Eq. (20); (2) is by the fact that $\mathcal{S}$ is exchangable with $\overline{\mathsf{P}}_{>\ell}$ (c.f. Section C); (3) is by the fact that $\overline{\mathsf{P}}_{>\ell}$ is a projection operator; (4) is by the fact that f_d is $\mathcal{G}_d$-invariant; (5) is by Jensen's inequality; (6) is by orthogonal decomposition; (7) is by Eq. (19). This concludes the proof.

A.5. Details of numerical simulations

A.5.1. SYNTHETIC DATA

We consider the standard (inner-product) kernel $H_d(\boldsymbol{x},\boldsymbol{y}) = h_{\rm NTK}(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d)$ to be the neural tangent kernel associated to a depth-5 neural network with fully connected layers and ReLu activation $\sigma(x) = \max(x,0)$. This can be obtained iteratively as follow (see Jacot et al. (2018) and Bietti and Bach (2020)): define for $u \in [-1,1]$,

$$h_0(u) = \frac{1}{\pi}(\pi - \arccos(u)), \qquad h_1(u) = u\cdot h_0(u) + \frac{1}{\pi}\sqrt{1-u^2}\,,$$

and $h_{\rm NTK}(u) = h^5_{\rm NTK}(u)$ with $h^1_{\rm NTK}(u) = h^1(u) = u$ and for $k = 2,\dots,5$,

$$\begin{aligned} h^k(u) &= h_1(h^{k-1}(u))\,,\\ h^k_{\rm NTK}(u) &= h^{k-1}_{\rm NTK}(u)h_0(h^{k-1}(u)) + h^k(u)\,. \end{aligned}$$

We compute the cyclic invariant kernel by summing over all cyclic translations $g \in {\rm Cyc}_d$:

$$H_{d,{\rm inv}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{g\in{\rm Cyc}_d} h_{\rm NTK}(\langle \boldsymbol{x}, g\cdot\boldsymbol{y}\rangle/d)\,.$$

A.5.2. CYCLIC INVARIANT MNIST DATA SET

We consider the MNIST data set of 28×28 grayscale images ($d = 784$) of handwritten digits, which contains 60000 training images and 10000 testing images. We pre-process the images in three steps:

(a) We compute the discrete Fourier transform of the images in the training set and compute the average absolute value of the frequency components (see left frame of Fig. 3). For each $T \in \{20, 70, 120, 200, 400, 784\}$, we select $\Omega_T \subset [28]\times[28]$ to be the set of the top T frequencies (i.e., the T frequencies with highest absolute value averaged on the training set).

(b) For each T, we construct a train and test sets in which we project each image onto Ω_T (i.e., we set all the frequency components not in Ω_T to 0). We displayed in Fig. 4 two digits and their projection on the top T frequencies Ω_T for different T.

(c) For each image in the training and test sets, we replace the image by a uniformly generated 2 dimensional (cyclic) translation of the image. We display some examples in Fig. 5.

We further normalize the images so that $\|\boldsymbol{x}\|_2 = 1$ and center the labels $y_i \in \mathcal{Y}$ where $\mathcal{Y} = \{-4.5, -3.5, \dots, 3.5, 4.5\}$. In order to compute the classification error, we round the prediction value to the nearest label in $\mathcal{Y}$.

We use the inner-product kernel $H_d(\boldsymbol{x},\boldsymbol{y}) = h_{\rm NTK}(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d)$ where $h_{\rm NTK}$ is the neural tangent kernel associated to a 2-layers neural network with fully connected layers and ReLu activation $\sigma(x) = \max(x,0)$, which given by

$$h_{\rm NTK}(u) = u\cdot\Big(\pi - \frac{\arccos(u)}{\pi}\Big) + \frac{1}{\pi}\sqrt{1-u^2}\,.$$

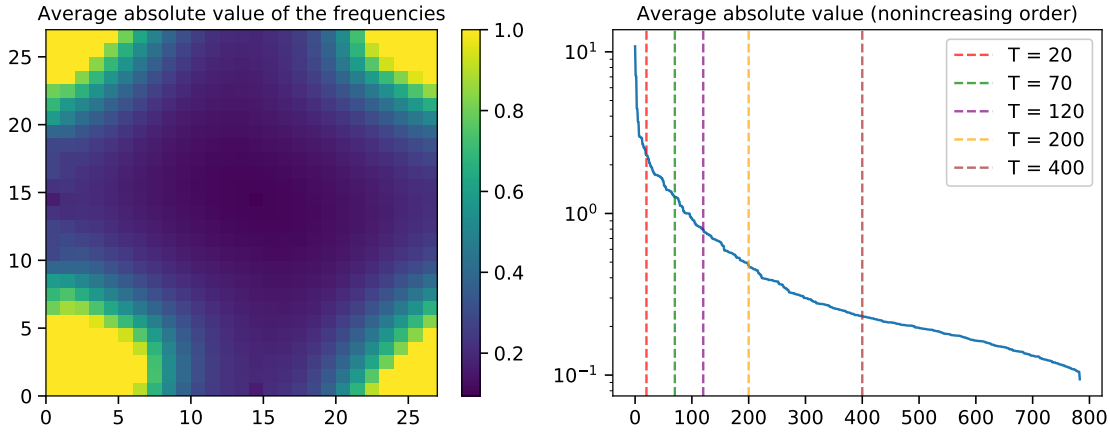

Figure 3: Left frame: the absolute value of the frequency components of MNIST images averaged over the training set (threshold at 1 in the figure). Coordinates on the bottom left-hand side correspond to lower frequency components while coordinates closer to the top right-hand side represent the high frequency directions. Right frame: average absolute value of the frequencies in nonincreasing order. The vertical lines correspond to the different T chosen ($T = 784$ corresponds to keeping all the frequencies).

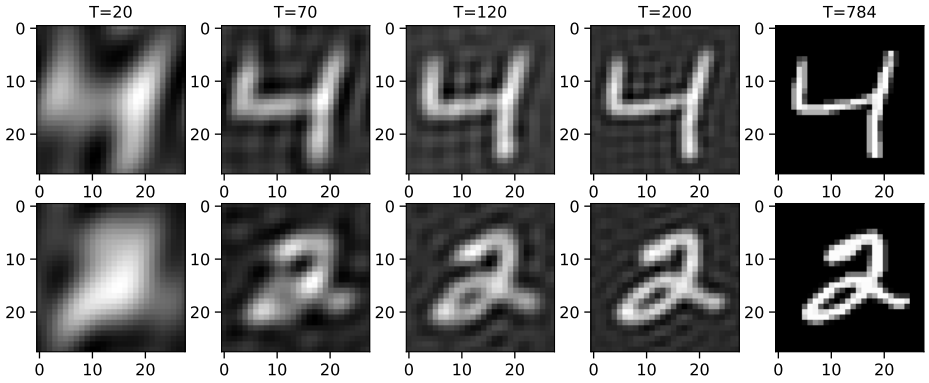

Figure 4: Examples of two images projected on the top T frequencies.

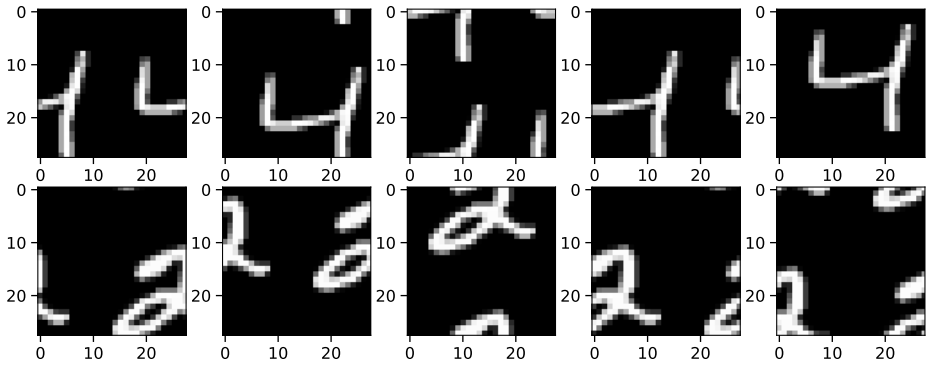

Figure 5: Examples of random 2-dimensional cyclic translations of the images (for $T = 784$).

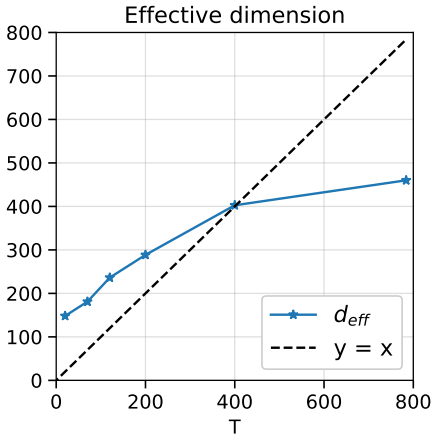

Figure 6: Estimated sample size gap between standard and invariant kernel methods, for the translationally invariant MNIST dataset, as a function of the frequency content T.

The cyclic invariant kernel is computed by summing over all two-dimensional cyclic translations $g_{ij} \in \mathrm{Cyc2D}_{28,28}$:

$$H_{d,\mathrm{inv}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{28^2} \sum_{i,j=0}^{27} h_{\mathrm{NTK}}(\langle \boldsymbol{x}, g_{ij} \cdot \boldsymbol{y} \rangle / d) \, .$$

For each T, we estimate the effective dimension d_{eff} by fitting two parallel lines through the classification error points of the standard and cyclic kernels at the same time (keeping only the points where the curves decrease). The estimated (log) effective dimension is then given by the difference of the offsets. We report these estimates for different T in Fig. 6.

Appendix B. Proof of the main theorems

In this section, we present the proofs of Theorem 2 and 5 stated in the main text. The rest of the appendices are organized as follow:

- Appendix C presents key properties of the decomposition of invariant functions, while Appendix H reviews some technical background on the functional spaces on the sphere and the hypercube.
- Appendix D proves that the examples of symmetry group listed in Section 2 (one and two-dimensional cyclic groups and band-limited functions) have degeneracy 1.
- Appendix E presents a key concentration result on the diagonal elements of polynomial invariant kernels. In particular, the results of Appendix E are the only ones required in the proofs of Theorems 2 and 5 in the case of polynomial activations for general symmetry group $\mathcal{G}_d$ of degeneracy $\alpha \le 1$.
- Appendices F and G provides necessary results to extend the proofs to non-polynomial activations in the case of $(\mathcal{A}_d, \mathcal{G}_d) = (\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$.

B.1. Notations

For a positive integer, we denote by $[n]$ the set $\{1, 2, \dots, n\}$. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we denote $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = u_1 v_1 + \ldots + u_d v_d$ their scalar product, and $\|\boldsymbol{u}\|_2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}$ the ℓ_2 norm. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, we denote $\|\boldsymbol{A}\|_{\mathrm{op}} = \max_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{u}\|_2$ its operator norm and by $\|\boldsymbol{A}\|_F = \left(\sum_{i,j} A_{ij}^2\right)^{1/2}$ its Frobenius norm. If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a square matrix, the trace of $\boldsymbol{A}$ is denoted by $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i \in [n]} A_{ii}$.

We use $O_d(\,\cdot\,)$ (resp. $o_d(\,\cdot\,)$) for the standard big-O (resp. little-o) relations, where the subscript d emphasizes the asymptotic variable. Furthermore, we write $f = \Omega_d(g)$ if $g(d) = O_d(f(d))$, and $f = \omega_d(g)$ if $g(d) = o_d(f(d))$. Finally, $f = \Theta_d(g)$ if we have both $f = O_d(g)$ and $f = \Omega_d(g)$.

We use $O_{d,\mathbb{P}}(\,\cdot\,)$ (resp. $o_{d,\mathbb{P}}(\,\cdot\,)$) the big-O (resp. little-o) in probability relations. Namely, for $h_1(d)$ and $h_2(d)$ two sequences of random variables, $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}(|h_1(d)/h_2(d)| > C_\varepsilon) \le \varepsilon, \qquad \forall d \ge d_\varepsilon,$$

and respectively: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability. Similarly, we will denote $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = O_{d,\mathbb{P}}(h_1(d))$, and $h_1(d) = \omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = o_{d,\mathbb{P}}(h_1(d))$. Finally, $h_1(d) = \Theta_{d,\mathbb{P}}(h_2(d))$ if we have both $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ and $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$.

B.2. Proof of Theorem 2

Let $\mathcal{G}_d$ be a group of degeneracy $\alpha \le 1$. Consider $\boldsymbol{x}, \boldsymbol{\theta} \sim \mathrm{Unif}(\mathcal{A}_d)$, $d^{\mathsf{s}-\alpha+\delta_0} \le n \le d^{\mathsf{s}-\alpha+1-\delta_0}$, $d^{\mathsf{S}-\alpha+\delta_0} \le N \le d^{\mathsf{S}-\alpha+1-\delta_0}$ and an activation function σ that satisfies Assumption 1 at level (s, S). Denote

$$\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}) = \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{x} \rangle / \sqrt{d}) \pi_d(\mathrm{d}g).$$

Theorem 2 is a consequence of Theorem 1 in Mei et al. (2021) where we take $\mathcal{X}_d = \Omega_d = \mathcal{A}_d$, $\nu_d = \tau_d = \mathrm{Unif}(\mathcal{A}_d)$ and $\mathcal{D}_d = \mathcal{V}_d = L^2(\mathcal{A}_d, \mathcal{G}_d) \subset L^2(\mathcal{A}_d)$. The proof amounts to checking that $\overline{\sigma}$ indeed verifies the feature map concentration and spectral gap assumptions (see Section 2.2 in Mei et al. (2021)). We borrow some of the notations introduced in Mei et al. (2021) and refer the reader to their Section 2.1.

Proof [Proof of Theorem 2] For the sake of simplicity, we consider the overparametrized case $N(d) \ge n(d) d^{\delta}$ for some $\delta > 0$, and therefore $\mathsf{S} \ge \mathsf{s}$. The underparametrized case $d^{\delta} N(d) \le n(d)$ is treated analogously.

Step 1. Diagonalization of the activation function $\overline{\sigma}$ and choosing $\mathsf{m} = \mathsf{m}(d)$, $\mathsf{M} = \mathsf{M}(d)$.

We can decompose the inner product activation σ in the basis of Gegenbauer polynomials (see Section H for definitions):

$$\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d}) = \sum_{k=0}^{\infty} \xi_{d,k} B(\mathcal{A}_d; k) Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle) \,,$$

where (with $\boldsymbol{e} \in \mathcal{A}_d$ arbitrary)

$$\xi_{d,k}(\sigma) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\mathcal{A}_d)} [\sigma(\langle \boldsymbol{e}, \boldsymbol{\theta} \rangle / \sqrt{d}) Q_k^{(d)}(\langle \boldsymbol{e}, \boldsymbol{\theta} \rangle)] \,.$$

From Assumption 1.(a) that $|\sigma(x)| \le c_0 \exp(c_1 x^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$ (which is trivially verified for a polynomial activation function), there exists a constant $C > 0$ such that (see for example Lemma 5 in Ghorbani et al. (2021))

$$\|\sigma(\langle \boldsymbol{e}, \, \cdot \, \rangle / \sqrt{d})\|_{L^2(\mathcal{A}_d)} = \sum_{k=1}^{\infty} \xi_{d,k}^2 B(\mathcal{A}_d; k) \le C. \tag{22}$$

We have for fixed k, $B(\mathcal{A}_d; k) = \Theta(d^k)$. Furthermore, for non-polynomial activation functions in the case of $(\mathcal{A}_d, \mathcal{G}_d) = (\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$, we use $\sup_{k > \mathsf{s}} B(\mathbb{S}^{d-1}; k)^{-1} = O_d(d^{-\mathsf{s}-1})$ (Lemma 1 in Ghorbani et al. (2021)). We deduce that

$$\sup_{k > \mathsf{s}} \xi_{d,k}^2 = O_d(d^{-\mathsf{s}-1}) \,, \tag{23}$$

$$\sup_{k > \mathsf{S}} \xi_{d,k}^2 = O_d(d^{-\mathsf{S}-1}) \,. \tag{24}$$

From the correspondence between Gegenbauer and Hermite polynomials when $d \to \infty$ (see Eq. (111) in Section H.1.3), Assumption 1.(b) implies that $\xi_{d,k}^2 = \Theta_d(d^{-k})$ for $k = 0, \dots, \mathsf{s}$.

Let us diagonalize $\overline{\sigma}$ in the basis of $\mathcal{G}_d$-invariant polynomials $\{\overline{Y}_{kl}\}_{k \ge 0, \ell \in [D(\mathcal{A}_d; k)]}$ (see Section C for definitions). From Lemma 11 stated in Section C.3, we have

$$\begin{aligned} \overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}) &= \sum_{k=0}^{\infty} \xi_{d,k} B(\mathcal{A}_d; k) \int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{x}, g \cdot \boldsymbol{\theta} \rangle) \pi_d(\mathrm{d}g) \\ &= \sum_{k=0}^{\infty} \xi_{d,k} \sum_{l=1}^{D(\mathcal{A}_d; k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{x}) \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta}) \,. \end{aligned} \tag{25}$$

Denote $(\lambda_{d,j})_{j \ge 1}$ the eigenvalues of $\overline{\sigma}$ in non increasing order of their absolute value (namely, the $\xi_{d,k}$'s which have degeneracies $D(\mathcal{A}_d; k)$). Set m and M to be the number of eigenvalues $\lambda_{d,j}^2$

that are bigger than $d^{-\mathsf{s}-1+\delta}$ and $d^{-\mathsf{S}-1+\delta}$ respectively, for a constant $\delta > 0$ that will be set sufficiently small (see Step 4). From the above discussion, $(\lambda_{d,j})_{j\leq \mathsf{m}}$ corresponds exactly to all the eigenvalues associated to invariant polynomials of degree less of equal to s, while $(\lambda_{d,j})_{j\leq \mathsf{M}}$ does not contain any eigenvalues associated to invariant polynomials of degree bigger or equal to $\mathsf{S}+1$. Hence,

$$\mathsf{m} = \sum_{k=0}^{\mathsf{s}} D(\mathcal{A}_d; k) = \Theta_d(d^{\mathsf{s}-\alpha}), \qquad \mathsf{M} \leq \sum_{k=0}^{\mathsf{S}} D(\mathcal{A}_d; k) = O_d(d^{\mathsf{S}-\alpha}), \tag{26}$$

where we used that $\mathcal{G}_d$ has degeneracy α so that $D(\mathcal{A}_d; k) = \Theta_d(d^{-\alpha}) \cdot B(\mathcal{A}_d; k)$.

Step 2. Diagonal elements of the truncated kernel.

We introduce the kernel associated to activation $\overline{\sigma}$:

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{\theta}}[\overline{\sigma}(\boldsymbol{x}_1; \boldsymbol{\theta})\overline{\sigma}(\boldsymbol{x}_2; \boldsymbol{\theta})] = \sum_{k=0}^{\infty} \xi_{d,k}^2 D(\mathcal{A}_d; k) \Upsilon_k^{(d)}(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

where we denote

$$\Upsilon_k^{(d)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{D(\mathcal{A}_d; k)} \sum_{l=1}^{D(\mathcal{A}_d;k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{x}_1) \overline{Y}_{kl}^{(d)}(\boldsymbol{y}_1).$$

Similarly, we introduce a kernel in the feature space

$$U_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}}[\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}_1)\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}_2)] = \sum_{k=0}^{\infty} \xi_{d,k}^2 D(\mathcal{A}_d; k) \Upsilon_k^{(d)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

We denote $\mathbb{H}_d, \mathbb{U}_d : L^2(\mathcal{A}_d, \mathcal{G}_d) \to L^2(\mathcal{A}_d, \mathcal{G}_d)$ the kernel operators with kernel representation H_d and U_d, and denote $\mathbb{H}_{d,>\mathsf{m}}$ and $\mathbb{U}_{d,>\mathsf{M}}$ the kernel operators where the biggest m and M eigenvalues respectively are set to 0. Recalling the discussion on the choice of m and M, denote $E = \{k : \xi_{d,k}^2 \leq d^{-\mathsf{S}-1+\delta}\}$: E contain all integers bigger or equal to $\mathsf{S}+1$ and none smaller or equal to s.

The diagonal elements of the truncated kernels are then given by

$$\begin{aligned} H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x}) &= \sum_{k=\mathsf{s}+1}^{\infty} \xi_{d,k}^2 D(\mathcal{A}_d; k) \Upsilon_k^{(d)}(\boldsymbol{x}, \boldsymbol{x}), \\ U_{d,>\mathsf{M}}(\boldsymbol{\theta}, \boldsymbol{\theta}) &= \sum_{k\in E} \xi_{d,k}^2 D(\mathcal{A}_d; k) \Upsilon_k^{(d)}(\boldsymbol{\theta}, \boldsymbol{\theta}), \end{aligned} \tag{27}$$

and

$$\begin{aligned} \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) &= \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x})] = \sum_{k=\mathsf{s}+1}^{\infty} \xi_{d,k}^2 D(\mathcal{A}_d; k), \\ \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) &= \mathbb{E}_{\boldsymbol{\theta}}[U_{d,>\mathsf{M}}(\boldsymbol{\theta}, \boldsymbol{\theta})] = \sum_{k\in E} \xi_{d,k}^2 D(\mathcal{A}_d; k). \end{aligned}$$

From Assumption 1.(c), σ is not a polynomial of degree less or equal to S. Hence, there exists $\ell > \mathsf{S}$ such that $\mu_\ell(\sigma) \neq 0$ and therefore $\xi_{d,\ell}^2 D(\mathcal{A}_d; \ell) = \Theta(d^{-\alpha})$. Furthermore, from Eq. (22) and the assumption that $\mathcal{G}_d$ is of degeneracy α (for polynomial activation functions, see Proposition 22 in Section G for general σ in the case of $(\mathcal{A}_d, \mathcal{G}_d) = (\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$), we have

$$\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \Theta(d^{-\alpha}), \qquad \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) = \Theta(d^{-\alpha}). \tag{28}$$

Step 3. Checking the feature map concentration property at level $\{N(d), \mathsf{M}(d), n(d), \mathsf{m}(d)\}_{d\ge 1}$.

Let us first consider the case of a polynomial activation function σ. Denote D its degree and $u = u(d)$ the total (finite) number of nonzero eigenvalues of $\overline{\sigma}$ (which are associated to invariant polynomials of degree less or equal to D). Let us verify the feature map concentration property (Assumption 1 in Mei et al. (2021)) with sequence $u(d) \ge \max(\mathsf{m}, \mathsf{M})$. Note that $u \ge \max(\mathsf{m}, \mathsf{M})$, part (b) and (c) of the property are trivially verified in that case.

(a) *(Hypercontractivity of finite eigenspaces on $\mathcal{D}_d$.)* The subspace of polynomials of degree less or equal to D on the hypercube and the sphere verifies the hypercontractivity property (see Lemmas 35 and 36 in Section H.3).

(d) *(Concentration of diagonal elements.)* From Eq. (27) and Proposition 18 stated in Section E, we have

$$\begin{aligned}&\sup_{i\in[n]}\left|H_{d,>\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{x})]\right|\\ &\le \sum_{k=\mathsf{s}+1}^{D} \xi_{d,k}^2 D(\mathcal{A}_d;k)\sup_{i\in[n]}\left|\Upsilon_k^{(d)}(\boldsymbol{x}_i,\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Upsilon_k^{(d)}(\boldsymbol{x},\boldsymbol{x})]\right| = o_{d,\mathbb{P}}(1)\cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{x})].\end{aligned}$$

A similar computation shows the concentration of the diagonal elements of $U_{d,>\mathsf{M}}$.

Let us now consider a non polynomial activation function σ in the case of $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$ and $\mathcal{G}_d = \mathrm{Cyc}_d$ (of degeneracy 1). Let us choose $\ell > 2\mathsf{S} + 10$ such that $\mu_\ell(\sigma) \neq 0$ (it must exists otherwise σ would be a polynomial) and therefore $\xi_{d,\ell}^2 = \Theta_d(d^{-\ell})$. Consider $u = u(d)$ to be the number of eigenvalues such that $\lambda_{d,j}^2$ is strictly bigger than $\xi_{d,\ell}^2$. Then, $(\lambda_{d,j})_{j\le u}$ do not contain any eigenvalues $\xi_{d,k}$ for $k \ge \ell$ and contain all $\xi_{d,k}$ for $k \le \mathsf{s}$. In particular, $u \ge \max(\mathsf{m}, \mathsf{M})$. Denote $E = \{k : \xi_{d,k}^2 \le \xi_{d,\ell}^2\}$: E contain all integers bigger or equal to ℓ.

Let us verify the feature map concentration property with the sequence $u(d)$ (part (a) is the same with D replaced by $\ell - 1$).

(b) *(Properly decaying eigenvalues.)* We have

$$\begin{aligned}&\mathrm{Tr}(\mathbb{H}_{d,>u}) \ge \xi_{d,\ell}^2 D(\mathbb{S}^{d-1};\ell) = \Omega_d(d^{-\alpha})\,,\\ &\mathrm{Tr}(\mathbb{H}_{d,>u}^2) = \sum_{k\in E}\xi_{d,k}^4 D(\mathbb{S}^{d-1};k) \le \xi_{d,\ell}^2 \mathrm{Tr}(\mathbb{H}_{d,>u})\,.\end{aligned}$$

Hence,

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>u})^2}{\mathrm{Tr}(\mathbb{H}_{d,>u}^2)} \ge \xi_{d,\ell}^{-2}\cdot \mathrm{Tr}(\mathbb{H}_{d,>u}) = \Omega_d(1)\cdot d^{2\mathsf{S}+9} \ge \max(n,N)^{2+\delta}.$$

(c) *(Hypercontractivity of the high degree part.)* Denote $\overline{\sigma}_{>u} = \overline{\mathsf{P}}_E\overline{\sigma}$ the activation $\overline{\sigma}$ obtained by setting the first u eigenvalues to 0 (i.e., setting coefficients $k \notin E$ to zero in Eq. (25)). From Eq. (28), we need to show that for p as defined in Assumption 1.(a), we have

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\overline{\sigma}_{>u}(\boldsymbol{x};\boldsymbol{\theta})^{2p}]^{1/(2p)} = O_d(d^{-1/2+\delta}).$$

Denote $E_{\leq 4p} = E \cap \{0, \dots, 4p\}$ (recall that E contains all $k \geq \ell$) and decompose $\overline{\sigma}_{>u} = \overline{\mathsf{P}}_{E_{\leq 4p}}\overline{\sigma} + \overline{\mathsf{P}}_{>4p}\overline{\sigma}$. Then by triangle inequality we have,

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\overline{\sigma}_{>u}(\boldsymbol{x};\boldsymbol{\theta})^{2p}]^{1/(2p)} \leq \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\overline{\mathsf{P}}_{E_{\leq 4p}}\overline{\sigma}(\boldsymbol{x};\boldsymbol{\theta})^{2p}]^{1/(2p)} + \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\overline{\mathsf{P}}_{>4p}\overline{\sigma}(\boldsymbol{x};\boldsymbol{\theta})^{2p}]^{1/(2p)}.$$

Using hypercontractivity of polynomials of degree less or equal to $4p$, the first term is bounded by $O_d(d^{-1/2})$, while the second term is bounded in Proposition 29 in Section G.

(d) *(Concentration of diagonal elements.)* This is proved in Proposition 22 in Section F.

Step 4. Checking the spectral gap property at level $\{N(d), \mathsf{M}(d), n(d), \mathsf{m}(d)\}_{d\geq 1}$.
Let us now check the spectral gap property (Assumption 2 in Mei et al. (2021)).

(a) *(Number of samples.)* First by Eq. (26) and the assumption $d^{\mathsf{s}-\alpha+\delta_0} \leq n \leq d^{\mathsf{s}+1-\alpha-\delta_0}$, we have $\mathsf{m} \leq n^{1-\delta}$ for $\delta > 0$ chosen sufficiently small. By the choice of m and recalling Eq. (28), we have

$$\lambda_{\mathsf{m}+1}^{-2}\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \sup_{k\geq \mathsf{s}+1}\{\xi_{d,k}^{-2}\}\cdot \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \Omega_d(d^{\mathsf{s}+1-\alpha}) \geq n^{1+\delta}\,,$$
$$\lambda_{\mathsf{m}}^{-2}\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \xi_{d,\mathsf{s}}^{-2}\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = O_d(1)\cdot d^{\mathsf{s}-\alpha} \leq n^{1-\delta}\,,$$

with $\delta > 0$ chosen sufficiently small.

(b) *(Number of features.)* By construction $\mathsf{M} \geq \mathsf{m}$. Furthermore, recalling Eq. (26) and the assumption $d^{\mathsf{S}-\alpha+\delta_0} \leq N \leq d^{\mathsf{S}+1-\alpha-\delta_0}$, we have $\mathsf{M} \leq N^{1-\delta}$ for $\delta > 0$ chosen sufficiently small. By choice of M, $\lambda_{d,\mathsf{M}+1}^2 \leq d^{-\mathsf{S}-1+\delta}$. Hence,

$$\lambda_{\mathsf{M}+1}^{-2}\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) = \Omega_d(1)\cdot d^{\mathsf{S}+1-\alpha-\delta} \geq N^{1+\delta},$$

for $\delta > 0$ chosen sufficiently small.

Finally notice that we used a different parametrization of λ in Eq. (6) and the condition in Mei et al. (2021) becomes $\lambda/d^\alpha = O_d(1)\cdot \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$, i.e., $\lambda = O_d(1)$. This concludes the proof. ■

B.3. Proof of Theorem 5

We consider the same setting as in the previous section and consider

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{\theta}}[\overline{\sigma}(\boldsymbol{x}_1;\boldsymbol{\theta})\overline{\sigma}(\boldsymbol{x}_2;\boldsymbol{\theta})]\,.$$

Theorem 5 is a consequence of Theorem 4 in Mei et al. (2021) and the proof amounts to checking that H_d verifies the kernel concentration properties and eigenvalue condition (see Section 3.2 in Mei et al. (2021)). Note that some of the conditions were already covered in the proof of Theorem 2 and we will only mention the ones that still need to be verified. Furthermore, by the spectral gap property proven in Section B.2, the bound in Theorem 4 in Mei et al. (2021) (which is in term of a shrinkage operator) can indeed be rewritten as

$$R_{\mathrm{KR,inv}}(f_d, \boldsymbol{X}, \lambda) = \|\overline{\mathsf{P}}_{>\mathsf{s}} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

Proof [Proof of Theorem 5] We choose m as in the proof of Theorem 2.

Step 1. Checking the kernel concentration property at level $\{n(d), \mathsf{m}(d)\}_{d\ge 1}$.

First notice that

$$\mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x})] = \sum_{k=\mathsf{s}+1}^{\infty} \xi_{d,k}^4 D(\mathcal{A}_d; k) \Upsilon_k^{(d)}(\boldsymbol{x}_i, \boldsymbol{x}_i),$$

and the concentration of the diagonal elements in the case of a polynomial activation function follows from the same argument as in Section B.2.

Hence, we only need to check this property in the case non polynomial activation function σ ($\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$ and $\mathcal{G}_d = \mathrm{Cyc}_d$ of degeneracy 1). Let us choose u as in the proof of the feature map concentration property in Theorem 2.

- *(Properly decaying eigenvalues.)* We have

$$\begin{aligned}
&\mathrm{Tr}(\mathbb{H}_{d,>u}^2) \ge \xi_{d,\ell}^4 D(\mathbb{S}^{d-1}; \ell) = \Omega_d(1) \cdot d^{-\ell-1}\,,\\
&\mathrm{Tr}(\mathbb{H}_{d,>u}^4) \le \sup_{j\le u}\{\lambda_{d,j}^6\} \mathrm{Tr}(\mathbb{H}_{d,>u}) = O_d(1) \cdot d^{-3\ell}\,.
\end{aligned}$$

  Hence,

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>u}^2)^2}{\mathrm{Tr}(\mathbb{H}_{d,>u}^4)} = \Omega_d(1) \cdot d^{\ell-2} = \Omega_d(d^{2\mathsf{s}}) \ge n^{2+\delta}.$$

- *(Concentration of the diagonal elements of the kernel.)* This is proven in Proposition 23 in Section F.

Step 2. Checking the eigenvalue condition at level $\{n(d), \mathsf{m}(d)\}_{d\ge 1}$.

By the choice of m, we have

$$\lambda_{d,\mathsf{m}+1}^{-4} \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}^2) = \frac{\sum_{k\ge \mathsf{s}+1} \xi_{d,k}^4 D(\mathcal{A}_d; k)}{\sup_{k\ge \mathsf{s}+1} \xi_{d,k}^4} \ge D(\mathcal{A}_d; \mathsf{s}+1) = \Omega_d(d^{\mathsf{s}+1-\alpha}) \ge n^{1+\delta}.$$

Again notice that we used a different parametrization of λ in Eq. (11) and the condition in Mei et al. (2021) becomes $\lambda/d^\alpha = O_d(1) \cdot \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$, i.e., $\lambda = O_d(1)$. This concludes the proof. ■

Appendix C. Decomposition of invariant functions

In this section, we take $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathscr{Q}^d\}$, and $\mathcal{G}_d$ to be any group that is isomorphic to a subgroup of $\mathcal{O}(d)$ and that preserves $\mathcal{A}_d$. This section is mostly built on the technical background presented in Appendix H.

C.1. The invariant function class and the symmetrization operator

Let $L^2(\mathcal{A}_d)$ be the class of L^2 functions on $\mathcal{A}_d$ equipped with uniform probability measure $\mathrm{Unif}(\mathcal{A}_d)$. We define the invariant function class to be

$$L^2(\mathcal{A}_d, \mathcal{G}_d) = \Big\{ f \in L^2(\mathcal{A}_d) : f(\boldsymbol{x}) = f(g \cdot \boldsymbol{x}), \ \ \forall \boldsymbol{x} \in \mathcal{A}_d, \ \ \forall g \in \mathcal{G}_d \Big\}.$$

We define the symmetrization operator $\mathcal{S} : L^2(\mathcal{A}_d) \to L^2(\mathcal{A}_d, \mathcal{G}_d)$ to be

$$(\mathcal{S}f)(\boldsymbol{x}) = \int_{\mathcal{G}_d} f(g \cdot \boldsymbol{x}) \pi_d(\mathrm{d}g).$$

C.2. Orthogonal polynomials on invariant function class

For either $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathscr{Q}^d\}$, we define $V_{d,\le k} \subseteq L^2(\mathcal{A}_d)$ to be the subspace spanned by all the degree ℓ polynomials, $V_{d,>k} \equiv V_{d,\le k}^{\perp} \subseteq L^2(\mathcal{A}_d)$ to be the orthogonal complement of $V_{d,\le k}$, and $V_{d,k} = V_{d,\le k} \cap V_{d,\le k-1}^{\perp}$. In words, $V_{d,k}$ contains all degree k polynomials that orthogonal to all polynomials of degree at most $k-1$. We further define $V_{d,<k} = V_{d,\le k-1}$ and $V_{d,\ge k} = V_{d,>k-1}$.

Let $\overline{\mathsf{P}}_{\le \ell}$ to be the projection operator on $L^2(\mathcal{A}_d, \mathrm{Unif})$ that project a function onto $V_{d,\le \ell}$, the space spanned by all the degree ℓ polynomials. Then it is easy to see that $\overline{\mathsf{P}}_{\le \ell}$ and $\mathcal{S}$ operator commute. This means, for any $f \in L^2(\mathcal{A}_d)$, we have

$$\overline{\mathsf{P}}_{\le \ell}[\mathcal{S}(f)] = \mathcal{S}[\overline{\mathsf{P}}_{\le \ell}(f)].$$

Similarly, we can define $\overline{\mathsf{P}}_{\ell}$, $\overline{\mathsf{P}}_{<\ell}$, $\overline{\mathsf{P}}_{>\ell}$, $\overline{\mathsf{P}}_{\ge \ell}$, which commute with $\mathcal{S}$. We denote $V_{d,\ell}(\mathcal{G}_d) \equiv \mathcal{P}_\ell(\mathcal{A}_d, \mathcal{G}_d)$ to be the space of polynomials in the images of $\overline{\mathsf{P}}_\ell \mathcal{S}$ (which is consistent with the definition of $V_{d,\ell}(\mathcal{G}_d)$ in Definition 1). Then we have

$$\mathcal{P}_\ell(\mathcal{A}_d, \mathcal{G}_d) = \overline{\mathsf{P}}_\ell(L^2(\mathcal{A}_d, \mathcal{G}_d)) = \mathcal{S}[\overline{\mathsf{P}}_\ell(L^2(\mathcal{A}_d))].$$

We denote $D(\mathcal{A}_d; k) = D(\mathcal{A}_d; \mathcal{G}_d; k) \equiv \dim(\mathcal{P}_k(\mathcal{A}_d, \mathcal{G}_d))$ to be the dimension of $\mathcal{P}_k(\mathcal{A}_d, \mathcal{G}_d)$. We denote $\{\overline{Y}_{kl}^{(d)}\}_{l \in [D(\mathcal{A}_d;k)]}$ to be a set of orthonormal polynomial basis in $\mathcal{P}_k(\mathcal{A}_d, \mathcal{G}_d)$. That means

$$\mathbb{E}_{\boldsymbol{x} \sim \mathrm{Unif}(\mathcal{A}_d)}[\overline{Y}_{k_1 l_1}^{(d)}(\boldsymbol{x}) \overline{Y}_{k_2 l_2}^{(d)}(\boldsymbol{x})] = \mathbf{1}\{k_1 = k_2, l_1 = l_2\},$$

and

$$\overline{Y}_{kl}^{(d)}(\boldsymbol{x}) = \overline{Y}_{kl}^{(d)}(g \cdot \boldsymbol{x}), \ \ \forall \boldsymbol{x} \in \mathcal{A}_d, \ \ \forall g \in \mathcal{G}_d.$$

C.3. A representation lemma

We have the following representation lemma. This lemma is important in the proofs of counting the degeneracy of groups (See Section D).

Lemma 11 (Convolution representation of projection operator) *Let $Q_k^{(d)}$ be the k-th Gegenbauer polynomial, or the k-th hypercubic Gegenbauer polynomial. For any fixed integer k, we have*

$$\frac{1}{D(\mathcal{A}_d;k)}\sum_{l=1}^{D(\mathcal{A}_d;k)}\overline{Y}_{kl}^{(d)}(\boldsymbol{x})\overline{Y}_{kl}^{(d)}(\boldsymbol{y}) = \frac{B(\mathcal{A}_d;k)}{D(\mathcal{A}_d;k)}\int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{x}, g\cdot\boldsymbol{y}\rangle)\pi_d(\mathrm{d}g). \tag{29}$$

Proof [Proof of Lemma 11] Define

$$\Gamma_{1k}(\boldsymbol{x},\boldsymbol{y}) = \sum_{l=1}^{D(\mathcal{A}_d;k)}\overline{Y}_{kl}^{(d)}(\boldsymbol{x})\overline{Y}_{kl}^{(d)}(\boldsymbol{y}),$$

and

$$\Gamma_{2k}(\boldsymbol{x},\boldsymbol{y}) = B(\mathcal{A}_d;k)\int_{\mathcal{G}_d} Q_k^{(d)}(\langle g\cdot\boldsymbol{x},\boldsymbol{y}\rangle)\pi_d(\mathrm{d}g).$$

Then Γ_{1k} and Γ_{2k} define two operators $\mathbb{T}_{1k},\mathbb{T}_{2k}: L^2(\mathcal{A}_d)\to L^2(\mathcal{A}_d)$, i.e., for $j=1,2$,

$$\mathbb{T}_{jk}f(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim\mathrm{Unif}(\mathcal{A}_d)}[\Gamma_{jk}(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y})].$$

Recall that $Q_k^{(d)}$ is a representation of the projector onto the subspace of degree-k spherical harmonics (see Eq. (104) in Section H.1.2). We deduce that

$$\mathbb{T}_{2k}f(\boldsymbol{x}) = \mathcal{S}\mathbb{E}_{\boldsymbol{y}}[B(\mathcal{A}_d;k)Q_k^{(d)}(\langle\boldsymbol{x},\boldsymbol{y}\rangle)f(\boldsymbol{y})] = \mathcal{S}\overline{\mathsf{P}}_k f(\boldsymbol{x}),$$

and therefore $\mathbb{T}_{2k} = \mathcal{S}\overline{\mathsf{P}}_k$.

Furthermore, we have $\mathbb{T}_{1k} = \overline{\mathsf{P}}_k\mathcal{S}$. Indeed, the images of both $\mathbb{T}_{1k}$ and $\overline{\mathsf{P}}_k\mathcal{S}$ are $\mathcal{P}_k(\mathcal{A}_d,\mathcal{G}_d)$, the space $\mathcal{P}_k(\mathcal{A}_d,\mathcal{G}_d)^\perp$ is the null space of both $\mathbb{T}_{1k}$ and $\overline{\mathsf{P}}_k\mathcal{S}$, and $\mathbb{T}_{1k}\overline{Y}_{kp}^{(d)}(\boldsymbol{x}) = \overline{\mathsf{P}}_k\mathcal{S}\overline{Y}_{kp}^{(d)}(\boldsymbol{x}) = \overline{Y}_{kp}^{(d)}(\boldsymbol{x})$.

By the commutativity of $\overline{\mathsf{P}}_k$ and $\mathcal{S}$ operator, we have $\mathbb{T}_{1k} = \overline{\mathsf{P}}_k\mathcal{S} = \mathcal{S}\overline{\mathsf{P}}_k = \mathbb{T}_{2k}$, and hence $\Gamma_{1k} = \Gamma_{2k}$. ∎

C.4. Gegenbauer decomposition of invariant features and kernels

By Section H, for either $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathscr{Q}^d\}$, for any activation function $\sigma \in L^2([-\sqrt{d},\sqrt{d}],\tau_d^1)$ (where τ_d^1 is the distribution of $\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle$ when $\boldsymbol{x}_1,\boldsymbol{x}_2 \sim_{iid} \mathrm{Unif}(\mathcal{A}_d)$), we can define its coefficients $\xi_{d,k}(\sigma)$ defined by

$$\xi_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]}\sigma(x)Q_k^{(d)}(\sqrt{d}x)\tau_d^1(\mathrm{d}x), \tag{30}$$

so that we have the following equation holds in $L^2([-\sqrt{d}, \sqrt{d}], \tau_d^1)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \xi_{d,k}(\sigma) B(\mathcal{A}_d; k) Q_k^{(d)}(\sqrt{d}x).$$

For any group $\mathcal{G}_d$ that is a subgroup of $\mathcal{O}(d)$, we define

$$\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}) \equiv \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{x}, g \cdot \boldsymbol{\theta} \rangle / \sqrt{d}) \pi_d(\mathrm{d}g).$$

Then, by the representation lemma (Lemma 11), we have

$$\begin{aligned}
\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}) &\equiv \sum_{k=0}^{\infty} \xi_{d,k}(\sigma) B(\mathcal{A}_d; k) \int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{x}, g \cdot \boldsymbol{\theta} \rangle) \pi_d(\mathrm{d}g) \\
&= \sum_{k=0}^{\infty} \xi_{d,k}(\sigma) \sum_{l=1}^{D(\mathcal{A}_d; k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{x}) \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta}).
\end{aligned}$$

As a consequence, suppose we define

$$H_d(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\mathcal{A}_d)} [\overline{\sigma}(\boldsymbol{x}; \boldsymbol{\theta}) \overline{\sigma}(\boldsymbol{y}; \boldsymbol{\theta})].$$

Then we have

$$H_d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=0}^{\infty} \xi_{d,k}(\sigma)^2 \sum_{l=1}^{D(\mathcal{A}_d; k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{x}) \overline{Y}_{kl}^{(d)}(\boldsymbol{y}).$$

Appendix D. Counting the degeneracy

D.1. Counting the degeneracy of Cyc_d and $\mathrm{Cyc2D}_{d_1,d_2}$ (Example 1 and 2)

Proposition 12 *Let $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$ with $d = d_1 \times d_2$. Let $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathscr{Q}^d\}$. Then for any fixed $k \geq 1$, we have*

$$\dim(\mathcal{P}_k(\mathcal{A}_d, \mathcal{G}_d)) \equiv D(\mathcal{A}_d; k) = \Theta_d(d^{k-1}).$$

D.1.1. PROOF OF PROPOSITION 12

Here we state a key lemma that is used to prove Proposition 12.

Lemma 13 *Let $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$ with $d = d_1 \times d_2$. Denote*

$$F_k(\boldsymbol{z}) = \int_{\mathcal{G}_d} (\langle \boldsymbol{z}, g \cdot \boldsymbol{z} \rangle / d)^k \pi_d(\mathrm{d}g).$$

Then for any fixed $k \geq 1$, we have

$$\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[F_k(\boldsymbol{z})] = \Theta_d(d^{-1}), \tag{31}$$

$$\mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\mathscr{Q}^d)}[F_k(\boldsymbol{\theta})] = \Theta_d(d^{-1}), \tag{32}$$

$$\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{S}^{d-1}(\sqrt{d})}[F_k(\boldsymbol{\theta})] = \Theta_d(d^{-1}). \tag{33}$$

Proof [Proof of Lemma 13] We prove Eq. (31) and (33). The proof for Eq. (32) is similar to the proof of Eq. (31).

Let $\{L_\ell\}_{0 \leq \ell \leq d-1}$ be the matrix representation of group elements of Cyc_d or $\mathrm{Cyc2D}_{d_1,d_2}$: when $\mathcal{G}_d = \mathrm{Cyc}_d$, $g_\ell \in \mathrm{Cyc}_d$ gives matrix representation L_ℓ for $0 \leq \ell \leq d-1$; when $\mathcal{G}_d = \mathrm{Cyc2D}_{d_1,d_2}$, $g_{st} \in \mathrm{Cyc2D}_{d_1,d_2}$ gives matrix representation $L_{s \times d_2 + t}$ for $0 \leq s \leq d_1 - 1$, $0 \leq t \leq d_2 - 1$. As a consequence, for either $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$, $L_0 = \mathbf{I}_d$ is the identity matrix. This gives

$$F_k(\boldsymbol{z}) = \|\boldsymbol{z}\|_2^{2k}/d^{k+1} + \sum_{l=1}^{d-1} \langle \boldsymbol{z}, L_l \boldsymbol{z} \rangle^k / d^{k+1}.$$

Step 1. The case $k = 1$. For either $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$, we have $\mathbb{E}[\langle \boldsymbol{z}, L_l \boldsymbol{z} \rangle] = 0$ for $1 \leq l \leq d-1$. As a consequence, we have

$$\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[F_1(\boldsymbol{z})] = \mathbb{E}[\|\boldsymbol{z}\|_2^2/d^2] + \sum_{l=1}^{d-1} \mathbb{E}[\langle \boldsymbol{z}, L_l \boldsymbol{z} \rangle / d^2] = \frac{1}{d}. \tag{34}$$

Step 2. The case $k = 2$. Note we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_2(\boldsymbol{z})] &= \mathbb{E}[\|\boldsymbol{x}\|_2^4/d^3] + \sum_{l=1}^{d-1}\mathbb{E}[\langle \boldsymbol{x}, L_l\boldsymbol{x}\rangle^2/d^3] \\ &= \mathbb{E}\Big[\Big(\sum_{i=1}^d x_i^2\Big)^2\Big]/d^3 + \sum_{l=1}^{d-1}\mathbb{E}\Big[\Big(\sum_{i=1}^d x_i(L_l\boldsymbol{x})_i\Big)^2\Big]/d^3 \\ &= \sum_{i,j=1}^d \mathbb{E}[x_i^2x_j^2]/d^3 + \sum_{l=1}^{d-1}\sum_{i,j=1}^d \mathbb{E}[x_i(L_l\boldsymbol{x})_i x_j(L_l\boldsymbol{x})_j]/d^3 \\ &= \Big(\frac{1}{d} + \frac{2}{d^2}\Big) + \sum_{l=1}^{d-1}\sum_{i,j=1}^d \mathbb{E}[x_i(L_l\boldsymbol{x})_i x_j(L_l\boldsymbol{x})_j]/d^3.\end{aligned}$$

Note that for either $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$, for any $i \in [d]$ and $1 \le l \le d-1$, the random variable $(L_l\boldsymbol{x})_i$ is independent from x_i. This gives

$$0 \le \sum_{l=1}^{d-1}\sum_{i,j=1}^d \mathbb{E}[x_i(L_l\boldsymbol{x})_i x_j(L_l\boldsymbol{x})_j]/d^3 \le \frac{2(d-1)d}{d^3} = \Theta_d(d^{-1}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_2(\boldsymbol{z})] = \Theta_d(d^{-1}). \tag{35}$$

Step 3. The case $k \ge 3$. By the moment formula of the χ^2 distribution, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[(\|\boldsymbol{z}\|_2^2/d)^k] = 1 + o_d(1).$$

Moreover, for either $\mathcal{G}_d \in \{\mathrm{Cyc}_d, \mathrm{Cyc2D}_{d_1,d_2}\}$, for any $l \ne 0$, we have

$$\mathbb{E}[\langle \boldsymbol{z}, L_l\boldsymbol{z}\rangle]/d = 0.$$

As a consequence, by the Hanson-Wright inequality as in Lemma 14, for any fixed $k \ge 3$ and $\varepsilon > 0$, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\Big[\sup_{1\le l\le d-1}(\langle \boldsymbol{z}, L_l\boldsymbol{z}\rangle/d)^k\Big] = O_d(d^{-k/2+\varepsilon}).$$

Therefore, for $k \ge 3$, we have

$$\Big|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] - \frac{1}{d}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[(\|\boldsymbol{z}\|_2^2/d)^k]\Big| \le \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\Big[\sup_{1\le l\le d-1}(\langle \boldsymbol{z}, L_l\boldsymbol{z}\rangle/d)^k\Big] = o_d(1/d),$$

so that

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] = 1/d + o_d(1/d). \tag{36}$$

Combining Eq. (34), (35), and (36) proves Eq. (31).

Step 4. From Gaussian to spherical. Note that when $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, we have $\|\boldsymbol{z}\|_2^2 \sim \chi^2(d)$ which is independent of $\sqrt{d}\cdot\boldsymbol{z}/\|\boldsymbol{z}\|_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Hence, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] &= \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{S}^{d-1}(\sqrt{d}),\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{\theta})(\|\boldsymbol{z}\|_2^{2k}/d^k)] \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{S}^{d-1}(\sqrt{d})}[F_k(\boldsymbol{\theta})]\cdot\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\|\boldsymbol{z}\|_2^{2k}/d^k].\end{aligned}$$

Note that for fixed $k \geq 1$, the moment formula for χ^2 distribution gives

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\|\boldsymbol{z}\|_2^{2k}/d^k] = 1 + o_d(1).$$

Combining with Eq. (31), we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{S}^{d-1}(\sqrt{d})}[F_k(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})]/\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\|\boldsymbol{z}\|_2^{2k}/d^k] = \Theta_d(d^{-1}).$$

This proves Eq. (33). ■

Proof [Proof of Proposition 12] Denote

$$P_k(\boldsymbol{\theta}) \equiv \frac{1}{B(\mathcal{A}_d;k)} \sum_{l=1}^{D(\mathcal{A}_d;k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta})^2 = \int_{\mathcal{G}_d} Q_k^{(d)}(\langle\boldsymbol{\theta}, g\cdot\boldsymbol{\theta}\rangle)\pi_d(\mathrm{d}g).$$

By Lemma 11, for any fixed $k \geq 1$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{A}_d)}[P_k(\boldsymbol{\theta})] = \frac{D(\mathcal{A}_d;k)}{B(\mathcal{A}_d;k)}. \tag{37}$$

By Lemma 15, we have

$$P_k(\boldsymbol{\theta}) = \sum_{m=0}^{k} a_{d,k,m} F_m(\boldsymbol{\theta}),$$

where $|a_{d,k,m}| \leq C_{k,m}/d^{(k-m)/2}$. As a result, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{A}_d)}[P_k(\boldsymbol{\theta})] = \sum_{m=0}^{k} a_{d,k,m}\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{A}_d)}[F_k(\boldsymbol{\theta})] = \Theta(d^{-1}).$$

Combining with Eq. (37) shows that $D(\mathcal{A}_d;k) = \Theta(d^{-1}B(\mathcal{A}_d;k)) = \Theta(d^{k-1})$. This concludes the proof. ■

D.1.2. AUXILIARY LEMMAS

Lemma 14 (Hanson-Wright inequality) *There exists a universal constant $c > 0$, such that for any $t > 0$ and $d \in \mathbb{N}$, and any permutation matrix $L \in \mathbb{R}^{d\times d}$ be any permutation matrix, when $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ or $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$, we have*

$$\mathbb{P}\Big(\Big|\langle\boldsymbol{x}, L\cdot\boldsymbol{x}\rangle - \mathbb{E}[\langle\boldsymbol{x}, L\cdot\boldsymbol{x}\rangle]\Big|/d \geq t\Big) \leq 2\cdot\exp\{-cd\cdot\min(t^2,t)\}.$$

Proof [Proof of Lemma 14] Note that for any permutation matrix L, we have $\|L\|_F \leq \sqrt{d}$, and $\|L\|_{\mathrm{op}} \leq 1$. By the Hanson-Wright inequality of vectors with independent sub-Gaussian entries (for example, see Theorem 1.1 of Rudelson et al. (2013)), we have

$$\mathbb{P}\Big(\Big|\langle\boldsymbol{x}, L\boldsymbol{x}\rangle - \mathbb{E}[\langle\boldsymbol{x}, L\boldsymbol{x}\rangle]\Big|/d > t\Big) \leq 2\exp\{-cd\cdot\min(t^2,t)\}.$$

This concludes the proof ■

Lemma 15 *Let $Q_k^{(d)}$ be either the k'th Gegenbauer polynomial or the k'th hypercubic Gegenbauer polynomial (as defined in Section H). Let coefficients of monomials in $Q_k^{(d)}(d \cdot x)$ to be $\{a_{d,k,m}\}_{0\le m\le k}$. That is, we have*

$$Q_k^{(d)}(x) = \sum_{m=0}^{k} a_{d,k,m}(x/d)^m.$$

Then, for any fixed k, there exists constant $C(k)$, such that

$$|a_{d,k,m}| \le C(k)/d^{(k-m)/2}.$$

Moreover, we have

$$\lim_{d\to\infty} a_{d,k,k} = 1.$$

Finally, for k and m in different parity, we have

$$a_{d,k,m} = 0.$$

Proof [Proof of Lemma 15] The proof holds by the following equation

$$\lim_{d\to\infty} \mathrm{Coeff}\Big\{B(\mathcal{A}_d;k)^{1/2}Q_k^{(d)}(\sqrt{d}\cdot x)\Big\} = \mathrm{Coeff}\Big\{\frac{1}{\sqrt{k!}}\mathrm{He}_k(x)\Big\}.$$

when $Q_k^{(d)}$ is either Gegenbauer polynomial or Hypercubic Gegenbauer polynomial (See Eq. (110) and Eq. (112)). ■

D.2. Counting the degeneracy of band-limited function class (Example 3)

Proposition 16 *Follow the notations of Example 3. Then for any fixed $k \ge 1$, we have*

$$D(\mathbb{S}^{(d-1)};k) = \Theta_d(d^{k-1}).$$

Here we state Lemma 17 that is used to prove Proposition 16. Given Lemma 17, the proof of Proposition 16 is the same as the proof of Proposition 12.

Lemma 17 *Follow the notations of Example 3. Denote*

$$F_k(\boldsymbol{z}) = \int_{\mathrm{Sft}_d} (\langle \boldsymbol{z}, g\cdot \boldsymbol{z}\rangle/d)^k \pi_d(\mathrm{d}g).$$

Then for any fixed $k \ge 1$, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] = \Theta_d(d^{-1}).$$

Proof [Proof of Lemma 17]

We prove the lemma for the case when d is odd. We denote $u_1 = z_1^2$, and $u_i = z_{2i}^2 + z_{2i+1}^2$ for $i = 2, \dots, (d-1)/2$. Then we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] = d^{-k}\cdot\mathbb{E}_{\boldsymbol{z}}\Big\{\int_{[0,1]}\Big(\sum_{j=0}^{(d-1)/2} u_j\cos(2\pi jt)\Big)^k \mathrm{d}t\Big\}\\
&= d^{-k}\cdot\mathbb{E}_{\boldsymbol{z}}\Big\{\int_{[0,1]}\sum_{j_1,\dots,j_k=0}^{(d-1)/2}\Big(\prod_{s\in[k]} u_{j_s}\cos(2\pi j_s t)\Big)\mathrm{d}t\Big\} \qquad (38)\\
&= d^{-k}\cdot\sum_{j_1,\dots,j_k=0}^{(d-1)/2}\mathbb{E}_{\boldsymbol{z}}\Big\{\prod_{s\in[k]} u_{j_s}\Big\}\Big(\int_{[0,1]}\prod_{s\in[k]}\cos(2\pi j_s t)\mathrm{d}t\Big).
\end{aligned}$$

Step 1. Bound Z function. First, we denote

$$Z(j_1,\dots,j_k) = \mathbb{E}_{\boldsymbol{z}}\Big\{\prod_{s\in[k]} u_{j_s}\Big\}.$$

We have

$$\begin{aligned}
&\sup_{j_1,\dots,j_k} Z(j_1,\dots,j_k) \le \sup_{j_1,\dots,j_k}\prod_{s\in[k]}\mathbb{E}[u_{j_s}^{2k}]^{1/(2k)}\\
&\le \sup_{j\in\{0,1,\dots,(d-1)/2\}}\mathbb{E}[u_j^{2k}]^{1/2} \le 2^k\cdot\mathbb{E}_{G\sim\mathcal{N}(0,1)}[G^{2k}]^{1/2} \equiv M_k. \qquad (39)
\end{aligned}$$

Moreover, we have

$$\inf_{j_1,\dots,j_k} Z(j_1,\dots,j_k) \ge \mathbb{E}_{G\sim\mathcal{N}(0,1)}[G^2] = 1. \qquad (40)$$

Step 2. Bound $|\mathcal{I}|$. Further, we denote

$$\mathcal{I} = \Big\{(j_1,\dots,j_k)\in\{0,\dots,(d-1)/2\}^k : \exists(\varepsilon_i)_{i\in[k]}\in\{\pm1\}^k, \sum_{i=1}^k \varepsilon_i j_i = 0\Big\},$$

Then it is easy to see that

$$[(d+1)/2]^{k-1} \le |\mathcal{I}| \le 2\cdot(d+1)^{k-1}. \qquad (41)$$

Step 3. Bound E function. Next, we denote

$$E(j_1,\dots,j_k) = \int_{[0,1]}\prod_{s\in[k]}\cos(2\pi j_s t)\mathrm{d}t.$$

It is easy to see that

$$\sup_{j_1,\dots,j_k}|E(j_1,\dots,j_k)| \le 1. \qquad (42)$$

Moreover, for any $(j_1,\dots,j_k)\notin\mathcal{I}$, we have $E(j_1,\dots,j_k) = 0$. For any $(j_1,\dots,j_k)\in\mathcal{I}$, we have

$$E(j_1,\dots,j_k) = \frac{1}{2^k}\int_{[0,1]}\prod_{s\in[k]}[\exp(\mathrm{i}2\pi j_s t) + \exp(-\mathrm{i}2\pi j_s t)]\mathrm{d}t \ge 1/2^k. \qquad (43)$$

The last inequality used the fact that $(j_1, \ldots, j_k) \in \mathcal{I}$.

Step 4. Concludes the proof. Therefore, combining Eq. (38) (39) (41) (42), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] &\leq d^{-k}\cdot M_k\cdot \sum_{j_1,\ldots,j_k=0}^{(d-1)/2} |E(j_1,\ldots,j_k)| \\ &\leq d^{-k}\cdot M_k\cdot |\mathcal{I}| = O_d(d^{-1}).\end{aligned}$$

Combining Eq. (38) (40) (41) (43), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{z})] &\geq d^{-k}\cdot \sum_{j_1,\ldots,j_k=0}^{(d-1)/2} |E(j_1,\ldots,j_k)| \\ &\geq d^{-k}\cdot |\mathcal{I}|/2^k = \Omega_d(d^{-1}).\end{aligned}$$

This concludes the proof. ■

Appendix E. Concentration for invariant groups with degeneracy $\alpha \le 1$

Let $Q_k^{(d)}$ be the k'th Gegenbauer polynomial on $\mathcal{A}_d \in \{\mathbb{S}^{d-1}(\sqrt{d}), \mathscr{Q}^d\}$ (see Section H for definitions). Let $\mathcal{G}_d$ be an invariant group with degeneracy α. That means, for any fixed $k \ge \alpha$, we have $B(\mathcal{A}_d; k)/[D(\mathcal{A}_d; k)d^\alpha] = \Theta_d(1)$. For $k \in \mathbb{N}_{\ge 0}$, we denote

$$\Upsilon_k(\boldsymbol{\theta}) = \frac{1}{D(\mathcal{A}_d; k)} \sum_{l=1}^{D(\mathcal{A}_d;k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta})\overline{Y}_{kl}^{(d)}(\boldsymbol{\theta}) = \frac{B(\mathcal{A}_d; k)}{D(\mathcal{A}_d; k)} \int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta}\rangle)\pi_d(\mathrm{d}g). \tag{44}$$

Then we have

$$\mathbb{E}[\Upsilon_k(\boldsymbol{\theta})] = 1.$$

In this section, we show that Υ_k concentration around its mean, for any fixed $k \ge 2$ and $\alpha \le 1$.

E.1. Main proposition

Proposition 18 *Let $\mathcal{G}_d$ be an invariant group with degeneracy $\alpha \le 1$. Let $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathcal{A}_d)$ where $N = O_d(d^p)$ for some fixed integer p. Let Υ_k be as defined in Eq. (44). Then for any fixed $k \ge 2$, we have*

$$\sup_{i\in[N]} \Big|\Upsilon_k(\boldsymbol{\theta}_i) - 1\Big| = o_{d,\mathbb{P}}(1).$$

Proof [Proof of Proposition 18] Let us first focus on the sphere case $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$. Let $(\boldsymbol{x}_i)_{i\in[N]} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Without loss of generality, we assume $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_i$ are coupled such that $\boldsymbol{\theta}_i = \sqrt{d} \cdot \boldsymbol{x}_i/\|\boldsymbol{x}_i\|_2$. Denote

$$F_k(\boldsymbol{z}) = \int_{\mathcal{G}_d} (\langle \boldsymbol{z}, g \cdot \boldsymbol{z}\rangle/d)^k \pi_d(\mathrm{d}g).$$

Let $\{a_{d,k,m}\}_{0\le m\le k}$ be the coefficients of monomials in $Q_k^{(d)}(d \cdot x)$. That is, we have

$$Q_k^{(d)}(x) = \sum_{m=0}^{k} a_{d,k,m}(x/d)^m.$$

Then

$$\int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta}\rangle)\pi_d(\mathrm{d}g) = \sum_{m=0}^{k} a_{d,k,m}F_m(\boldsymbol{\theta}).$$

Moreover, by Lemma 15, we have $|a_{d,k,m}| \le C_{k,m}/d^{(k-m)/2}$, $\lim_{d\to\infty} a_{d,k,k} = 1$, and $a_{d,k,m} = 0$ for k and m of different parity.

Then we have

$$\begin{aligned}
&\sup_{i\in[N]} |\Upsilon_k(\boldsymbol{\theta}_i) - \mathbb{E}[\Upsilon_k(\boldsymbol{\theta}_i)]| \\
&= \frac{B(\mathbb{S}^{d-1};k)}{D(\mathbb{S}^{d-1};k)} \sup_{i\in[N]} \Big| \int_{\mathcal{G}_d} Q_k(\langle \boldsymbol{\theta}_i, g\cdot\boldsymbol{\theta}_i\rangle)\pi_d(\mathrm{d}g) - \mathbb{E}\Big[\int_{\mathcal{G}_d} Q_k(\langle \boldsymbol{\theta}_i, g\cdot\boldsymbol{\theta}_i\rangle)\pi_d(\mathrm{d}g)\Big]\Big| \\
&\le C\times d^{\alpha}\times \sum_{m=1}^{k} a_{d,k,m} \times \sup_{i\in[N]} \Big|F_m(\boldsymbol{\theta}_i) - \mathbb{E}[F_m(\boldsymbol{\theta}_i)]\Big| \\
&\le C\times d^{\alpha}\times \sum_{m=1}^{k} a_{d,k,m} \times \sup_{i\in[N]} \Big|F_m(\boldsymbol{x}_i) - \mathbb{E}[F_m(\boldsymbol{x}_i)]\Big| \cdot [d^m/\|\boldsymbol{x}_i\|_2^{2m}].
\end{aligned}$$

By the concentration of χ^2-distribution, for any $\varepsilon > 0$, the following event happens with high probability

$$\mathcal{E}_1 \equiv \Big\{ \sup_{i\in[N]} \big|\|\boldsymbol{x}_i\|_2^2/d - 1\big| \le 1/d^{1/2-\varepsilon}\Big\}.$$

Moreover, combining Lemma 19 with Lemma 20, for any fixed $m \ge 2$, we have

$$\mathbb{E}[(F_m(\boldsymbol{x}) - \mathbb{E}[F_m(\boldsymbol{x})])^2] \le C_m d^{-1-3\alpha/2}.$$

By the hypercontractivity property of Gaussian distribution as per Lemma 37, for any $\varepsilon > 0$, taking q sufficiently large, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{i\in[N]} \Big|F_m(\boldsymbol{x}_i) - \mathbb{E}[F_m(\boldsymbol{x}_i)]\Big|\Big] \le \mathbb{E}\Big[\sum_{i=1}^{N}\Big(F_m(\boldsymbol{x}_i) - \mathbb{E}[F_m(\boldsymbol{x}_i)]\Big)^{2q}\Big]^{1/(2q)} \\
&\le C(q)\cdot d^{p/(2q)}\cdot \mathbb{E}[(F_m(\boldsymbol{x}) - \mathbb{E}[F_m(\boldsymbol{x})])^2]^{1/2} \le C d^{-1-3\alpha/2+\varepsilon}
\end{aligned}$$

By Markov's inequality, we deduce that the following event happens with high probability

$$\mathcal{E}_2 \equiv \Big\{\forall 2\le m\le k, \sup_{i\in[N]} \Big|F_m(\boldsymbol{x}_i) - \mathbb{E}[F_m(\boldsymbol{x}_i)]\Big| \le C d^{-1-3\alpha/2+\varepsilon}\Big\}.$$

Finally, by Lemma 19, we have

$$\mathbb{E}[F_1(\boldsymbol{x})^2] \le C d^{-2\alpha},$$

and by the hypercontractivity property of low degree polynomials with Gaussian measure (Lemma 37), for any $\varepsilon > 0$, taking q sufficiently large, we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{i\in[N]} \Big|F_1(\boldsymbol{x}_i) - \mathbb{E}[F_1(\boldsymbol{x}_i)]\Big|\Big] &\le c\mathbb{E}\Big[\sum_{i=1}^{N} F_1(\boldsymbol{x}_i)^{2q}\Big]^{1/(2q)} + \mathbb{E}[F_1(\boldsymbol{x})^2]^{1/2} \\
&\le C(q)\cdot d^{p/(2q)}\cdot \mathbb{E}[F_1(\boldsymbol{x})^2]^{1/2} + \mathbb{E}[F_1(\boldsymbol{x})^2]^{1/2} \le C d^{-\alpha+\varepsilon}.
\end{aligned}$$

As a result, the following event happens with high probability

$$\mathcal{E}_3 \equiv \Big\{ \sup_{i\in[N]} \Big|F_1(\boldsymbol{x}_i) - \mathbb{E}[F_1(\boldsymbol{x}_i)]\Big| \le C d^{-\alpha+\varepsilon}\Big\}.$$

When all the events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ happen, for any $k \geq 2$, we have

$$\begin{aligned}
&\sup_{i\in[N]} |\Upsilon_k(\boldsymbol{\theta}_i) - \mathbb{E}[\Upsilon_k(\boldsymbol{\theta}_i)]| \\
&\leq C \times d^{\alpha} \times \sum_{m=1}^{k} a_{d,k,m} \times \sup_{i\in[N]} \left| F_m(\boldsymbol{x}_i) - \mathbb{E}[F_m(\boldsymbol{x}_i)] \right| \cdot [d^m/\|\boldsymbol{x}_i\|_2^{2m}] \\
&\leq C \times d^{\alpha} \times \left[d^{-(k-1)/2} d^{-\alpha+\varepsilon} + \sum_{m=2}^{k} d^{-(k-m)/2} \times d^{-1-3\alpha/2+\varepsilon} \right] = o_d(1).
\end{aligned}$$

The case of the hypercube $\mathcal{A}_d \sim \mathscr{Q}^d$ follows similarly without introducing the gaussian measure and using Lemma 21 instead of Lemma 20. ■

E.2. Auxiliary Lemmas

Lemma 19 *Let $\mathcal{G}_d$ be an invariant group with degeneracy $\alpha \leq 1$. Denote*

$$F_k(\boldsymbol{z}) = \int_{\mathcal{G}_d} (\langle \boldsymbol{z}, g \cdot \boldsymbol{z} \rangle / d)^k \pi_d(\mathrm{d}g).$$

Then for any fixed $s \in [1, \infty)$ and integer $k \geq 1$, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{x})^s]^{1/s} = O_d(d^{-\alpha}), \tag{45}$$

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{A}_d)}[F_k(\boldsymbol{\theta})^s]^{1/s} = O_d(d^{-\alpha}). \tag{46}$$

Proof [Proof of Lemma 19]

For $\boldsymbol{\theta} \in \mathcal{A}_d$, denote

$$P_k(\boldsymbol{\theta}) \equiv \frac{1}{B(\mathcal{A}_d; k)} \sum_{l=1}^{D(\mathcal{A}_d;k)} \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta})^2 = \int_{\mathcal{G}_d} Q_k^{(d)}(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle) \pi_d(\mathrm{d}g).$$

By Lemma 11 and by the assumption that $\mathcal{G}_d$ is an invariant group with degeneracy $\alpha \leq 1$, i.e., $B(\mathcal{A}_k; k)/[D(\mathcal{A}_k; k) d^{\alpha}] = \Theta_d(1)$, for any fixed $k \geq 1$, we have

$$\mathbb{E}[P_k(\boldsymbol{\theta})] = \frac{D(\mathcal{A}_d; k)}{B(\mathcal{A}_d; k)} = O_d(d^{-\alpha}).$$

Throughout the proof, we will denote $L^s = L^s(\mathcal{A}_d)$ to be the L^s space with respect to distribution $\boldsymbol{\theta} \sim \mathrm{Unif}(\mathcal{A}_d)$.

By the hypercontractivity of low degree polynomials on the sphere and the hypercube, as per Lemmas 35 and 36, for any $s \geq 1$, we have

$$\begin{aligned}
\|P_k\|_{L^s} &= \frac{D(\mathcal{A}_d;k)}{B(\mathcal{A}_d;k)}\Big\|\frac{B(\mathcal{A}_d;k)}{D(\mathcal{A}_d;k)}P_k\Big\|_{L^s} \leq \frac{C_{k,s}}{d^\alpha}\Big\|\frac{B(\mathcal{A}_d;k)}{D(\mathcal{A}_d;k)}P_k\Big\|_{L^2} \\
&= \frac{C_{k,s}}{d^\alpha}\Big[\frac{1}{D(\mathcal{A}_d;k)^2}\sum_{l_1,l_2=1}^{D(\mathcal{A}_d;k)}\mathbb{E}\big[\overline{Y}_{kl_1}^{(d)}(\boldsymbol{\theta})^2\overline{Y}_{kl_2}^{(d)}(\boldsymbol{\theta})^2\big]\Big]^{1/2} \\
&\leq \frac{C_{k,s}}{d^\alpha}\Big[\frac{1}{D(\mathcal{A}_d;k)^2}\sum_{l_1,l_2=1}^{D(\mathcal{A}_d;k)}\mathbb{E}\big[\overline{Y}_{kl_1}^{(d)}(\boldsymbol{\theta})^4\big]^{1/2}\mathbb{E}\big[\overline{Y}_{kl_2}^{(d)}(\boldsymbol{\theta})^4\big]^{1/2}\Big]^{1/2} \\
&\leq \frac{C_{k,s}}{d^\alpha}\Big[\frac{1}{D(\mathcal{A}_d;k)^2}\sum_{l_1,l_2=1}^{D(\mathcal{A}_d;k)}\mathbb{E}\big[\overline{Y}_{kl_1}^{(d)}(\boldsymbol{\theta})^2\big]\mathbb{E}\big[\overline{Y}_{kl_2}^{(d)}(\boldsymbol{\theta})^2\big]\Big]^{1/2} = \frac{C_{k,s}}{d^\alpha}.
\end{aligned} \tag{47}$$

Let $\{a_{d,k,m}\}_{0\leq m\leq k}$ be the coefficients of monomials in $Q_k^{(d)}(d\cdot x)$. That is, we have

$$Q_k^{(d)}(x) = \sum_{m=0}^{k} a_{d,k,m}(x/d)^m.$$

Then

$$P_k = \sum_{m=0}^{k} a_{d,k,m}F_m. \tag{48}$$

Moreover, by Lemma 15, we have $|a_{d,k,m}| \leq C_{k,m}/d^{(k-m)/2}$, $\lim_{d\to\infty} a_{d,k,k} = 1$, and $a_{d,k,m} = 0$ for k and m have different parity.

We conclude the proof by induction over k. Note we have $F_0(\boldsymbol{\theta}) \equiv 1$. Moreover, for any $s \geq 1$, by Eq. (47) and (48) (and note that $a_{d,1,0} = 0$ and $\lim_{d\to\infty} a_{d,1,1} \to 1$), we have

$$\|F_1\|_{L^s} = \frac{1}{a_{d,1,1}}\|P_1\|_{L^s} \leq C_s/d^\alpha.$$

Fix a $k \geq 2$. Assume that, for any $1 \leq u \leq k-1$, we have $\|F_u\|_{L^s} \leq C_{u,s}/d^\alpha$ for $s \geq 1$, by Eq. (48) and (47), and the fact that $|a_{d,k,m}| \leq C_{k,m}/d^{(k-m)/2}$ and $\lim_{d\to\infty} a_{d,k,k} = 1$, we have

$$\begin{aligned}
\|F_k\|_{L^s} &= \Big\|\frac{1}{a_{d,k,k}}P_k - \sum_{m=1}^{k-1} a_{d,k,m}F_m - a_{d,k,0}\Big\|_{L^s} \\
&\leq \Big\|\frac{1}{a_{d,k,k}}P_k\Big\|_{L^s} + \sum_{m=1}^{k-1}|a_{d,k,m}|\Big\|F_m\Big\|_{L^s} + |a_{d,k,0}| \\
&\leq C/d^\alpha + \Big[\sum_{m=1}^{k-1} C/d^{(k-m)/2}\Big]\cdot C/d^\alpha + C/d^{k/2} \leq C_{k,s}/d^\alpha,
\end{aligned}$$

where we recall that we assume $\alpha \leq 1$.

Finally, for the case of $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$, recalling that we can write

$$F_k(\boldsymbol{x}) = (\|\boldsymbol{x}\|_2^2/d)^k F_k(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = \boldsymbol{x}/\|\boldsymbol{x}_i\|_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})$ is independent of $\|\boldsymbol{x}_i\|_2$ in the case of $\boldsymbol{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. Hence, we get by Cauchy-Schwarz inequality

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{x})^s]^{1/s} = \|F_k\|_{L^{2s}} \cdot \mathbb{E}_{\boldsymbol{x}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}\Big[(\|\boldsymbol{x}\|_2^2/d)^{2sk}\Big]^{1/(2s)} \le C_{k,s}/d^{\alpha},$$

by hypercontractivity of low degree polynomials for Gaussian measure (Lemma 37). ■

Lemma 20 *Let $\boldsymbol{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. Let $\mathcal{G}_d$ be a general invariant group. Let $F_k(\boldsymbol{z})$ be defined as in Lemma 19. Then for any fixed $k \ge 1$, there exists a constant C_k, such that*

- *If k is odd, then*
$$\mathrm{Var}_{\boldsymbol{x}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{x})] \le \frac{C_k}{d}\mathbb{E}_{\boldsymbol{x}}[F_{k-1}(\boldsymbol{x})^2].$$
- *If k is even, then*
$$\mathrm{Var}_{\boldsymbol{x}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}[F_k(\boldsymbol{x})] \le \frac{C_k}{d}\Big(\mathbb{E}_{\boldsymbol{x}}[F_{k-2}(\boldsymbol{x})^2] \wedge \mathbb{E}_{\boldsymbol{x}}[F_k(\boldsymbol{x})^2]^{(2k-1)/(2k)}\Big).$$

Proof [Proof of Lemma 20]

By the Gaussian Poincaré inequality, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}[(F_k(\boldsymbol{x}) - \mathbb{E}[F_k(\boldsymbol{x})])^2] \le \mathbb{E}[\|\nabla F_k(\boldsymbol{x})\|_2^2].$$

We have

$$\nabla F_k(\boldsymbol{x}) = k\int_{\mathcal{G}_d} (\langle \boldsymbol{x}, g\cdot\boldsymbol{x}\rangle/d)^{k-1}[(g\cdot\boldsymbol{x} + g^{-1}\cdot\boldsymbol{x})/d]\pi_d(\mathrm{d}g),$$

which gives

$$\begin{aligned}&\mathbb{E}[\|\nabla F_k(\boldsymbol{x})\|_2^2]\\ &= \frac{4k^2}{d}\int_{\mathcal{G}_d\times\mathcal{G}_d}\mathbb{E}\Big[(\langle \boldsymbol{x}, g_1\cdot\boldsymbol{x}\rangle/d)^{k-1}(\langle \boldsymbol{x}, g_2\cdot\boldsymbol{x}\rangle/d)^{k-1}\langle \boldsymbol{x}, g_1g_2\cdot\boldsymbol{x}\rangle/d\Big]\pi_d(\mathrm{d}g_1)\pi_d(\mathrm{d}g_2).\end{aligned} \tag{49}$$

Case 1: Odd k. When k is odd, we have

$$\begin{aligned}&\mathbb{E}[\|\nabla F_k(\boldsymbol{x})\|_2^2]\\ &\le \frac{4k^2}{d}\int_{\mathcal{G}_d\times\mathcal{G}_d}\mathbb{E}\Big[(\langle \boldsymbol{x}, g_1\cdot\boldsymbol{x}\rangle/d)^{k-1}(\langle \boldsymbol{x}, g_2\cdot\boldsymbol{x}\rangle/d)^{k-1}\|\boldsymbol{x}\|_2^2/d\Big]\pi_d(\mathrm{d}g_1)\pi_d(\mathrm{d}g_2)\\ &= \frac{4k^2}{d}\mathbb{E}\big[F_{k-1}(\boldsymbol{x})^2(\|\boldsymbol{x}\|_2^2/d)\big] \le \frac{4k^2}{d}\mathbb{E}\big[F_{k-1}(\boldsymbol{x})^4\big]^{1/2}\mathbb{E}\big[(\|\boldsymbol{x}\|_2^2/d)^2\big]^{1/2} \le \frac{C_k}{d}\mathbb{E}[F_{k-1}(\boldsymbol{x})^2],\end{aligned}$$

where we used in the second line Cauchy-Schwarz inequality and that the matrix representations of g are orthogonal matrices, and in the last inequality the hypercontractivity of low degree polynomials for Gaussian measures (Lemma 37).

Case 2: Even k. Bound 1. When k is even, we have the following first bound

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_k(\boldsymbol{x})\|_2^2] \\
&\le \frac{4k^2}{d}\int_{\mathcal{G}_d\times\mathcal{G}_d}\mathbb{E}\Big[(\langle \boldsymbol{x}, g_1\cdot\boldsymbol{x}\rangle/d)^{k-2}(\langle \boldsymbol{x}, g_2\cdot\boldsymbol{x}\rangle/d)^{k-2}\|\boldsymbol{x}\|_2^6/d^3\Big]\pi_d(\mathrm{d}g_1)\pi_d(\mathrm{d}g_2) \\
&= \frac{4k^2}{d}\mathbb{E}\big[F_{k-2}(\boldsymbol{x})^2(\|\boldsymbol{x}\|_2^6/d^3)\big] \le \frac{4k^2}{d}\mathbb{E}\big[F_{k-2}(\boldsymbol{x})^4\big]^{1/2}\mathbb{E}\big[(\|\boldsymbol{x}\|_2^6/d^3)^2\big]^{1/2} \le \frac{C_k}{d}\mathbb{E}\big[F_{k-2}(\boldsymbol{x})^2\big]\,.
\end{aligned}$$

Case 3: Even k. Bound 2. When k is even, we have the following second bound, which follows by Hölder's inequality $\mathbb{E}[XY]\le\mathbb{E}[|X|^{k/(k-1)}]^{(k-1)/k}\cdot\mathbb{E}[|Y|^k]^{1/k}$,

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_k(\boldsymbol{x})\|_2^2] \\
&= \frac{4k^2}{d}\int_{\mathcal{G}_d\times\mathcal{G}_d}\mathbb{E}\Big[(\langle \boldsymbol{x}, g_1\cdot\boldsymbol{x}\rangle/d)^{k-1}(\langle \boldsymbol{x}, g_2\cdot\boldsymbol{x}\rangle/d)^{k-1}\langle \boldsymbol{x}, g_1g_2\cdot\boldsymbol{x}\rangle/d\Big]\pi_d(\mathrm{d}g_1)\pi_d(\mathrm{d}g_2) \\
&\le \frac{4k^2}{d}\mathbb{E}\Big[\int_{\mathcal{G}_d\times\mathcal{G}_d}(\langle \boldsymbol{x}, g_1\cdot\boldsymbol{x}\rangle/d)^k(\langle \boldsymbol{x}, g_2\cdot\boldsymbol{x}\rangle/d)^k\pi_d(\mathrm{d}g_1)\pi_d(\mathrm{d}g_2)\Big]^{(k-1)/k} \\
&\quad\times\mathbb{E}\Big[\int_{\mathcal{G}_d}(\langle \boldsymbol{x}, g\cdot\boldsymbol{x}\rangle/d)^k\pi_d(\mathrm{d}g)\Big]^{1/k} \\
&= \frac{4k^2}{d}\mathbb{E}\big[F_k(\boldsymbol{x})^2\big]^{(k-1)/k}\times\mathbb{E}\big[F_k(\boldsymbol{x})\big]^{1/k} \le \frac{4k^2}{d}\mathbb{E}\big[F_k(\boldsymbol{x})^2\big]^{(2k-1)/(2k)}.
\end{aligned}$$

Combining these two bounds yields the result for k even. ■

Lemma 21 *Let $\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{Q}^d)$. Let $\mathcal{G}_d$ be a general invariant group that preserves $\mathcal{Q}^d$. Let $F_k(\boldsymbol{z})$ be defined as in Lemma 19. Then for any fixed $k\ge 1$, there exists a constant C_k, such that*

- *If k is odd, then*
$$\mathrm{Var}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{Q}^d)}[F_k(\boldsymbol{\theta})] \le \frac{C_k}{d}\mathbb{E}_{\boldsymbol{\theta}}[F_{k-1}(\boldsymbol{\theta})^2] + \frac{C_k}{d^3}.$$
- *If k is even, then*
$$\mathrm{Var}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{Q}^d)}[F_k(\boldsymbol{\theta})] \le \frac{C_k}{d}\Big(\mathbb{E}_{\boldsymbol{\theta}}[F_{k-2}(\boldsymbol{\theta})^2]\wedge\mathbb{E}_{\boldsymbol{\theta}}[F_k(\boldsymbol{\theta})^2]^{(2k-1)/(2k)}\Big) + \frac{C_k}{d^3}.$$

Proof [Proof of Lemma 21]

The proof is similar to the proof of Lemma 20. By the discrete Poincaré inequality, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathcal{Q}_d)}\big[\big(F_k(\boldsymbol{\theta})-\mathbb{E}[F_k(\boldsymbol{\theta})]\big)^2\big] \le \mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{i=1}^d D_iF_k(\boldsymbol{\theta})^2\Big]\,,$$

where D_i denote the discrete derivative defined as

$$D_if(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_{-i})}{2},$$

with $\boldsymbol{\theta}_{-i} = (\theta_1, \dots, \theta_{i-1}, -\theta_i, \theta_{i+1}, \dots, \theta_d)$. Let $\varphi_g(\boldsymbol{\theta}) = (\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle / d)^k$, then

$$\begin{aligned} D_i \varphi_g(\boldsymbol{\theta}) &= \frac{(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle / d)^k - (\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta} \rangle / d)^k}{2} + \frac{(\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta} \rangle / d)^k - (\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta}_{-i} \rangle / d)^k}{2} \\ &=: D_{i,1} \varphi_g(\boldsymbol{\theta}) + D_{i,2} \varphi_g(\boldsymbol{\theta}). \end{aligned}$$

We have $\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta} \rangle = \langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle - 2\theta_i (g \cdot \boldsymbol{\theta})_i$. By Taylor expansion, the first term verifies (recall that $g \cdot \boldsymbol{\theta} \in \mathscr{Q}^d$ and $\theta_i^2 (g \cdot \boldsymbol{\theta})_i^2 = 1$)

$$D_{i,1} \varphi_g(\boldsymbol{\theta}) = k(\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle / d)^{k-1} (\theta_i (g \cdot \boldsymbol{\theta})_i / d) - k(k-1) X_{i,1}(\boldsymbol{\theta}, g)^{k-2} / d^2,$$

where $X_{i,1}(\boldsymbol{\theta}, g)$ is on the line segment between $\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta} \rangle / d$ and $\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta} \rangle / d$. Similarly, Taylor expansion on the second term yields

$$D_{i,2} \varphi_g(\boldsymbol{\theta}) = k(\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta}_{-i} \rangle / d)^{k-1} (\theta_i (g^{-1} \cdot \boldsymbol{\theta}_{-i})_i / d) + k(k-1) X_{i,2}(\boldsymbol{\theta}, g)^{k-2} / d^2,$$

where $X_{i,2}(\boldsymbol{\theta}, g)$ is on the line segment between $\langle \boldsymbol{\theta}_{-i}, g \cdot \boldsymbol{\theta}_{-i} \rangle / d$ and $\langle \boldsymbol{\theta}, g \cdot \boldsymbol{\theta}_{-i} \rangle / d$.

Using Jensen's inequality to separate each of the 4 terms in $D_i \varphi_g(\boldsymbol{\theta})$, using that $\boldsymbol{\theta}_{-i}$ and $\boldsymbol{\theta}$ have the same distribution, we get

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}} \Big[\sum_{i=1}^d D_i F_k(\boldsymbol{\theta})^2 \Big] \\ &\le \frac{32k^2}{d} \int_{\mathcal{G}_d \times \mathcal{G}_d} \mathbb{E} \Big[(\langle \boldsymbol{\theta}, g_1 \cdot \boldsymbol{\theta} \rangle / d)^{k-1} (\langle \boldsymbol{\theta}, g_2 \cdot \boldsymbol{\theta} \rangle / d)^{k-1} \langle \boldsymbol{\theta}, g_1 g_2 \cdot \boldsymbol{\theta} \rangle / d \Big] \pi_d(\mathrm{d}g_1) \pi_d(\mathrm{d}g_2) \\ &\quad + \frac{16k^2(k-1)^2}{d^4} \sum_{s \in \{1,2\}} \sum_{i=1}^d \int_{\mathcal{G}_d \times \mathcal{G}_d} \mathbb{E} \Big[X_{i,s}(\boldsymbol{\theta}, g_1)^{k-2} X_{i,s}(\boldsymbol{\theta}, g_2)^{k-2} \Big] \pi_d(\mathrm{d}g_1) \pi_d(\mathrm{d}g_2) \,. \end{aligned}$$

Noticing that $\sup_{i,s,\boldsymbol{\theta},g} |X_{i,s}(\boldsymbol{\theta}, g)| \le 1$, the second term in the above equation can be bounded by C_k / d^3. The first term in the above equation can be bounded using the same way as bounding the right hand side of Eq. (49) as in the proof of Lemma 20. This concludes the proof. ■

Appendix F. Kernel concentration for the cyclic group and general σ

Throughout this section, we will always take $\mathcal{G}_d = \mathrm{Cyc}_d$ to be the cyclic group, and $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$ to be the sphere. We will write in short $B(d,k) = B(\mathbb{S}^{d-1}(\sqrt{d}); k)$ and $D(d,k) = D(\mathbb{S}^{d-1}(\sqrt{d}); \mathrm{Cyc}_d; k)$. We recall that the cyclic group has degeneracy 1, i.e., for each integers $k \geq 1$, $B(d,k)/D(d,k) = \Theta_d(d)$.

F.1. Main propositions

Let the Gegenbauer decomposition of σ be

$$\sigma(x) = \sum_{k=0}^{\infty} \xi_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x).$$

For $S \subseteq \mathbb{N}$, we define

$$\sigma_{d,S}(x) = \sum_{k \in S} \xi_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x). \tag{50}$$

For any $\|\boldsymbol{\theta}_1\|_2 = \|\boldsymbol{\theta}_2\|_2 = \sqrt{d}$ and any $S \subseteq \mathbb{Z}_{\geq 0}$, denote

$$h_{d,S}(\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \rangle / d) \equiv \mathbb{E}_{\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma_{d,S}(\langle \boldsymbol{\theta}_1, \boldsymbol{x} \rangle / \sqrt{d}) \sigma_{d,S}(\langle \boldsymbol{\theta}_2, \boldsymbol{x} \rangle / \sqrt{d})].$$

Proposition 22 *Let $\ell \geq 2$ be a fixed integer. Assume that $\sigma \in C^{\ell \vee 3}(\mathbb{R})$ be a $\ell \vee 3$'th continuously differentiable function with derivatives satisfy $\sup_{0 \leq k \leq \ell \vee 3} \sigma^{(k)}(u) \leq c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.*

Define $H_{d,S} : \mathbb{S}^{d-1}(\sqrt{d}) \times \mathbb{S}^{d-1}(\sqrt{d}) \to \mathbb{R}$ via

$$H_{d,S}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv \int_{\mathrm{Cyc}_d} h_{d,S}(\langle \boldsymbol{\theta}_1, g \cdot \boldsymbol{\theta}_2 \rangle / d) \pi_d(\mathrm{d}g). \tag{51}$$

Then, for $N = d^p$ for any fixed p, letting $(\boldsymbol{\theta}_i)_{i \in [N]} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$\sup_{i \in [N]} \left| H_{d,\geq \ell}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) - \mathbb{E} H_{d,\geq \ell}(\boldsymbol{\theta}, \boldsymbol{\theta}) \right| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E} H_{d,\geq \ell}(\boldsymbol{\theta}, \boldsymbol{\theta}). \tag{52}$$

Moreover, we have $\mathbb{E} H_{d,\geq \ell}(\boldsymbol{\theta}, \boldsymbol{\theta}) = O_d(d^{-1})$.

Proof [Proof of Proposition 22] We let C, C_k, $C_{k,\ell}$ be constants that depend on σ, k, and ℓ but independent of dimension d. The exact values of these constant can change from line to line.

Step 1. Finite subset $S \subseteq \{2, 3, \ldots\}$. Note we have

$$H_{d,S}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv \sum_{k \in S} \xi_{d,k}(\sigma)^2 B(d,k) \int_{\mathrm{Cyc}_d} Q_k^{(d)}(\langle \boldsymbol{\theta}_1, g \cdot \boldsymbol{\theta}_2 \rangle) \pi_d(\mathrm{d}g).$$

By Lemma 11 and Proposition 12, for any $S \subseteq \mathbb{N}$ with finite cardinality $|S| < \infty$, we have

$$\mathbb{E}[H_{d,S}(\boldsymbol{\theta}, \boldsymbol{\theta})] = \sum_{k \in S} \xi_{d,k}(\sigma)^2 D(d,k) = \Theta_d(d^{-1}).$$

Moreover, by Proposition 18, we have

$$\sup_{i\in[N]}\Big|H_{d,S}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)-\mathbb{E}[H_{d,S}(\boldsymbol{\theta},\boldsymbol{\theta})]\Big| \le \sum_{k\in S}\xi_{d,k}(\sigma)^2 D(d,k)\sup_{i\in[N]}\Big|\Upsilon_k(\boldsymbol{\theta}_i)-1\Big| = o_d(1)\cdot\mathbb{E}[H_{d,S}(\boldsymbol{\theta},\boldsymbol{\theta})].$$

Step 2. For general set $S=\{u:u\ge\ell\}$.

By Lemma 27, we have $\sup_{d\ge1}\sup_{\gamma\in[-1,1]}|h^{(\ell)}_{d,\ge\ell}(\gamma)|\le C_\ell$. Therefore, for any $\gamma\in[-1,1]$, we have

$$\Big|h_{d,\ge\ell}(\gamma)-\sum_{k=0}^{\ell-1}\frac{1}{k!}h^{(k)}_{d,\ge\ell}(0)\gamma^k\Big|\le C_\ell\cdot|\gamma|^{\ell+1}. \tag{53}$$

By Lemma 28, for any $k\le\ell-1$, we have

$$\Big|h^{(k)}_{d,\ge\ell}(0)\Big|\le C_{k,\ell}\cdot d^{-(\ell-k)/2}. \tag{54}$$

Moreover, by the Hanson-Wright inequality as in Lemma 14, since N is at most polynomial in d, then for any $\delta>0$, we have

$$\sup_{1\le k\le\ell+1}\sup_{g\in\mathrm{Cyc}_d\setminus\mathbf{I}}\sup_{i\in[N]}\Big|\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle^k\Big|\cdot d^{-k/2-\delta}=o_{d,\mathbb{P}}(1), \tag{55}$$

and

$$\sup_{1\le k\le\ell+1}\sup_{g\in\mathrm{Cyc}_d\setminus\mathbf{I}}\mathbb{E}\Big[\Big|\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle^k\Big|\Big]\cdot d^{-k/2-\delta}=o_d(1). \tag{56}$$

Therefore, by Eq. (53), (54), (55) and (56), we have

$$\sup_{g\in\mathrm{Cyc}_d\setminus\mathbf{I}}\sup_{i\in[N]}\Big|h_{d,\ge\ell}(\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle/d)\Big|=O_{d,\mathbb{P}}(d^{-\ell/2+\delta}),$$

and

$$\sup_{g\in\mathrm{Cyc}_d\setminus\mathbf{I}}\mathbb{E}\Big[\Big|h_{d,\ge\ell}(\langle\boldsymbol{\theta},g\cdot\boldsymbol{\theta}\rangle/d)\Big|\Big]=O_d(d^{-\ell/2+\delta}).$$

As a result, for any $\ell\ge3$, we have

$$\begin{aligned}
&\sup_{i\in[N]}\Big|H_{d,\ge\ell}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)-\mathbb{E}H_{d,\ge\ell}(\boldsymbol{\theta},\boldsymbol{\theta})\Big|\\
=&\sup_{i\in[N]}\Big|\int_{\mathrm{Cyc}_d\setminus\mathbf{I}}h_{d,\ge\ell}(\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle/d)\pi_d(\mathrm{d}g)-\mathbb{E}\int_{\mathrm{Cyc}_d\setminus\mathbf{I}}h_{d,\ge\ell}(\langle\boldsymbol{\theta},g\cdot\boldsymbol{\theta}\rangle/d)\pi_d(\mathrm{d}g)\Big|\\
\le&\sup_{i\in[N]}\Big|\int_{\mathrm{Cyc}_d\setminus\mathbf{I}}h_{d,\ge\ell}(\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle/d)\pi_d(\mathrm{d}g)\Big|+\Big|\mathbb{E}\int_{\mathrm{Cyc}_d\setminus\mathbf{I}}h_{d,\ge\ell}(\langle\boldsymbol{\theta},g\cdot\boldsymbol{\theta}\rangle/d)\pi_d(\mathrm{d}g)\Big|\\
\le&O_{d,\mathbb{P}}(d^{-\ell/2+\delta})=o_{d,\mathbb{P}}(d^{-1}).
\end{aligned}$$

By the arguments in Step 1, for any $\ell\ge2$, we have

$$\sup_{i\in[N]}\Big|H_{d,\ge\ell}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)-\mathbb{E}H_{d,\ge\ell}(\boldsymbol{\theta},\boldsymbol{\theta})\Big|\le o_{d,\mathbb{P}}(d^{-1}).$$

Finally, for any $\ell \geq 2$, for any σ such that $\sigma_{d,\geq \ell}$ that is non-trivial (if $\sigma_{d,\geq \ell} = 0$, this proposition holds trivially), we have

$$\mathbb{E} H_{d,\geq \ell}(\boldsymbol{\theta}, \boldsymbol{\theta}) = \Theta_d(d^{-1}).$$

This proves the proposition. ■

Proposition 23 *Let $\ell \geq 2$ be a fixed integer. Assume that $\sigma \in C(\mathbb{R})$ be a continuous function with $|\sigma(u)| \leq c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.*

Define $H_{d,S} : \mathbb{S}^{d-1}(\sqrt{d}) \times \mathbb{S}^{d-1}(\sqrt{d}) \to \mathbb{R}$ via

$$H_{d,S}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv \int_{\mathrm{Cyc}_d} h_{d,S}(\langle \boldsymbol{\theta}_1, g \cdot \boldsymbol{\theta}_2 \rangle / d) \pi_d(\mathrm{d}g). \tag{57}$$

Then, for $N = O(d^p)$ for any fixed p, letting $(\boldsymbol{\theta}_i)_{i \in [N]} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$\sup_{i \in [N]} \left| \mathbb{E}_{\boldsymbol{\theta}}[H_{d,\geq \ell}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2] - \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\geq \ell}(\boldsymbol{\theta}', \boldsymbol{\theta})^2] \right| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\geq \ell}(\boldsymbol{\theta}', \boldsymbol{\theta})^2]. \tag{58}$$

Proof [Proof of Proposition 23]

Denoting $\mu_k(\sigma) = \mathbb{E}_{G \sim \mathcal{N}(0,1)}[\sigma(G) \mathrm{He}_k(G)]$. Let $q = \min\{k \geq \ell : \mu_k(\sigma) \neq 0\}$ and let $u = q + 2$. We consider the case when $q < \infty$, since for $q = \infty$, the claim holds trivially. We have the expression

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[H_{d,\geq \ell}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2] &= \sum_{k=\ell}^{\infty} \xi_{d,k}(\sigma)^4 B(d,k) \int_{\mathrm{Cyc}_d} Q_k^{(d)}(\langle \boldsymbol{\theta}_i, g \cdot \boldsymbol{\theta}_i \rangle) \pi_d(\mathrm{d}g) \\
&= \mathbb{E}_{\boldsymbol{\theta}}[H_{d,[\ell,u)}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2] + \mathbb{E}_{\boldsymbol{\theta}}[H_{d,\geq u}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2].
\end{aligned}$$

Step 1. Upper bounding $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\geq u}(\boldsymbol{\theta}', \boldsymbol{\theta})^2]$ and $\mathbb{E}_{\boldsymbol{\theta}}[H_{d,\geq u}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2]$. We have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_i} \mathbb{E}_{\boldsymbol{\theta}}[H_{d,\geq u}(\boldsymbol{\theta}_i, \boldsymbol{\theta})^2] \\
&= \sup_{\boldsymbol{\theta}_i} \sum_{k=u}^{\infty} \xi_{d,k}(\sigma)^4 \sum_{l \in [D(d,k)]} \overline{Y}_{kl}^{(d)}(\boldsymbol{\theta}_i)^2 \leq \sup_{\boldsymbol{\theta}_i} \sum_{k=u}^{\infty} \xi_{d,k}(\sigma)^4 \sum_{l \in [B(d,k)]} Y_{kl}^{(d)}(\boldsymbol{\theta}_i)^2 \\
&= \sum_{k=u}^{\infty} \xi_{d,k}(\sigma)^4 B(d,k) Q_k^{(d)}(d) = \sum_{k=u}^{\infty} \xi_{d,k}(\sigma)^4 B(d,k) \\
&\leq \Big[\sup_{k \geq u} B(d,k)^{-1} \Big] \cdot \Big[\sum_{k=u}^{\infty} \xi_{d,k}(\sigma)^2 B(d,k) \Big]^2 \\
&= B(d,u)^{-1} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d})^2] = \Theta_d(d^{-u}).
\end{aligned} \tag{59}$$

This also gives

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\geq u}(\boldsymbol{\theta}', \boldsymbol{\theta})^2] = \Theta_d(d^{-u}). \tag{60}$$

Step 2. Upper bounding $\sup_{i\in[N]}|\mathbb{E}_{\boldsymbol{\theta}}[H_{d,[\ell,u)}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2]-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,[\ell,u)}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]|$**.** By Proposition 18, we have

$$\begin{aligned}
&\sup_{i\in[N]}\left|\mathbb{E}_{\boldsymbol{\theta}}[H_{d,[\ell,u)}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2]-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,[\ell,u)}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]\right|\\
&\le \sum_{k=\ell}^{u-1}\xi_{d,k}(\sigma)^4 D(d,k)\sup_{i\in[N]}\left|\frac{B(d,k)}{D(d,k)}\int_{\mathrm{Cyc}_d}Q_k^{(d)}(\langle\boldsymbol{\theta}_i,g\cdot\boldsymbol{\theta}_i\rangle)\pi_d(\mathrm{d}g)-1\right| \qquad (61)\\
&= o_{d,\mathbb{P}}(1)\cdot\Big[\sum_{k=\ell}^{u-1}\xi_{d,k}(\sigma)^4 D(d,k)\Big] = o_{d,\mathbb{P}}(1)\cdot\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,[\ell,u)}(\boldsymbol{\theta}',\boldsymbol{\theta})^2].
\end{aligned}$$

Step 3. Lower bounding $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\ge\ell}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]$**.** We have

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\ge\ell}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]=\sum_{k=\ell}^{\infty}\xi_{d,k}(\sigma)^4 D(d,k)\ge\xi_{d,q}(\sigma)^4 D(d,q)=\Theta_d(d^{-q-1}). \qquad (62)$$

The last equality is by Proposition 12, and the fact that

$$\lim_{d\to\infty}\xi_{d,q}(\sigma)^2 B(d,q)=\mu_q(\sigma)^2/q!>0.$$

Step 4. Complete the proof. By Eq. (59), (60), (61) and (62), we have

$$\begin{aligned}
&\sup_{i\in[N]}\left|\mathbb{E}_{\boldsymbol{\theta}}[H_{d,\ge\ell}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2]-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\ge\ell}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]\right|\\
&\le \sup_{i\in[N]}\left|\mathbb{E}_{\boldsymbol{\theta}}[H_{d,[\ell,u)}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2]-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,[\ell,u)}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]\right|\\
&\quad+\sup_{i\in[N]}\left|\mathbb{E}_{\boldsymbol{\theta}}[H_{d,\ge\ell}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2]\right|+\left|\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\ge\ell}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]\right|\\
&\le o_{d,\mathbb{P}}(1)\cdot\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,[\ell,u)}(\boldsymbol{\theta}',\boldsymbol{\theta})^2]+\Theta_d(d^{-u})=o_{d,\mathbb{P}}(1)\cdot\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'}[H_{d,\ge\ell}(\boldsymbol{\theta}',\boldsymbol{\theta})^2].
\end{aligned}$$

This completes the proof. ∎

F.2. Auxiliary lemmas

The following lemma is a reformulation of (Ghorbani et al., 2021, Lemma 5).

Lemma 24 *Assume $\sigma\in C(\mathbb{R})$ with $\sigma(u)^2\le c_0\exp(c_1 u^2/2)$ for some constants $c_0>0$ and $c_1<1$. Then*

(a) $\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)^2]<\infty$.

(b) *Let $\|\boldsymbol{x}\|_2=\sqrt{d}$. Then there exists $d_0=d_0(c_1)$ such that, for $\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})$,*

$$\sup_{d\ge d_0}\mathbb{E}_{\boldsymbol{w}}[\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)^2]<\infty\,. \qquad (63)$$

(c) *Let* $\|\boldsymbol{x}\|_2 = \sqrt{d}$, $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ *and* $\tau \sim \chi(d)/\sqrt{d}$. *Then we have*

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w},\tau}\Big[\Big(\sigma(\tau\langle \boldsymbol{w},\boldsymbol{x}\rangle) - \sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\Big)^2\Big] = 0. \tag{64}$$

Lemma 25 *Assume that* $\psi, \phi \in C(\mathbb{R})$ *with* $\psi(u)^2, \phi(u)^2 \le c_0 \exp(c_1 u^2/2)$ *for some constants* $c_0 > 0$ *and* $c_1 < 1$. *Denote*

$$E_d[\psi,\phi](\gamma) \equiv \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\psi(\langle \boldsymbol{x},\boldsymbol{w}\rangle)\phi(\langle \boldsymbol{x}',\boldsymbol{w}\rangle)],$$
$$E[\psi,\phi](\gamma) \equiv \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I}_d/d)}[\psi(\langle \boldsymbol{x},\boldsymbol{g}\rangle)\phi(\langle \boldsymbol{x}',\boldsymbol{g}\rangle)],$$

where $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}'\|_2 = \sqrt{d}$ *such that* $\langle \boldsymbol{x},\boldsymbol{x}'\rangle/d = \gamma$ *(by an invariance argument,* E_d *and* E *do not depend on the choice of* $\boldsymbol{x}$ *and* $\boldsymbol{x}'$*). Then we have*

$$\lim_{d\to\infty} \sup_{\gamma\in[-1,1]} \Big|E_d[\psi,\phi](\gamma) - E[\psi,\phi](\gamma)\Big| = 0, \tag{65}$$

and

$$\sup_{\gamma\in[-1,1]} \Big|E[\psi,\phi](\gamma)\Big| < \infty. \tag{66}$$

Proof [Proof of Lemma 25] Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d/d)$, $\boldsymbol{w} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$ and $\tau = \|\boldsymbol{g}\|_2$. Then we have $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $\tau \sim \chi(d)/\sqrt{d}$ independently. We further denote

$$\overline{E}_d[\psi,\phi](\gamma) \equiv \mathbb{E}_{\boldsymbol{w},\tau}[\psi(\tau\langle \boldsymbol{x},\boldsymbol{w}\rangle)\phi(\langle \boldsymbol{x}',\boldsymbol{w}\rangle)]$$

where $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}'\|_2 = \sqrt{d}$ such that $\langle \boldsymbol{x},\boldsymbol{x}'\rangle/d = \gamma$. Note we have

$$\begin{aligned}
&\lim_{d\to\infty} \sup_{\gamma\in[-1,1]} \Big|E_d[\psi,\phi](\gamma) - \overline{E}_d[\psi,\phi](\gamma)\Big| \\
\le &\lim_{d\to\infty} \sup_{\gamma\in[-1,1]} \Big|\mathbb{E}_{\boldsymbol{w},\tau}\Big\{\Big[\psi(\tau\langle \boldsymbol{x},\boldsymbol{w}\rangle) - \psi(\langle \boldsymbol{x},\boldsymbol{w}\rangle)\Big]\phi(\langle \boldsymbol{x}',\boldsymbol{w}\rangle)\Big\}\Big| \\
\le &\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w},\tau}\Big\{\Big[\psi(\tau\langle \boldsymbol{x},\boldsymbol{w}\rangle) - \psi(\langle \boldsymbol{x},\boldsymbol{w}\rangle)\Big]^2\Big\}^{1/2} \mathbb{E}_{\boldsymbol{w}}[\phi(\langle \boldsymbol{x}',\boldsymbol{w}\rangle)^2]^{1/2} = 0,
\end{aligned}$$

where the last equality is by (b) and (c) in Lemma 24. Moreover, we have

$$\begin{aligned}
&\lim_{d\to\infty} \sup_{\gamma\in[-1,1]} \Big|\overline{E}_d[\psi,\phi](\gamma) - E[\psi,\phi](\gamma)\Big| \\
\le &\lim_{d\to\infty} \sup_{\gamma\in[-1,1]} \Big|\mathbb{E}_{\boldsymbol{w},\tau}\Big\{\Big[\phi(\tau\langle \boldsymbol{x},\boldsymbol{w}\rangle) - \phi(\langle \boldsymbol{x},\boldsymbol{w}\rangle)\Big]\psi(\tau\langle \boldsymbol{x}',\boldsymbol{w}\rangle)\Big\}\Big| \\
\le &\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w},\tau}\Big\{\Big[\phi(\tau\langle \boldsymbol{x},\boldsymbol{w}\rangle) - \phi(\langle \boldsymbol{x},\boldsymbol{w}\rangle)\Big]^2\Big\}^{1/2} \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\psi(G)^2]^{1/2} = 0,
\end{aligned}$$

where the last equality is by (a) and (c) in Lemma 24. Combining the two equations above proves Eq. (65).

Finally, note that we have

$$\begin{aligned}
\sup_{\gamma\in[-1,1]} \Big|E[\psi,\phi](\gamma)\Big| &\le \mathbb{E}_{\boldsymbol{g}}[\psi(\langle \boldsymbol{x},\boldsymbol{g}\rangle)^2]^{1/2} \mathbb{E}_{\boldsymbol{g}}[\phi(\langle \boldsymbol{x}',\boldsymbol{g}\rangle)^2]^{1/2} \\
&= \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\psi(G)^2]^{1/2} \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\phi(G)^2]^{1/2} < \infty.
\end{aligned}$$

This proves Eq. (66). ∎

Lemma 26 *Assume that $\sigma \in C^\ell(\mathbb{R})$ with derivatives satisfy $\sup_{0\le k\le \ell} |\sigma^{(k)}(u)|^2 \le c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$. Denote*

$$h_d(\gamma) \equiv \mathbb{E}_{\boldsymbol{w}\sim \mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma(\langle \boldsymbol{x}', \boldsymbol{w}\rangle)],$$
$$h(\gamma) \equiv \mathbb{E}_{\boldsymbol{g}\sim \mathcal{N}(\boldsymbol{0},\mathbf{I}_d/d)}[\sigma(\langle \boldsymbol{x}, \boldsymbol{g}\rangle)\sigma(\langle \boldsymbol{x}', \boldsymbol{g}\rangle)],$$

where $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}'\|_2 = \sqrt{d}$ such that $\langle \boldsymbol{x}, \boldsymbol{x}'\rangle/d = \gamma$ (by an invariance argument, h_d and h do not depend on the choice of $\boldsymbol{x}$ and $\boldsymbol{x}'$). Then we have

$$\lim_{d\to\infty} \sup_{0\le k\le \ell} \sup_{\gamma\in[-1,1]} \left| h_d^{(k)}(\gamma) - h^{(k)}(\gamma)\right| = 0,$$

and

$$\sup_{0\le k\le \ell} \sup_{\gamma\in[-1,1]} \left| h^{(k)}(\gamma)\right| < \infty.$$

Proof [Proof of Lemma 26]

For $k = 0$, the result is implied by Lemma 25 by observing that $h_d' = E_d[\sigma, \sigma]$ and $h' = E[\sigma, \sigma]$.

For $k = 1$, the result is implied by Lemma 25 by the fact that $h_d' = E_d[u\sigma(u), \sigma'(u)]$ and $h' = E[u\sigma(u), \sigma'(u)]$, and there exist constants $c_0 > 0$ and $c_1 < 1$ such that $\sigma'(u), u\sigma(u) \le c_0 e^{c_1 u^2/2}$. Indeed, for $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}'\|_2 = \sqrt{d}$ such that $\langle \boldsymbol{x}, \boldsymbol{x}'\rangle/d = \gamma$, we have (similarly for h')

$$\begin{aligned} h_d'(\gamma) &= \lim_{\delta\to 0} \delta^{-1}\Big\{\mathbb{E}_{\boldsymbol{w}}\Big[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma(\langle (1-\delta^2)^{1/2}\boldsymbol{x}' + \delta \boldsymbol{x}, \boldsymbol{w}\rangle)\Big] - \mathbb{E}_{\boldsymbol{w}}\Big[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma(\langle \boldsymbol{x}', \boldsymbol{w}\rangle)\Big]\Big\} \\ &= \mathbb{E}_{\boldsymbol{w}}\Big[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma'(\langle \boldsymbol{x}', \boldsymbol{w}\rangle)\langle \boldsymbol{x}, \boldsymbol{w}\rangle\Big] = E_d[u\sigma(u), \sigma'(u)](\gamma). \end{aligned}$$

By an induction argument, for any fixed k, $h_d^{(k)}$ can be identified by a fixed number of combinations of $E_d[\psi, \phi]$ with $\psi, \phi \in \Lambda_k \equiv \{u^s\sigma^{(t)}(u)\}_{0\le s,t\le k}$. Further, for any fixed k, there exists $c_{0,k} > 0$ and $c_{1,k} < 1$ such that, for any $\psi \in \Lambda_k$, we have $\psi(u) \le c_{0,k} e^{c_{1,k}u^2/2}$. Applying Lemma 25 proves the lemma. ∎

Lemma 27 *Assume that $\sigma \in C^k(\mathbb{R})$ with derivatives satisfy $\sup_{0\le s\le k} |\sigma^{(s)}(u)|^2 \le c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$. For any $\|\boldsymbol{\theta}_1\|_2 = \|\boldsymbol{\theta}_2\|_2 = \sqrt{d}$ and any $\ell \ge 1$, denote*

$$h_{d,S}(\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle/d) \equiv \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma_{d,S}(\langle \boldsymbol{\theta}_1, \boldsymbol{x}\rangle/\sqrt{d})\sigma_{d,S}(\langle \boldsymbol{\theta}_2, \boldsymbol{x}\rangle/\sqrt{d})].$$

where $\sigma_{d,S}$ is given in Eq. (50). Then we have

$$\sup_{d\ge 1} \sup_{\gamma\in[-1,1]} \left| h_{d,\ge \ell}^{(k)}(\gamma)\right| \le C_{k,\ell}.$$

Proof [Proof of Lemma 27] Note we have

$$h_{d,\ge \ell}(\gamma) = h_d(\gamma) - h_{d,<\ell}(\gamma).$$

By Lemma 26, we have

$$\sup_{d\ge 1} \sup_{\gamma\in[-1,1]} \left| h_d^{(k)}(\gamma)\right| \le C_k.$$

Moreover, since $h_{d,<\ell}(\gamma)$ is a degree $\ell-1$ polynomial on $[-1,1]$ and its coefficients converge to the coefficients of $h_{<\ell}$ with $h_{<\ell}(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle/d)=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\sigma_{d,<\ell}(\langle\boldsymbol{\theta}_1,\boldsymbol{x}\rangle/\sqrt{d})\sigma_{d,<\ell}(\langle\boldsymbol{\theta}_2,\boldsymbol{x}\rangle/\sqrt{d})]$. Then, it is easy to see that

$$\sup_{d\geq 1}\sup_{\gamma\in[-1,1]}\left|h^{(k)}_{d,<\ell}(\gamma)\right|\leq C_{k,\ell}.$$

This proves the lemma. ■

Lemma 28 *Assume that $\sigma\in C^{\ell}(\mathbb{R})$ with derivatives satisfy $\sup_{0\leq s\leq\ell}|\sigma^{(s)}(u)|^2\leq c_0\,\exp(c_1\,u^2/2)$ for some constants $c_0>0$ and $c_1<1$. Then there exists constant $C_{k,\ell}$, such that*

$$\left|h^{(k)}_{d,\geq\ell}(0)\right|\leq C_{k,\ell}\cdot d^{-(\ell-k)/2}.$$

Proof [Proof of Lemma 28]

We let C, C_k, $C_{k,\ell}$ be constants that depend on σ, k, and ℓ but independent of dimension d. The exact values of these constant can change from line to line.

We let $\tilde{\tau}_d$ be the measure of $\langle\boldsymbol{e}_1,\boldsymbol{x}\rangle$ when $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ (hence converging weakly to a standard Gaussian), and $\tilde{Q}^{(d)}_k(x)=\sqrt{B(d,k)}Q^{(d)}_k(x/\sqrt{d})$ to be the rescaled Gegenbauer polynomials, forming an orthonormal system with respect to $\tilde{\tau}_d$. In particular $\tilde{Q}^{(d)}_k$ converges to the k-th Hermite polynomial. We let $\langle\,\cdot\,,\,\cdot\,\rangle$ denote the scalar product with respect to $\tilde{\tau}_d$.

By the definition of $h_{d,\geq\ell}$, we have

$$\langle h_{d,\geq\ell}(\,\cdot\,/\sqrt{d}),\tilde{Q}^{(d)}_k\rangle=0,\quad\forall k\leq\ell-1. \tag{67}$$

Let $\tilde{h}_{d,\geq\ell}(x)$ be obtained from $h_{d,\geq\ell}(x)$ by removing its Taylor expansion up to term $x^{\ell-1}$, i.e., we have

$$\tilde{h}_{d,\geq\ell}(x)=h_{d,\geq\ell}(x)-\sum_{k=0}^{\ell-1}\frac{h^{(k)}_{d,\geq\ell}(0)}{k!}x^k.$$

Then Eq. (67) gives

$$\begin{aligned}&\sum_{j=0}^{\ell-1}\langle\tilde{Q}^{(d)}_k,x\rangle^j\Big(\frac{h^{(k)}_{d,\geq\ell}(0)}{j!d^{j/2}}\Big)=-\frac{\Delta_k(d)}{d^{\ell/2}},\quad\forall k\leq\ell-1,\\&\Delta_k(d)\equiv d^{\ell/2}\langle\tilde{h}_{d,\geq\ell}(\,\cdot\,/\sqrt{d}),\tilde{Q}_k\rangle.\end{aligned} \tag{68}$$

We claim that $\sup_{d\geq1}|\Delta_k(d)|\leq C_{k,\ell}$. Indeed, by Rodrigues formula, there exist non-negative constants $A_{d,k}$, $\tilde{A}_{d,k}$ with $\sup_{d\geq1}A_{d,k}\vee\tilde{A}_{d,k}\leq C_k$, such that

$$\begin{aligned}\Delta_k(d)&=(-1)^k d^{\ell/2}A_{d,k}\int_{-1}^{1}\tilde{h}_{d,\geq\ell}\big(x/\sqrt{d}\big)\,\frac{\mathrm{d}^k}{\mathrm{d}x^k}\Big(1-\frac{x^2}{d}\Big)^{\frac{d-3}{2}+k}\mathrm{d}x\\&=A_{d,k}\,d^{(\ell-k)/2}\int_{-1}^{1}\tilde{h}^{(k)}_{d,\geq\ell}\big(x/\sqrt{d}\big)\Big(1-\frac{x^2}{d}\Big)^{\frac{d-3}{2}+k}\mathrm{d}x\\&=\tilde{A}_{d,k}\,d^{(\ell-k)/2}\cdot\mathbb{E}_{X_d\sim\tilde{\tau}_d}\Big\{\tilde{h}^{(k)}_{d,\geq\ell}\big(X_d/\sqrt{d}\big)\Big(1-\frac{X_d^2}{d}\Big)^k\Big\}.\end{aligned} \tag{69}$$

By the definition of $\tilde{h}_{d,\ge\ell}$, using the Taylor expansion in the integral form, we have

$$\tilde{h}_{d,\ge\ell}(\gamma) = \int_0^\gamma h^{(\ell)}_{d,\ge\ell}(u)\frac{(\gamma-u)^{\ell-1}}{(\ell-1)!}\mathrm{d}u,$$

and hence for any $k \le \ell-1$, we have

$$\tilde{h}^{(k)}_{d,\ge\ell}(\gamma) = \int_0^\gamma h^{(\ell)}_{d,\ge\ell}(u)\frac{(\gamma-u)^{\ell-1-k}}{(\ell-1-k)!}\mathrm{d}u,$$

so that for any $\gamma \in [-1,1]$, we have

$$\sup_{d\ge1}|\tilde{h}^{(k)}_{d,\ge\ell}(\gamma)| \le C_{k,\ell}\cdot\sup_{d\ge1}\sup_{u\in[-1,1]}|h^{(\ell)}_{d,\ge\ell}(u)|\cdot|\gamma|^{\ell-k} \le C_{k,\ell}\cdot|\gamma|^{\ell-k}.$$

The last inequality is by Lemma 27 (here we used the assumption that $\sigma \in C^\ell(\mathbb{R})$). Therefore, by Eq. (69), we have (note X_d converges in distribution to a standard Gaussian random variable)

$$|\Delta_k(d)| \le C_{k,\ell}\cdot\mathbb{E}_{X_d\sim\tilde{\tau}_d}\{|X_d|^{\ell-k}\} \le C_{k,\ell}. \tag{70}$$

To conclude, we reconsider Eq. (68). Let $\boldsymbol{M}(d) = (M_{k,q}(d))_{0\le k,q\le\ell-1} \in \mathbb{R}^{\ell\times\ell}$ be the matrix with entries $M_{k,q}(d) \equiv \langle\tilde{Q}^{(d)}_k, x^q\rangle$, $\boldsymbol{\xi}(d) = (\xi_q(d))_{0\le q\le\ell-1} \in \mathbb{R}^\ell$ the vector with entries $\xi_q(d) \equiv h^{(q)}_{d,\ge\ell}(0)/(q!d^{q/2})$, and $\boldsymbol{\Delta}(d) = (\Delta_0(d),\dots,\Delta_{\ell-1}(d))^{\mathsf{T}} \in \mathbb{R}^\ell$. We can therefore rewrite this equation as

$$\boldsymbol{M}(d)\boldsymbol{\xi}(d) = \boldsymbol{\Delta}(d)/d^{\ell/2}. \tag{71}$$

As $d\to\infty$, $\boldsymbol{M}(d)$ converges entrywise to $\boldsymbol{M}(\infty) = (M_{k,q}(\infty))_{0\le k,q\le\ell-1}$, whereby

$$M_{k,q}(\infty) \equiv \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\mathrm{He}_k(G)G^q]/\sqrt{k!}.$$

Since $\boldsymbol{M}(\infty)$ is non-singular (because the Hermite polynomials are a basis), it follows that $\sigma_{\min}(\boldsymbol{M}(d))$ is bounded away from zero for d large enough, and therefore $\sup_{d\ge1}\sigma_{\max}(\boldsymbol{M}(d)^{-1}) < \infty$. Therefore combining with Eq. (70), we get

$$\|\boldsymbol{\xi}(d)\|_2 \le C_\ell\cdot\|\boldsymbol{\Delta}(d)\|_2\cdot d^{-\ell/2} \le C_\ell\cdot d^{-\ell/2}. \tag{72}$$

Therefore, for any $0\le k\le\ell-1$, we have

$$\left|h^{(k)}_{d,\ge\ell}(0)\right| \le k!d^{k/2}|\xi_k(d)| \le C_{k,\ell}\cdot d^{-(\ell-k)/2}. \tag{73}$$

This proves the lemma. ∎

Appendix G. Hypercontractivity of general activation σ for $(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$

Let us consider an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and denote for $\boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d})$ and $\boldsymbol{w} \in \mathbb{S}^{d-1}(1)$,

$$\overline{\sigma}(\boldsymbol{x}; \boldsymbol{w}) = \int_{\mathrm{Cyc}_d} \sigma(\langle \boldsymbol{x}, g \cdot \boldsymbol{w} \rangle) \pi_d(\mathrm{d}g) = \frac{1}{d} \sum_{i=0}^{d-1} \sigma(\langle \boldsymbol{x}, \boldsymbol{L}^i \boldsymbol{w} \rangle),$$

where $\boldsymbol{L} \in \mathbb{R}^{d \times d}$ is the cyclic permutation matrix that shifts the coordinates by one (hence $\boldsymbol{L}^i$ shifts the coordinates by i).

Denote $\overline{\sigma}_{>\ell} = \overline{\mathsf{P}}_{>\ell} \overline{\sigma}$ the projection of $\overline{\sigma}$ orthogonal to cyclic polynomials of degree less or equal to ℓ. From the discussion in Section C.2, we have

$$\overline{\mathsf{P}}_{>\ell} \overline{\sigma}(\,\cdot\,; \boldsymbol{w}) = \overline{\mathsf{P}}_{>\ell} \mathcal{S}[\sigma(\langle\,\cdot\,, \boldsymbol{w}\rangle/\sqrt{d})] = \mathcal{S} \overline{\mathsf{P}}_{>\ell}[\sigma(\langle\,\cdot\,, \boldsymbol{w}\rangle/\sqrt{d})]\,,$$

where $\mathcal{S} : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Cyc}_d)$ is the symmetrization operator defined in Section C.1 and $\overline{\mathsf{P}}_{>\ell} : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ is the projection orthogonal to (general) polynomials of degree less or equal to ℓ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ (see Section H). Hence, denoting $\sigma_{>\ell} = \overline{\mathsf{P}}_{>\ell} \sigma$, we have

$$\overline{\sigma}_{>\ell}(\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{d} \sum_{i=0}^{d-1} \sigma_{>\ell}(\langle \boldsymbol{x}, \boldsymbol{L}^i \boldsymbol{w} \rangle)\,. \tag{74}$$

Proposition 29 *Consider fixed integers $m \geq 1$ and $\ell \geq 4m$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a differentiable activation function such that $|\sigma(x)|, |\sigma'(x)| \leq c_0 \exp(c_1 x^2/(8m))$ for some constants $c_0 > 0$ and $c_1 < 1$. Let $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, then for any $\varepsilon > 0$,*

$$\mathbb{E}_{\boldsymbol{x}, \boldsymbol{w}} \left[\overline{\sigma}_{>\ell}(\boldsymbol{x}; \boldsymbol{w})^{2m} \right]^{1/(2m)} = d^{\varepsilon - 1/2} \cdot O_d(1). \tag{75}$$

G.1. Proof of Proposition 29

The goal of this proof is to replace $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ by $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_d)$ and using Proposition 30 (stated in Section G.2), which is the Gaussian equivalent of Proposition 29.

Recall that $\sigma_{>\ell}$ is defined as the projection of σ orthogonal to degree ℓ polynomials with respect to the distribution $\langle \boldsymbol{x}, \boldsymbol{e} \rangle$ with $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\|\boldsymbol{e}\|_2 = 1$ arbitrary. We can write it explicitly in terms of Gegenbauer polynomials:

$$\sigma_{>\ell}(x) = \sigma(x) - \sum_{k=0}^{\ell} \xi_{d,k} B(\mathbb{S}^{d-1}; k) Q_k(\sqrt{d} x).$$

Let us introduce $\varphi_{>\ell}$ defined as the projection of σ orthogonal to degree ℓ polynomials with respect to the Gaussian measure. It is given explicitely by

$$\varphi_{>\ell}(x) = \sigma(x) - \sum_{k=0}^{\ell} \frac{\mu_k(\sigma)}{k!} \mathrm{He}_k(x),$$

where He_k denote the k-th Hermite polynomial (see Section H.1.3 for definitions).

Consider the symmetrized activation functions

$$\overline{\sigma}_{>\ell}(\boldsymbol{x};\boldsymbol{w}) = \overline{\sigma}(\boldsymbol{x};\boldsymbol{w}) - \sum_{k=0}^{\ell} \xi_{d,k} B(\mathbb{S}^{d-1};k)\overline{Q}_k(\boldsymbol{x};\boldsymbol{w})\,,$$
$$\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w}) = \overline{\sigma}(\boldsymbol{g};\boldsymbol{w}) - \sum_{k=0}^{\ell} \frac{\mu_k(\sigma)}{k!}\overline{\mathrm{He}}_k(\boldsymbol{g};\boldsymbol{w})\,,$$

where we denoted the symmetrized polynomials

$$\overline{Q}_k(\boldsymbol{x};\boldsymbol{w}) = \frac{1}{d}\sum_{i=0}^{d-1} Q_k(\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{L}^i\boldsymbol{w}\rangle)\,,$$
$$\overline{\mathrm{He}}_k(\boldsymbol{g};\boldsymbol{w}) = \frac{1}{d}\sum_{i=0}^{d-1} \mathrm{He}_k(\langle \boldsymbol{g}, \boldsymbol{L}^i\boldsymbol{w}\rangle)\,.$$

Consider $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_d)$ and $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$. Because $\langle \boldsymbol{x}, \boldsymbol{e}\rangle$ converges in distribution to a normal distribution, we expect the moments of $\overline{\sigma}_{>\ell}(\boldsymbol{x};\boldsymbol{w})$ to converge to the moments of $\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})$. Let us show that this convergence occurs with rate $O_d(d^{\varepsilon-1/2})$. By triangle inequality, we have

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\big(\overline{\sigma}_{>\ell}(\sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2;\boldsymbol{w}) - \overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})\big)^{2m}\Big]^{1/(2m)} \le R_1 + R_2 + R_3 + R_4\,,$$

with

$$R_1 = \mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\big(\overline{\sigma}(\sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2;\boldsymbol{w}) - \overline{\sigma}(\boldsymbol{g};\boldsymbol{w})\big)^{2m}\Big]^{1/(2m)},$$
$$R_2 = \mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\big(A_{\le 2}(\sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2;\boldsymbol{w}) - B_{\le 2}(\boldsymbol{g};\boldsymbol{w})\big)^{2m}\Big]^{1/(2m)},$$
$$R_3 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{w}}\Big[A_{[3:\ell]}(\boldsymbol{x};\boldsymbol{w})^{2m}\Big]^{1/(2m)},$$
$$R_4 = \mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[B_{[3:\ell]}(\boldsymbol{g};\boldsymbol{w})^{2m}\Big]^{1/(2m)},$$

where we denoted $[3:\ell] = \{3,\dots,\ell\}$ and for any subset $S \subset \{0,\dots,\ell\}$,

$$A_S(\boldsymbol{x};\boldsymbol{w}) = \sum_{k\in S} \xi_{d,k} B(\mathbb{S}^{d-1};k)\overline{Q}_k(\boldsymbol{x};\boldsymbol{w})\,,$$
$$B_S(\boldsymbol{g};\boldsymbol{w}) = \sum_{k\in S} \frac{\mu_k(\sigma)}{k!}\overline{\mathrm{He}}_k(\boldsymbol{g};\boldsymbol{w})\,.$$

Step 1. Bound on R_1.

Denote $\tau = \|\boldsymbol{g}\|_2/\sqrt{d}$ and $\boldsymbol{x} = \sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2$, such that τ and $\boldsymbol{x}$ are independent, and $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. By the mean value theorem, there exists $\tilde{\tau}$ on the line segment between 1 and τ such that

$$\overline{\sigma}(\tau\cdot\boldsymbol{x};\boldsymbol{w}) - \overline{\sigma}(\boldsymbol{x};\boldsymbol{w}) = (\tau-1)\langle \nabla_{\boldsymbol{x}}\overline{\sigma}(\tilde{\tau}\cdot\boldsymbol{x};\boldsymbol{w}), \boldsymbol{x}\rangle.$$

By Cauchy-Schwarz inequality, we get

$$\begin{aligned}
R_1 &= \mathbb{E}_{\tau,\boldsymbol{x},\boldsymbol{w}}\Big[\big(\overline{\sigma}(\tau\cdot\boldsymbol{x};\boldsymbol{w})-\overline{\sigma}(\boldsymbol{x};\boldsymbol{w})\big)^{2m}\Big]^{1/(2m)}\\
&= \mathbb{E}_{\tau,\boldsymbol{x},\boldsymbol{w}}\Big[(\tau-1)^{2m}\langle\nabla_{\boldsymbol{x}}\overline{\sigma}(\tilde{\tau}\cdot\boldsymbol{x};\boldsymbol{w}),\boldsymbol{x}\rangle^{2m}\Big]^{1/(2m)}\\
&\le \mathbb{E}_{\tau}\big[(\tau-1)^{4m}\big]^{1/(4m)}\cdot\mathbb{E}_{\tau,\boldsymbol{x},\boldsymbol{w}}\Big[\langle\nabla_{\boldsymbol{x}}\overline{\sigma}(\tilde{\tau}\cdot\boldsymbol{x};\boldsymbol{w}),\boldsymbol{x}\rangle^{4m}\Big]^{1/(4m)}
\end{aligned}$$

Let us bound the first term:

$$\mathbb{E}_{\tau}\big[(\tau-1)^{4m}\big]^{1/(4m)}\le\mathbb{E}_{\tau}\big[(\tau^2-1)^{4m}\big]^{1/(4m)}\le C_{4m}\mathbb{E}_{\tau}\big[(\tau^2-1)^2\big]^{1/2}=C_{4m}\sqrt{\frac{2}{d}}\,, \tag{76}$$

where we used in the first inequality that $|\tau-1|\le|\tau^2-1|$ for $\tau\ge 0$; in the second inequality that τ^2-1 is a degree 2 polynomial in $\boldsymbol{g}\sim\mathsf{N}(0,\mathbf{I}_d)$ and verifies the hypercontractivity property of Lemma 32; last equality, that $d\cdot\tau^2=\|\boldsymbol{g}\|_2^2$ follows a chisquared distribution of degree d.

For the second term, we have

$$\langle\nabla_{\boldsymbol{x}}\overline{\sigma}(\tilde{\tau}\cdot\boldsymbol{x};\boldsymbol{w}),\boldsymbol{x}\rangle=\frac{1}{d}\sum_{i=0}^{d-1}\langle\boldsymbol{x},\boldsymbol{L}^i\boldsymbol{w}\rangle\sigma^{(1)}(\langle\tilde{\tau}\cdot\boldsymbol{x},\boldsymbol{L}^i\boldsymbol{w}\rangle)\,.$$

Recall that $\tilde{\tau}\cdot\boldsymbol{x}$ is between $\tau\cdot\boldsymbol{x}$ and $\boldsymbol{x}$ which have marginal distributions $\boldsymbol{g}\sim\mathsf{N}(0,\mathbf{I}_d)$ and $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ respectively. Denote x_1 the first coordinate of $\boldsymbol{x}$ (therefore $\tau\cdot x_1\sim\mathsf{N}(0,1)$). By Jensen's inequality and using that by rotation $\langle\boldsymbol{x},\boldsymbol{L}^i\boldsymbol{w}\rangle$ has the same distribution as x_1, we get

$$\begin{aligned}
\mathbb{E}_{\tau,\boldsymbol{x},\boldsymbol{w}}\Big[\langle\nabla_{\boldsymbol{x}}\overline{\sigma}(\tilde{\tau}\cdot\boldsymbol{x};\boldsymbol{w}),\boldsymbol{x}\rangle^{4m}\Big]&\le\mathbb{E}_{\tau,x_1}\Big[x_1^{4m}\sigma'(\tilde{\tau}\cdot x_1)^{4m}\Big]\\
&\le C\cdot\mathbb{E}_{G\sim\mathsf{N}(0,1)}\Big[\max(G^{4m},1)\exp\Big\{c_1\max(G^2,1)/2\Big\}\Big]\\
&= O_d(1)\,,
\end{aligned} \tag{77}$$

where we used that $c_1<1$.

Combining Eqs. (76) and (77) yields

$$R_1=d^{-1/2}\cdot O_d(1). \tag{78}$$

Step 2. Bound on R_3.

We have

$$\begin{aligned}
R_3=\mathbb{E}_{\boldsymbol{x},\boldsymbol{w}}\Big[A_{[3:\ell]}(\boldsymbol{x};\boldsymbol{w})^{2m}\Big]^{1/(2m)}&\le C_{2m}\mathbb{E}_{\boldsymbol{x}}\Big[\mathbb{E}_{\boldsymbol{w}}\Big[A_{[3:\ell]}(\boldsymbol{x};\boldsymbol{w})^2\Big]^m\Big]^{1/(2m)}\\
&\le C_mC_{2m}\mathbb{E}_{\boldsymbol{x},\boldsymbol{w}}\Big[A_{[3:\ell]}(\boldsymbol{x};\boldsymbol{w})^2\Big]^{1/2}\,,
\end{aligned}$$

where in the first inequality we used hypercontractivity of low-degree polynomials on the sphere with respect to $\boldsymbol{w}$ (Lemma 36), and in the second we used hypercontractivity of low-degree symmetric functions with respect to $\boldsymbol{x}$ (Lemma 6 in Mei et al. (2021)). By Lemma 11, we have

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{w}}\Big[A_{[3:\ell]}(\boldsymbol{x};\boldsymbol{w})^2\Big]=\sum_{k=3}^{\ell}\xi_{d,k}^2B(\mathbb{S}^{d-1};k)\cdot\frac{D(\mathbb{S}^{d-1};k)}{B(\mathbb{S}^{d-1};k)}=O_d(d^{-1}).$$

We deduce that

$$R_3 = d^{-1/2} \cdot O_d(1). \tag{79}$$

Step 3. Bound on R_4.

Similarly to R_3, we have

$$\begin{aligned} R_4 = \mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\left[B_{[3:\ell]}(\boldsymbol{g};\boldsymbol{w})^{2m}\right]^{1/(2m)} &\le C_{2m}\mathbb{E}_{\boldsymbol{w}}\left[\mathbb{E}_{\boldsymbol{g}}\left[B_{[3:\ell]}(\boldsymbol{g};\boldsymbol{w})^{2}\right]^{m}\right]^{1/(2m)} \\ &\le C_m C_{2m}\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\left[B_{[3:\ell]}(\boldsymbol{g};\boldsymbol{w})^{2}\right]^{1/2}, \end{aligned}$$

where in the first inequality we used hypercontractivity of low-degree polynomials with respect to $\boldsymbol{g}$ (Lemma 32), and in the second we used hypercontractivity of low-degree symmetric functions with respect to $\boldsymbol{w}$.

Following the proof of Proposition 30, by setting $m = 1$ and $\overline{\varphi}_{>2}(\boldsymbol{g};\boldsymbol{w}) = B_{[3:\ell]}(\boldsymbol{g};\boldsymbol{w})$, we have for any $\varepsilon > 0$,

$$R_4 = d^{\varepsilon-1/2} \cdot O_d(1). \tag{80}$$

Step 4. Conclude.

The bound on R_2 is more technical and we defer it to Section G.3. By Lemma 33, we have

$$R_2 = d^{-1/2} \cdot O_d(1). \tag{81}$$

Hence combining the bounds (78), (79), (80) and (81), we obtain for any $\varepsilon > 0$,

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{w}}\left[\overline{\sigma}_{>\ell}(\boldsymbol{x};\boldsymbol{w})^{2m}\right]^{1/(2m)} \le \mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\left[\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})^{2m}\right]^{1/(2m)} + O_d(d^{\varepsilon-1/2})\,.$$

Using Proposition 30 concludes the proof.

G.2. Proof in the Gaussian case

Recall that we defined

$$\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w}) = \frac{1}{d}\sum_{i=0}^{d-1}\varphi_{>\ell}(\langle \boldsymbol{g}, \boldsymbol{L}^i\boldsymbol{w}\rangle), \tag{82}$$

where

$$\varphi_{>\ell}(x) = \sigma(x) - \sum_{k=0}^{\ell}\frac{\mu_k(\sigma)}{k!}\mathrm{He}_k(x). \tag{83}$$

Let us now state and prove the Gaussian version of Proposition 29.

Proposition 30 *Consider fixed integers $m \ge 1$ and $\ell \ge 4m$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function such that $|\sigma(x)| \le c_0 \exp(c_1 x^2/(8m))$ for some constants $c_0 > 0$ and $c_1 < 1$. Let $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_d)$ and $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, then*

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\left[\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})^{2m}\right]^{1/(2m)} = d^{-1/2} \cdot O_d(1). \tag{84}$$

Proof [Proof of Proposition 30] Let us expand $\overline{\varphi}_{>\ell}$ as in Eq. (82)

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[d^{2m}\cdot\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})^{2m}\Big]=\sum_{0\le i_1,\dots,i_{2m}\le d-1}\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\prod_{k\in[2m]}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^{i_k}\boldsymbol{w}\rangle)\Big].$$

Let us consider the event

$$\mathcal{A}_\varepsilon:=\Big\{\boldsymbol{w}\in\mathbb{S}^{d-1}(1);\ \sup_{k\in[d-1]}|\langle\boldsymbol{w},\boldsymbol{L}^k\boldsymbol{w}\rangle|\le Cd^{\varepsilon-1/2}\Big\},$$

and for each set of indices $\mathcal{I}=\{i_1,\dots,i_{2m}\}$, consider separately the expectation over $\mathcal{A}_\varepsilon$ and $\mathcal{A}_\varepsilon^c$:

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)\Big]=A+B,$$

where

$$A:=\mathbb{E}_{\boldsymbol{w}}\Big[\mathbf{1}_{\mathcal{A}_\varepsilon}\mathbb{E}_{\boldsymbol{g}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)\Big]\Big],$$
$$B:=\mathbb{E}_{\boldsymbol{w}}\Big[\mathbf{1}_{\mathcal{A}_\varepsilon^c}\mathbb{E}_{\boldsymbol{g}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)\Big]\Big].$$

By Cauchy-Schwarz and Jensen's inequality, we have

$$B\le\mathbb{P}(\mathcal{A}_\varepsilon^c)^{1/2}\cdot\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)^2\Big]^{1/2},$$

with

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)^2\Big]^{1/2}&\le\prod_{i\in\mathcal{I}}\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)^{4m}\Big]^{1/(4m)}\\&=\mathbb{E}_G\Big[\varphi_{>\ell}(G)^{4m}\Big]^{1/2}=O_d(1)\,,\end{aligned}$$

where we used Hölder's inequality and that $\varphi_{>\ell}$ is the sum of a degree ℓ polynomial and σ with $|\sigma(x)|\le c_0\exp(c_1x^2/(8m))$, with constants $c_0>0$ and $c_1<1$. Combining these bounds and Lemma 34, we deduce there exists a constant C independent of d and $\mathcal{I}$ such that

$$B\le C\exp(-cd^{2\varepsilon}).\tag{85}$$

Similarly, by Hölder's inequality, we have the following first bound on A:

$$A\le\prod_{i\in\mathcal{I}}\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)^{2m}\Big]^{1/(2m)}=\mathbb{E}_G\Big[\varphi_{>\ell}(G)^{2m}\Big]\le C\,.\tag{86}$$

Fix $\boldsymbol{w}\in\mathcal{A}_\varepsilon$. Denote $\mathcal{I}_0$ the set of distinct indices in $\mathcal{I}$ and $p=|\mathcal{I}_0|\le 2m$. Denote for each $i\in\mathcal{I}_0$, r_i the multiciplity of i in $\mathcal{I}$, and $g_i=\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle$. We have $\sup_{i\neq j}|\mathbb{E}[g_ig_j]|\le\sup_{k\in[d-1]}|\langle\boldsymbol{w},\boldsymbol{L}^k\boldsymbol{w}\rangle|\le Cd^{\varepsilon-1/2}$ and $\mathbb{E}[g_i^2]=1$. Hence, if there exists $i\in\mathcal{I}$ that appears only once, we have by taking $\psi(x)=\varphi_{>\ell}(\boldsymbol{x})$ and $q=2m\le\ell/2$ in Lemma 31 stated below

$$\Big|\mathbb{E}_{\boldsymbol{g}}\Big[\prod_{i\in\mathcal{I}}\varphi_{>\ell}(\langle\boldsymbol{g},\boldsymbol{L}^i\boldsymbol{w}\rangle)\Big]\Big|\le C'd^{(2m+1)(\varepsilon-1/2)}\,,$$

where C' is independent of $\boldsymbol{w}$. We deduce that

$$A \le C' d^{(2m+1)(\varepsilon-1/2)} \,. \tag{87}$$

There are at most $m^{2m} d^m$ sets of indices $\mathcal{I}$ with no isolated index. Hence, combining the bounds (85), (86) and (87), we get

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}[d^{2m} \cdot \overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})^{2m}] \le C d^{2m} \exp(-c d^{2\varepsilon}) + C m^{2m} d^m + C' d^{2m} \cdot d^{(2m+1)(\varepsilon-1/2)} \,.$$

Taking $\varepsilon \le 1/(4m+2)$, we get

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}[\overline{\varphi}_{>\ell}(\boldsymbol{g};\boldsymbol{w})^{2m}]^{1/(2m)} = d^{-1/2} \cdot O_d(1) \,,$$

which concludes the proof. ∎

The proof of Proposition 30 relies on the following key lemma:

Lemma 31 *Let $q, p, m \ge 1$ be three integers such that $p \le 2m$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a function such that $|\psi(x)| \le c_0 \exp(c_1 x^2/(4m))$ for some constants $c_0 > 0$ and $c_1 < 1$. Furthermore, for all $k = 0, \dots, 2q$,*

$$\mu_k(\psi) = \mathbb{E}_G[\psi(G)\mathrm{He}_k(G)] = 0,$$

where $G \sim \mathsf{N}(0,1)$, i.e., ψ is orthogonal to all polynomials of degree less or equal to q with respect to the standard normal distribution. Let $\boldsymbol{g} = (g_1, \dots, g_p) \sim \mathsf{N}(0, \boldsymbol{\Sigma})$ with $\Sigma_{11} = \dots = \Sigma_{pp} = 1$ and $\sup_{i \ne j} |\Sigma_{ij}| \le C d^{\varepsilon - 1/2}$. Let $(r_1, \dots, r_p)$ be p integers such that $r_1 + \dots + r_p = 2m$ and there exists k such that $r_k = 1$. Then there exists $C' > 0$ depending only on c_0, c_1, C, q, m such that

$$\Big| \mathbb{E}_{\boldsymbol{g}} \Big[\prod_{k \in [p]} \psi(g_k)^{r_k} \Big] \Big| \le C' d^{(q+1)(\varepsilon-1/2)}. \tag{88}$$

Proof [Proof of Lemma 31] Without loss of generality, let us assume that $r_1 = 1$. Let us rewrite the expectation with respect to $\tilde{\boldsymbol{g}} \sim \mathsf{N}(0, \mathbf{I}_p)$:

$$\mathbb{E}_{\boldsymbol{g}} \Big[\prod_{k \in [p]} \psi(g_k)^{r_k} \Big] = \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}} \mathbb{E}_{\tilde{\boldsymbol{g}}} \Big[\prod_{k \in [p]} \psi(\tilde{g}_k)^{r_k} \cdot \exp\Big\{ \tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}}/2 \Big\} \Big] \,, \tag{89}$$

where we denoted $\boldsymbol{M} = \mathbf{I}_p - \boldsymbol{\Sigma}^{-1}$.

By Taylor expansion around 0 at order $q+1$, there exists $\zeta(\tilde{\boldsymbol{g}})$ between 0 and $\tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}}/2$ such that

$$\exp\Big\{ \tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}}/2 \Big\} = \sum_{s=0}^{q} \frac{1}{2^s s!} (\tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}})^s + \frac{1}{2^{q+1}(q+1)!} \exp\{\zeta(\tilde{\boldsymbol{g}})\} \cdot (\tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}})^{q+1}.$$

Notice that the terms $s = 0, \dots, q$ are polynomials of degree smaller or equal to $2q$ in $\tilde{\boldsymbol{g}}$. By the assumption of orthonormality of ψ to polynomials of degree less or equal to $2q$, we deduce

$$\begin{aligned}
&\Big| \mathbb{E}_{\boldsymbol{g}} \Big[\prod_{k \in [p]} \psi(g_k)^{r_k} \Big] \Big| \\
&= \frac{1}{2^{q+1}(q+1)! \sqrt{\det(\boldsymbol{\Sigma})}} \Big| \mathbb{E}_{\tilde{\boldsymbol{g}}} \Big[\prod_{k \in [p]} \psi(\tilde{g}_k)^{r_k} \cdot \exp\{\zeta(\tilde{\boldsymbol{g}})\} \cdot (\tilde{\boldsymbol{g}}^{\mathsf{T}} \boldsymbol{M} \tilde{\boldsymbol{g}})^{q+1} \Big] \Big| \\
&\le \frac{\|\boldsymbol{M}\|_{\mathrm{op}}^{q+1}}{2^{q+1}(q+1)! \sqrt{\det(\boldsymbol{\Sigma})}} \mathbb{E}_{\tilde{\boldsymbol{g}}} \Big[\prod_{k \in [p]} |\psi(\tilde{g}_k)|^{r_k} \cdot \exp\{\|\boldsymbol{M}\|_{\mathrm{op}} \|\tilde{\boldsymbol{g}}\|_2^2\} \cdot \|\tilde{\boldsymbol{g}}\|_2^{2(q+1)} \Big] \,.
\end{aligned}$$

Furthermore, from the bound $|\psi(x)| \le c_0 \exp(c_1 x^2/(4m))$ and that $r_k \le 2m$, we have

$$\left|\mathbb{E}_{\boldsymbol{g}}\Big[\prod_{k\in[p]} \psi(g_k)^{r_k}\Big]\right| \le \frac{c_0^{2m} p^{2q} \|\boldsymbol{M}\|_{\rm op}^{q+1}}{2^{q+1}(q+1)!\sqrt{\det(\boldsymbol{\Sigma})}} \mathbb{E}_G\Big[G^{2q+2}\exp\{c_1 G^2/2 + \|\boldsymbol{M}\|_{\rm op} G^2\}\Big]^p . \tag{90}$$

From the assumptions on $\boldsymbol{\Sigma}$, we have $\|\boldsymbol{\Sigma} - \mathbf{I}_p\|_{\rm op} \le \|\boldsymbol{\Sigma} - \mathbf{I}_p\|_F \le p \sup_{i\neq j} |\Sigma_{ij}| = O_d(d^{\varepsilon - 1/2})$, and therefore $\|\boldsymbol{M}\|_{\rm op} = O_d(d^{\varepsilon-1/2})$ and $\det(\boldsymbol{\Sigma})^{-1/2} = O_d(1)$.

From the assumption that $c_1 < 1$ and taking d sufficiently large such that $\|\boldsymbol{M}\|_{\rm op} < (1-c_1)/4$, the expectation on the right hand side of Eq. (90) is bounded by a constant. We deduce that

$$\mathbb{E}_{\boldsymbol{g}}\Big[\prod_{k\in[p]} \psi(g_k)^{r_k}\Big] = \|\boldsymbol{M}\|_{\rm op}^{q+1} \cdot O_d(1) = d^{(q+1)(\varepsilon-1/2)} \cdot O_d(1)\,,$$

which concludes the proof. ∎

G.3. Technical lemmas

The first lemma is a straightforward consequence of the proof of Lemma 37 (we include a proof for completeness).

Lemma 32 *For any $\ell \in \mathbb{N}$ and $f \in L^2(\mathbb{R}^d, \gamma_d)$ to be a degree ℓ polynomial on $\mathbb{R}^d$, where $\gamma_d = \mathsf{N}(0, \mathbf{I}_d)$ is the isotropic Gaussian distribution. Then for any $q \ge 2$, we have*

$$\|f\|^2_{L^q(\mathbb{R}^d,\gamma_d)} \le (q-1)^\ell \cdot \|f\|^2_{L^2(\mathbb{R}^d,\gamma_d)}.$$

Proof [Proof of Lemma 32] Let $\boldsymbol{\varepsilon} = (\varepsilon_{i,j})_{i\in[d],j\in[D]} \sim \mathrm{Unif}(\mathscr{Q}^{dD})$ and define for $i = 1,\dots,d$,

$$G_i = \frac{\varepsilon_{i,1} + \ldots + \varepsilon_{i,D}}{\sqrt{D}}.$$

Consider f a degree ℓ polynomial on $\mathbb{R}^d$ and define

$$\tilde{f}(\boldsymbol{\varepsilon}) = f(G_1,\dots,G_d).$$

From hypercontractivity of low degree polynomials on the hypercube (Lemma 35), we have

$$\|\tilde{f}\|^2_{L^q(\mathscr{Q}^{d^2})} \le (q-1)^\ell \cdot \|\tilde{f}\|^2_{L^2(\mathscr{Q}^{d^2})}. \tag{91}$$

Furthermore, by the multivariate central limit theorem, as $D \to \infty$ for d fixed, $(G_1,\dots,G_d)$ converges in distribution to $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_d)$. By dominated convergence theorem, we have $\|\tilde{f}\|^2_{L^q(\mathscr{Q}^{dD})} \to \|f\|^2_{L^q(\mathbb{R}^d,\gamma_d)}$ and $\|\tilde{f}\|^2_{L^2(\mathscr{Q}^{dD})} \to \|f\|^2_{L^2(\mathbb{R}^d,\gamma_d)}$, and taking the limit in inequality (91) yields the result. ∎

Lemma 33 *Follow the notations in Section G.1. We have*

$$\mathbb{E}_{\boldsymbol{g},\boldsymbol{w}}\Big[\big(A_{\le 2}(\sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2; \boldsymbol{w}) - B_{\le 2}(\boldsymbol{g}; \boldsymbol{w})\big)^{2m}\Big]^{1/(2m)} = O_d(d^{-1/2}).$$

Proof [Proof of Lemma 33] Denote $\tau = \|\boldsymbol{g}\|_2/\sqrt{d}$ and $\boldsymbol{x} = \sqrt{d}\boldsymbol{g}/\|\boldsymbol{g}\|_2$. Recall that we defined

$$A_{\le 2}(\boldsymbol{x}, \boldsymbol{w}) = \xi_{d,0} + \xi_{d,1} B(\mathbb{S}^{d-1}; 1)\overline{Q}_1(\boldsymbol{x}; \boldsymbol{w}) + \xi_{d,2} B(\mathbb{S}^{d-1}; 2)\overline{Q}_2(\boldsymbol{x}; \boldsymbol{w})\,,$$
$$B_{\le 2}(\tau \cdot \boldsymbol{x}, \boldsymbol{w}) = \mu_0(\sigma) + \mu_1(\sigma)\overline{\mathrm{He}}_1(\tau \cdot \boldsymbol{x}; \boldsymbol{w}) + \frac{\mu_2(\sigma)}{2}\overline{\mathrm{He}}_2(\tau \cdot \boldsymbol{x}; \boldsymbol{w})\,,$$

Let us bound the difference of each term separately.

Step 1. Bound 0th order term.

Following the same argument as in the bound of R_1 in Section G.1, we have

$$\begin{aligned} c_0 := |\mu_0(\sigma) - \xi_{d,0}| &= \left|\mathbb{E}_{\tau, x_1}\left[\sigma(\tau \cdot x_1) - \sigma(x_1)\right]\right| \\ &\le \mathbb{E}_\tau\left[(\tau - 1)^2\right]^{1/2} \mathbb{E}_{\tau, x_1}\left[x_1^2 \sigma'(\tilde{\tau} \cdot x_1)^2\right]^{1/2} = O_d(d^{-1/2}). \end{aligned} \tag{92}$$

Step 2. Bound 1st order term.

We have $\mathrm{He}_1(x) = x$ and $B(\mathbb{S}^{d-1}; 1)^{1/2} \cdot Q_1(\sqrt{d}x) = x$. Hence,

$$\begin{aligned} c_1 &:= \mathbb{E}_{\boldsymbol{g}, \boldsymbol{w}}\left[\left(\mu_1(\sigma)\overline{\mathrm{He}}_1(\tau \cdot \boldsymbol{x}; \boldsymbol{w}) - \xi_{d,1} B(\mathbb{S}^{d-1}; 1)\overline{Q}_1(\boldsymbol{x}; \boldsymbol{w})\right)^{2m}\right] \\ &= \mathbb{E}_\tau\left[\left(\tau \cdot \mu_1(\sigma) - \xi_{d,1} B(\mathbb{S}^{d-1}; 1)^{1/2}\right)^{2m}\right] \cdot \mathbb{E}_{\boldsymbol{x}, \boldsymbol{w}}\left[B(\mathbb{S}^{d-1}; 1)^m \overline{Q}_1(\boldsymbol{x}; \boldsymbol{w})^{2m}\right]. \end{aligned}$$

Using the convergence of Gegenbauer coefficients to Hermite coefficients (see Eq. (110) in Section H.1.3), there exists a constant $C > 0$ such that

$$\mathbb{E}_\tau\left[\left(\tau \cdot \mu_1(\sigma) - \xi_{d,1} B(\mathbb{S}^{d-1}; 1)^{1/2}\right)^{2m}\right]^{1/(2m)} \le C\left[\mathbb{E}_\tau\left[\tau^{2m}\right]^{1/(2m)} + 1\right] = O_d(1)\,, \tag{93}$$

where we used for example that low-degree polynomials of τ^2 are hypercontractive (see the bound on R_1 in Section G.1). From the same argument as in the bound of R_3 in Section G.1, we have

$$\mathbb{E}_{\boldsymbol{x}, \boldsymbol{w}}\left[B(\mathbb{S}^{d-1}; 1)^m \overline{Q}_1(\boldsymbol{x}; \boldsymbol{w})^{2m}\right]^{1/(2m)} \le C_{2m} \frac{D(\mathbb{S}^{d-1}; 1)^{1/2}}{B(\mathbb{S}^{d-1}; 1)^{1/2}} = O_d(d^{-1/2}). \tag{94}$$

Combining Eqs. (93) and (94) yields

$$c_1 = O_d(d^{-1/2}). \tag{95}$$

Step 3. Bound 2nd order term.

We have $\mathrm{He}_2(x) = x^2 - 1$ and $B(\mathbb{S}^{d-1}; 2)^{1/2} \cdot Q_2(\sqrt{d}x) = a_{2,d} \cdot (x^2 - 1)$ with $a_{2,d} = \Theta_d(1)$. We can rewrite

$$\mathrm{He}_2(\tau \cdot x_1) = \tau^2 B(\mathbb{S}^{d-1}; 2)^{1/2} \cdot Q_2(\sqrt{d}x)/a_{2,d} + \tau^2 - 1.$$

Hence, by triangle inequality,

$$\begin{aligned} c_1 &:= \mathbb{E}_{\boldsymbol{g}, \boldsymbol{w}}\left[\left(\mu_2(\sigma)\overline{\mathrm{He}}_2(\tau \cdot \boldsymbol{x}; \boldsymbol{w})/2 - \xi_{d,2} B(\mathbb{S}^{d-1}; 2)\overline{Q}_2(\boldsymbol{x}; \boldsymbol{w})\right)^{2m}\right]^{1/(2m)} \\ &\le \mathbb{E}_\tau\left[\left(\tau \cdot \mu_2(\sigma)/(2a_{2,d}) - \xi_{d,2} B(\mathbb{S}^{d-1}; 2)^{1/2}\right)^{2m}\right]^{1/(2m)} \cdot \mathbb{E}_{\boldsymbol{x}, \boldsymbol{w}}\left[B(\mathbb{S}^{d-1}; 2)^m \overline{Q}_2(\boldsymbol{x}; \boldsymbol{w})^{2m}\right]^{1/(2m)} \\ &\quad + \frac{|\mu_2(\sigma)|}{2} \mathbb{E}_\tau\left[(\tau^2 - 1)^{2m}\right]^{1/(2m)}. \end{aligned}$$

The first term is bounded as the 1-st order term while the second term is bounded as the 0-th order term. Combining the two yields

$$c_2 = O_d(d^{-1/2}). \tag{96}$$

Step 4. Conclude.

Combining the bounds (92), (95) and (96), we get by triangle inequality

$$\mathbb{E}_{\tau,\boldsymbol{x},\boldsymbol{w}}\Big[\big(A_{\leq 2}(\boldsymbol{x};\boldsymbol{w}) - B_{\leq 2}(\tau\cdot\boldsymbol{x};\boldsymbol{w})\big)^{2m}\Big]^{1/(2m)} \leq c_0 + c_1 + c_2 = O_d(d^{-1/2})\,,$$

which concludes the proof. ■

Lemma 34 *Let $\varepsilon > 0$ and $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$. Then there exists $C, c > 0$ such that*

$$\mathbb{P}(\mathcal{A}_\varepsilon^c) \leq C\exp(-cd^{2\varepsilon}).$$

Proof [Proof of Lemma 34] Let us use the correspondence between uniform distribution and Gaussian distribution: $\boldsymbol{w} \sim \boldsymbol{z}/\|\boldsymbol{z}\|_2$, where $\boldsymbol{z} \sim \mathsf{N}(0, \mathbf{I}_d)$. We have for $k = 1, \dots, d-1$,

$$\mathbb{P}(|\langle \boldsymbol{w}, \boldsymbol{L}^k\boldsymbol{w}\rangle| \geq t) = \mathbb{P}(|\langle \boldsymbol{z}, \boldsymbol{L}^k\boldsymbol{z}\rangle/\|\boldsymbol{z}\|_2^2| \geq t) \leq \mathbb{P}(|\langle \boldsymbol{z}, \boldsymbol{L}^k\boldsymbol{z}\rangle/d| \geq t/2) + \mathbb{P}(\|\boldsymbol{z}\|_2^2 \leq d/2).$$

Note that for any $k \in [d-1]$, we have $\|\boldsymbol{L}^k\|_F \leq \sqrt{d}$ and $\|\boldsymbol{L}^k\|_{\mathrm{op}} \leq 1$. By the Hanson-Wright inequality, for any $k \neq 0$, we have

$$\mathbb{P}\Big(\big|\langle \boldsymbol{z}, \boldsymbol{L}^k\boldsymbol{z}\rangle/d\big| > t\Big) \leq 2\exp\{-cd\cdot\min(t^2, t)\}.$$

Furthermore, by standard concentration of the norm of Gaussian vectors, we have

$$\mathbb{P}(\|\boldsymbol{z}\|_2^2 \leq d/2) \leq C\exp(-cd).$$

Taking $t = Cd^{\varepsilon-1/2}$ and combining the above two bounds, we get

$$\mathbb{P}(|\langle \boldsymbol{w}, \boldsymbol{L}^k\boldsymbol{w}\rangle| \geq t) \leq 2\exp(-cd^{2\varepsilon}) + C\exp(-cd).$$

Taking the union bounds over $k \in [d-1]$ concludes the proof. ■

Appendix H. Technical background of function spaces

H.1. Functions on the sphere

H.1.1. FUNCTIONAL SPACES OVER THE SPHERE

For $d \geq 3$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by τ_d the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in this section are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d)$, with scalar product and norm denoted as $\langle \cdot\,, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$:

$$\langle f, g\rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\, g(\boldsymbol{x})\, \tau_d(\mathrm{d}\boldsymbol{x})\,. \tag{97}$$

For $\ell \in \mathbb{Z}_{\geq 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e. homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{98}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(\mathbb{S}^{d-1}; \ell) = \frac{2\ell + d - 2}{d-2}\binom{\ell + d - 3}{\ell}\,. \tag{99}$$

For each $\ell \in \mathbb{Z}_{\geq 0}$, the spherical harmonics $\{Y_{\ell,j}^{(d)}\}_{1 \leq j \leq B(\mathbb{S}^{d-1};\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)}\rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by $\overline{\mathsf{P}}_k$ the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d)$. This can be written in terms of spherical harmonics as

$$\overline{\mathsf{P}}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(\mathbb{S}^{d-1};k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{100}$$

We also define $\overline{\mathsf{P}}_{\leq \ell} \equiv \sum_{k=0}^{\ell} \overline{\mathsf{P}}_k$, $\overline{\mathsf{P}}_{>\ell} \equiv \mathbf{I} - \overline{\mathsf{P}}_{\leq \ell} = \sum_{k=\ell+1}^{\infty} \overline{\mathsf{P}}_k$, and $\overline{\mathsf{P}}_{<\ell} \equiv \overline{\mathsf{P}}_{\leq \ell-1}$, $\overline{\mathsf{P}}_{\geq \ell} \equiv \overline{\mathsf{P}}_{>\ell-1}$.

H.1.2. GEGENBAUER POLYNOMIALS

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell \geq 0}$ forms

an orthogonal basis on $L^2([-d,d],\tilde{\tau}_d^1)$, where $\tilde{\tau}_d^1$ is the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle$ when $\boldsymbol{x} \sim \tau_d$, satisfying the normalization condition:

$$\langle Q_k^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle), Q_j^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = \frac{1}{B(\mathbb{S}^{d-1};k)}\,\delta_{jk}\,. \tag{101}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript (d) in $Q_\ell^{(d)}$ when clear from the context.

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot\rangle)$.

We will use the following properties of Gegenbauer polynomials

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(\mathbb{S}^{d-1};k)}\delta_{jk}Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{102}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(\mathbb{S}^{d-1};k)}\sum_{i=1}^{B(\mathbb{S}^{d-1};k)} Y_{ki}^{(d)}(\boldsymbol{x})Y_{ki}^{(d)}(\boldsymbol{y}). \tag{103}$$

These properties imply that —up to a constant— $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree -k spherical harmonics

$$(\overline{\mathsf{P}}_k f)(\boldsymbol{x}) = B(\mathbb{S}^{d-1};k)\int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \tau_d(\mathrm{d}\boldsymbol{y})\,. \tag{104}$$

For a function $\sigma \in L^2([-\sqrt{d},\sqrt{d}],\tau_d^1)$ (where τ_d^1 is the distribution of $\langle \boldsymbol{e}_1, \boldsymbol{x}\rangle$ when $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\xi_{d,k}(\sigma)$ to be

$$\xi_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x)Q_k^{(d)}(\sqrt{d}x)\tau_d^1(\mathrm{d}x), \tag{105}$$

then we have the following equation holds in $L^2([-\sqrt{d},\sqrt{d}],\tau_d^1)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \xi_{d,k}(\sigma)B(\mathbb{S}^{d-1};k)Q_k^{(d)}(\sqrt{d}x).$$

For any rotationally invariant kernel $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$, with $h_d(\sqrt{d}\,\cdot) \in L^2([-\sqrt{d},\sqrt{d}],\tau_d^1)$, we can associate a self adjoint operator $\mathscr{H}_d : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ via

$$\mathscr{H}_d f(\boldsymbol{x}) \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} h_d(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d)\, f(\boldsymbol{x}_1)\, \tau_d(\mathrm{d}\boldsymbol{x}_1)\,. \tag{106}$$

By rotational invariance, the space V_k of homogeneous polynomials of degree k is an eigenspace of $\mathscr{H}_d$, and we will denote the corresponding eigenvalue by $\xi_{d,k}(h_d)$. In other words $\mathscr{H}_d f(\boldsymbol{x}) \equiv \sum_{k=0}^{\infty} \xi_{d,k}(h_d)\overline{\mathsf{P}}_k f$. The eigenvalues can be computed via

$$\xi_{d,k}(h_d) = \int_{[-\sqrt{d},\sqrt{d}]} h_d\big(x/\sqrt{d}\big) Q_k^{(d)}(\sqrt{d}x)\tau_d^1(\mathrm{d}x)\,. \tag{107}$$

H.1.3. HERMITE POLYNOMIALS

The Hermite polynomials $\{\mathrm{He}_k\}_{k\ge 0}$ form an orthogonal basis of $L^2(\mathbb{R},\gamma)$, where $\gamma(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\,\mathrm{He}_k(G)\big\} = k!\,\delta_{jk}\,. \tag{108}$$

As a consequence, for any function $g \in L^2(\mathbb{R},\gamma)$, we have the decomposition

$$g(x) = \sum_{k=0}^{\infty} \frac{\mu_k(g)}{k!}\,\mathrm{He}_k(x)\,, \qquad \mu_k(g) \equiv \mathbb{E}\big\{g(G)\,\mathrm{He}_k(G)\big\}\,. \tag{109}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k\ge 0}$ with respect to the measure $\tilde{\tau}_d^1$, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to γ. Since $\tilde{\tau}_d^1 \Rightarrow \gamma$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\,B(\mathbb{S}^{d-1};k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}\,. \tag{110}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty} \xi_{d,k}(\sigma)(B(\mathbb{S}^{d-1};k)k!)^{1/2}, \tag{111}$$

where $\mu_k(\sigma)$ and $\xi_{d,k}(\sigma)$ are given in Eq. (109) and (105).

H.2. Functions on the hypercube

Fourier analysis on the hypercube is a well studied subject O'Donnell (2014). The purpose of this section is to introduce some notations that make the correspondence with proofs on the sphere straightforward. For convenience, we will adopt the same notations as for their spherical case.

H.2.1. FOURIER BASIS

Denote $\mathscr{Q}^d = \{-1,+1\}^d$ the hypercube in d dimension. Let us denote τ_d to be the uniform probability measure on $\mathscr{Q}^d$. All the functions will be assumed to be elements of $L^2(\mathscr{Q}^d,\tau_d)$ (which

contains all the bounded functions $f : \mathscr{Q}^d \to \mathbb{R}$), with scalar product and norm denoted as $\langle \cdot, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$:

$$\langle f, g \rangle_{L^2} \equiv \int_{\mathscr{Q}^d} f(\boldsymbol{x}) g(\boldsymbol{x}) \tau_d(\mathrm{d}\boldsymbol{x}) = \frac{1}{2^n} \sum_{\boldsymbol{x} \in \mathscr{Q}^d} f(\boldsymbol{x}) g(\boldsymbol{x}).$$

Notice that $L^2(\mathscr{Q}^d, \tau_d)$ is a 2^n dimensional linear space. By analogy with the spherical case we decompose $L^2(\mathscr{Q}^d, \tau_d)$ as a direct sum of $d+1$ linear spaces obtained from polynomials of degree $\ell = 0, \dots, d$

$$L^2(\mathscr{Q}^d, \tau_d) = \bigoplus_{\ell=0}^{d} V_{d,\ell}.$$

For each $\ell \in \{0, \dots, d\}$, consider the Fourier basis $\{Y_{\ell,S}^{(d)}\}_{S \subseteq [d], |S|=\ell}$ of degree ℓ, where for a set $S \subseteq [d]$, the basis is given by

$$Y_{\ell,S}^{(d)}(\boldsymbol{x}) \equiv x^S \equiv \prod_{i \in S} x_i.$$

It is easy to verify that (notice that $x_i^k = x_i$ if k is odd and $x_i^k = 1$ if k is even)

$$\langle Y_{\ell,S}^{(d)}, Y_{k,S'}^{(d)} \rangle_{L^2} = \mathbb{E}[x^S \times x^{S'}] = \delta_{\ell,k} \delta_{S,S'}.$$

Hence $\{Y_{\ell,S}^{(d)}\}_{S \subseteq [d], |S|=\ell}$ form an orthonormal basis of $V_{d,\ell}$ and

$$\dim(V_{d,\ell}) = B(\mathscr{Q}^d; \ell) = \binom{d}{\ell}.$$

As above, we will omit the superscript (d) in $Y_{\ell,S}^{(d)}$ when clear from the context.

H.2.2. HYPERCUBIC GEGENBAUER

We consider the following family of polynomials $\{Q_\ell^{(d)}\}_{\ell=0,\dots,d}$ that we will call hypercubic Gegenbauer, defined as

$$Q_\ell^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) = \frac{1}{B(\mathscr{Q}^d; \ell)} \sum_{S \subseteq [d], |S|=\ell} Y_{\ell,S}^{(d)}(\boldsymbol{x}) Y_{\ell,S}^{(d)}(\boldsymbol{y}).$$

Notice that the right hand side only depends on $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and therefore these polynomials are uniquely defined. In particular,

$$\langle Q_\ell^{(d)}(\langle \mathbf{1}, \cdot \rangle), Q_k^{(d)}(\langle \mathbf{1}, \cdot \rangle) \rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)} \delta_{\ell k}.$$

Hence $\{Q_\ell^{(d)}\}_{\ell=0,\dots,d}$ form an orthogonal basis of $L^2(\{-d, -d+2, \dots, d-2, d\}, \tilde{\tau}_d^1)$ where $\tilde{\tau}_d^1$ is the distribution of $\langle \mathbf{1}, \boldsymbol{x} \rangle$ when $\boldsymbol{x} \sim \tau_d$, i.e., $\tilde{\tau}_d^1 \sim 2\mathrm{Bin}(d, 1/2) - d/2$.

We have

$$\langle Q_\ell^{(d)}(\langle \boldsymbol{x}, \cdot \rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot \rangle) \rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)} Q_k(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) \delta_{\ell k}.$$

For a function $\sigma(\cdot/\sqrt{d}) \in L^2(\{-d, -d+2, \dots, d-2, d\}, \tilde{\tau}_d^1)$, denote its hypercubic Gegenbauer coefficients $\xi_{d,k}(\sigma)$ to be

$$\xi_{d,k}(\sigma) = \int_{\{-d,-d+2,\dots,d-2,d\}} \sigma(x/\sqrt{d}) Q_k^{(d)}(x) \tilde{\tau}_d^1(\mathrm{d}x).$$

Notice that by weak convergence of $\langle \mathbf{1}, \boldsymbol{x} \rangle / \sqrt{d}$ to the normal distribution, we have also convergence of the (rescaled) hypercubic Gegenbauer polynomials to the Hermite polynomials, i.e., for any fixed k, we have

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(\mathscr{Q}^d; k)^{1/2}\} = \mathrm{Coeff}\left\{ \frac{1}{(k!)^{1/2}} \mathrm{He}_k(x) \right\}. \tag{112}$$

H.3. Hypercontractivity of Gaussian measure and uniform distributions on the sphere and the hypercube

By Holder's inequality, we have $\|f\|_{L^p} \le \|f\|_{L^q}$ for any f and any $p \le q$. The reverse inequality does not hold in general, even up to a constant. However, for some measures, the reverse inequality will hold for some sufficiently nice functions. These measures satisfy the celebrated hypercontractivity properties Gross (1975); Bonami (1970); Beckner (1975, 1992).

Lemma 35 (Hypercube hypercontractivity Beckner (1975)) *For any $\ell = \{0, \dots, d\}$ and $f_d \in L^2(\mathscr{Q}^d)$ to be a degree ℓ polynomial, then for any integer $q \ge 2$, we have*

$$\|f_d\|_{L^q(\mathscr{Q}^d)}^2 \le (q-1)^\ell \cdot \|f_d\|_{L^2(\mathscr{Q}^d)}^2.$$

Lemma 36 (Spherical hypercontractivity Beckner (1992)) *For any $\ell \in \mathbb{N}$ and $f_d \in L^2(\mathbb{S}^{d-1})$ to be a degree ℓ polynomial, for any $q \ge 2$, we have*

$$\|f_d\|_{L^q(\mathbb{S}^{d-1})}^2 \le (q-1)^\ell \cdot \|f_d\|_{L^2(\mathbb{S}^{d-1})}^2.$$

Lemma 37 (Gaussian hypercontractivity) *For any $\ell \in \mathbb{N}$ and $f \in L^2(\mathbb{R}, \gamma)$ to be a degree ℓ polynomial on $\mathbb{R}$, where γ is the standard Gaussian distribution. Then for any $q \ge 2$, we have*

$$\|f\|_{L^q(\mathbb{R},\gamma)}^2 \le (q-1)^\ell \cdot \|f\|_{L^2(\mathbb{R},\gamma)}^2.$$

The Gaussian hypercontractivity is a direct consequence of hypercube hypercontractivity.