Journal of Statistical Mechanics:
Theory and Experiment

PAPER

# When do neural networks outperform kernel methods?[*]

To cite this article: Behrooz Ghorbani *et al J. Stat. Mech.* (2021) 124009

View the article online for updates and enhancements.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: ML 2021

# When do neural networks outperform kernel methods?*

**Behrooz Ghorbani**[1], **Song Mei**[2], **Theodor Misiakiewicz**[3,**] **and Andrea Montanari**[1,3,4]

[1] Department of Electrical Engineering, Stanford University, United States of America
[2] Department of Statistics, University of California, Berkeley, United States of America
[3] Department of Statistics, Stanford University, United States of America
[4] Google Research, Brain Team, United States of America
E-mail: ghorbani@google.com, songmei@berkeley.edu, misiakie@stanford.edu and montanari@stanford.edu

Received 10 November 2021
Accepted for publication 14 November 2021
Published 29 December 2021

Online at stacks.iop.org/JSTAT/2021/124009
https://doi.org/10.1088/1742-5468/ac3a81

J. Stat. Mech. (2021) 124009

**Abstract.** For a certain scaling of the initialization of stochastic gradient descent (SGD), wide neural networks (NN) have been shown to be well approximated by reproducing kernel Hilbert space (RKHS) methods. Recent empirical work showed that, for some classification tasks, RKHS methods can replace NNs without a large loss in performance. On the other hand, two-layers NNs are known to encode richer smoothness classes than RKHS and we know of special examples for which SGD-trained NN provably outperform RKHS. This is true even in the wide network limit, for a different scaling of the initialization. How can we reconcile the above claims? For which tasks do NNs outperform RKHS? If covariates are nearly isotropic, RKHS methods suffer from the curse of dimensionality, while NNs can overcome it by learning the best low-dimensional representation. Here we show that this curse of dimensionality becomes milder if the covariates display the same low-dimensional structure as the target function, and we precisely characterize this tradeoff. Building on these results, we present the spiked covariates

[*] This article is an updated version of: Ghorbani B, Mei S, Misiakiewicz T and Montanari A 2020 When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems* vol 33 ed H Larochelle, M Ranzato, R Hadsell, M F Balcan and H Lin (New York: Curran Associates) pp 14820–30.
[**] Author to whom any correspondence should be addressed.

1742-5468/21/124009+110$33.00

model that can capture in a unified framework both behaviors observed in earlier work. We hypothesize that such a latent low-dimensional structure is present in image classification. We test numerically this hypothesis by showing that specific perturbations of the training distribution degrade the performances of RKHS methods much more significantly than NNs.

**Keywords:** deep learning, machine learning, learning theory

## Contents

## 1. Introduction

In supervised learning we are given data $\{(y_i, \boldsymbol{x}_i)\}_{i\leqslant n} \sim_{iid} \mathbb{P} \in \mathscr{P}(\mathbb{R}\times\mathbb{R}^d)$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ a covariate vector and $y_i \in \mathbb{R}$ the corresponding label, and would like to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ to predict future labels. In many applications, state-of-the-art systems use multi-layer neural networks (NN). The simplest such model is provided by two-layers fully-connected networks:

$$\mathcal{F}_{\mathrm{NN}}^{N} := \left\{ \hat{f}_{\mathrm{NN}}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}) = \sum_{i=1}^{N} b_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : b_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d, \quad \forall\, i \in [N] \right\}. \tag{1}$$

$\mathcal{F}_{\mathrm{NN}}^{N}$ is a non-linearly parametrized class of functions: while nonlinearity poses a challenge to theoreticians, it is often claimed to be crucial in order to learn rich representation of the data. Recent efforts to understand NN have put the spotlight on two linearizations of $\mathcal{F}_{\mathrm{NN}}^{N}$, the random features [1] and the neural tangent [2] classes

$$\mathcal{F}_{\mathrm{RF}}^{N}(\boldsymbol{W}) := \left\{ \hat{f}_{\mathrm{RF}}(\boldsymbol{x}; \boldsymbol{a}; \boldsymbol{W}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : a_i \in \mathbb{R}, \quad \forall\, i \in [N] \right\}, \tag{2}$$

$$\mathcal{F}_{\mathrm{NT}}^{N}(\boldsymbol{W}) := \left\{ \hat{f}_{\mathrm{NT}}(\boldsymbol{x}; \boldsymbol{S}, \boldsymbol{W}) = \sum_{i=1}^{N} \langle \boldsymbol{s}_i, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : \boldsymbol{s}_i \in \mathbb{R}^d, \quad \forall\, i \in [N] \right\}. \tag{3}$$

$\mathcal{F}_{\mathrm{RF}}^{N}(\boldsymbol{W})$ and $\mathcal{F}_{\mathrm{NT}}^{N}(\boldsymbol{W})$ are linear classes of functions, depending on the realization of the input-layer weights $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\leqslant N}$ (which are chosen randomly). The relation between NN and these two linear classes is given by the first-order Taylor expansion: $\hat{f}_{\mathrm{NN}}(\boldsymbol{x}; \boldsymbol{b} + \varepsilon\boldsymbol{a}, \boldsymbol{W} + \varepsilon\boldsymbol{S}) - \hat{f}_{\mathrm{NN}}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}) = \varepsilon\hat{f}_{\mathrm{RF}}(\boldsymbol{x}; \boldsymbol{a}; \boldsymbol{W}) + \varepsilon\hat{f}_{\mathrm{NT}}(\boldsymbol{x}; \boldsymbol{S}(\boldsymbol{b}); \boldsymbol{W}) + O(\varepsilon^2)$, where $\boldsymbol{S}(\boldsymbol{b}) = (b_i \boldsymbol{s}_i)_{i\leqslant N}$. A number of recent papers show that, if weights and stochastic gradient descent (SGD) updates are suitably scaled, and the network is sufficiently wide ($N$ sufficiently large), then SGD converges to a function $\hat{f}_{\mathrm{NN}}$ that is approximately in $\mathcal{F}_{\mathrm{RF}}^{N}(\boldsymbol{W}) + \mathcal{F}_{\mathrm{NT}}^{N}(\boldsymbol{W})$, with $\boldsymbol{W}$ determined by the SGD initialization [2–7]. This was termed the ‘lazy regime’ in [8].

Does this linear theory convincingly explain the successes of NN? Can the performances of NN be achieved by the simpler NT or RF models? Is there any fundamental difference between the two classes RF and NT? If the weights $(\boldsymbol{w}_i)_{i\leqslant N}$ are i.i.d. draws from a distribution $\nu$ on $\mathbb{R}^d$, the spaces $\mathcal{F}_{\mathrm{RF}}^{N}(\boldsymbol{W})$, $\mathcal{F}_{\mathrm{NT}}^{N}(\boldsymbol{W})$ can be thought as finite-dimensional approximations of a certain reproducing kernel Hilbert space (RKHS):

$$\mathcal{H}(h) := \mathrm{cl}\left(\left\{f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i h(\boldsymbol{x}, \boldsymbol{x}_i) : c_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^d, N \in \mathbb{N}\right\}\right), \tag{4}$$

where $\mathrm{cl}(\cdot)$ denotes closure. From this point of view, RF and NT differ in that they correspond to slightly different choices of the kernel: $h_{\mathrm{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \int \sigma(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle)\nu(\mathrm{d}\boldsymbol{w})$ versus $h_{\mathrm{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \int \sigma'(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_1)\sigma'(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_2)\nu(\mathrm{d}\boldsymbol{w})$. Multi-layer fully-connected NNs in the lazy regime can be viewed as randomized approximations to RKHS as well, with some changes in the kernel $h$. This motivates analogous questions for $\mathcal{H}(h)$: can the performances of NN be achieved by RKHS methods?

Recent work addressed the separation between NN and RKHS from several points of view, without providing a unified answer. Some empirical studies on various datasets showed that networks can be replaced by suitable kernels with limited drop in performances [9–16]. At least two studies reported a larger gap for convolutional networks and the corresponding kernels [17, 18]. On the other hand, theoretical analysis provided a number of separation examples, i.e. target functions $f_*$ that can be represented and possibly efficiently learnt using NN, but not in the corresponding RKHS [19–24]. For instance, if the target is a single neuron $f_*(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}_*, \boldsymbol{x}\rangle)$, then training a neural network with one hidden neuron learns the target efficiently from approximately $d \log d$ samples [25], while the corresponding RKHS has test error bounded away from zero for every sample size polynomial in $d$ [19, 21]. Further even in the infinite width limit, it is known that two-layers NN can actually capture a richer class of functions than the associated RKHS, provided SGD training is scaled differently from the lazy regime [26–30].

Can we reconcile empirical and theoretical results?

### 1.1. Overview

In this paper we introduce a stylized scenario—which we will refer to as the spiked covariates model—that can explain the above seemingly divergent observations in a unified framework. The spiked covariates model is based on two building blocks: (1) target functions depending on low-dimensional projections; (2) approximately low-dimensional covariates.

(a) *Target functions depending on low-dimensional projections.* We investigate the hypothesis that NNs are more efficient at learning target functions that depend on low-dimensional projections of the data (the signal covariates). Formally, we consider target functions $f_* : \mathbb{R}^d \to \mathbb{R}$ of the form $f_*(\boldsymbol{x}) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x})$, where $\boldsymbol{U} \in \mathbb{R}^{d\times d_0}$ is a semi-orthogonal matrix, $d_0 \ll d$, and $\varphi : \mathbb{R}^{d_0} \to \mathbb{R}$ is a suitably smooth function. This model captures an important property of certain applications. For instance, the labels in an image classification problem do not depend equally on the whole Fourier spectrum of the image, but predominantly on the low-frequency components.

As for the example of a single neuron $f_*(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}_*, \boldsymbol{x}\rangle)$, we expect RKHS to suffer from a curse of dimensionality in learning functions of low-dimensional projections. Indeed, this is well understood in low dimension or for isotropic covariates [20, 21].

(b) *Approximately low-dimensional covariates*. RKHS behave well on certain image classification tasks [10, 12, 17], and this seems to contradict the previous point. However, the example of image classification naturally brings up another important property of real data that helps to clarify this puzzle. Not only we expect the target function $f_*(\boldsymbol{x})$ to depend predominantly on the low-frequency components of image $\boldsymbol{x}$, but the image $\boldsymbol{x}$ itself to have most of its spectrum concentrated on low-frequency components (linear denoising algorithms exploit this very observation).

More specifically, we consider the case in which $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z}_1 + \boldsymbol{U}^{\perp}\boldsymbol{z}_2$, where $\boldsymbol{U} \in \mathbb{R}^{d\times d_0}$, $\boldsymbol{U}^{\perp} \in \mathbb{R}^{d\times(d-d_0)}$, and $[\boldsymbol{U}|\boldsymbol{U}^{\perp}] \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Moreover, we assume $\boldsymbol{z}_1 \sim \mathrm{Unif}(\mathbb{S}^{d_0-1}(r_1\sqrt{d_0}))$, $\boldsymbol{z}_2 \sim \mathrm{Unif}(\mathbb{S}^{d-d_0-1}(r_2\sqrt{d-d_0}))$, and $r_1^2 \geqslant r_2^2$. We find that, if $r_1/r_2$ (which we will denote later as the covariates signal-to-noise ratio) is sufficiently large, then the curse of dimensionality becomes milder for RKHS methods. We characterize precisely how the performance of these methods depend on the covariate signal-to-noise ratio $r_1/r_2$, the signal dimension $d_0$, and the ambient dimension $d$.

Notice that the spiked covariate model is highly stylized. For instance, while we expect real images to have a latent low-dimensional structure, this is best modeled in a nonlinear fashion (e.g. sparsity in wavelet domain [31]). Nevertheless the spiked covariate model captures the two basic mechanisms, and provides useful qualitative predictions. As an illustration, consider adding noise to the high-frequency components of images in a classification task. This will make the distribution of $\boldsymbol{x}$ more isotropic, and—according to our theory—deteriorate the performances of RKHS methods. On the other hand, NN should be less sensitive to this perturbation. (Notice that noise is added both to train and test samples.) In figure 1 we carry out such an experiment using FMNIST data ($d = 784$, $n = 60\,000$, 10 classes). We compare two-layers NN with the RF and NT models. We choose the architectures of NN, NT, RF as to match the number of parameters: namely we used $N = 4096$ for NN and NT and $N = 321\,126$ for RF. We also fit the corresponding RKHS models (corresponding to $N = \infty$) using kernel ridge regression (KRR), and two simple polynomial models: $f_\ell(\boldsymbol{x}) = \sum_{k=0}^{\ell}\langle \boldsymbol{B}_k, \boldsymbol{x}^{\otimes k}\rangle$, for $\ell \in \{1, 2\}$. In the unperturbed dataset, all of these approaches have comparable accuracies (except the linear fit). As noise is added, RF, NT, and RKHS methods deteriorate rapidly. While the accuracy of NN decreases as well, it significantly outperforms other methods.

### 1.2. Notations and outline

Throughout the paper, we use bold lowercase letters $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots\}$ to denote vectors and bold uppercase letters $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots\}$ to denote matrices. We denote by $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ the set of $d$-dimensional vectors with radius $r$ and $\mathrm{Unif}(\mathbb{S}^{d-1}(r))$ be the uniform probability distribution on $\mathbb{S}^{d-1}(r)$. Further, we let $N(\mu, \tau^2)$ be the Gaussian distribution with mean $\mu$ and variance $\tau^2$.

Let $O_d(\cdot)$ (respectively $o_d(\cdot)$, $\Omega_d(\cdot)$, $\omega_d(\cdot)$) denote the standard big-O (respectively little-o, big-omega, little-omega) notation, where the subscript $d$ emphasizes the asymptotic variable. We denote by $o_{d,\mathbb{P}}(\cdot)$ the little-o in probability notation: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability.

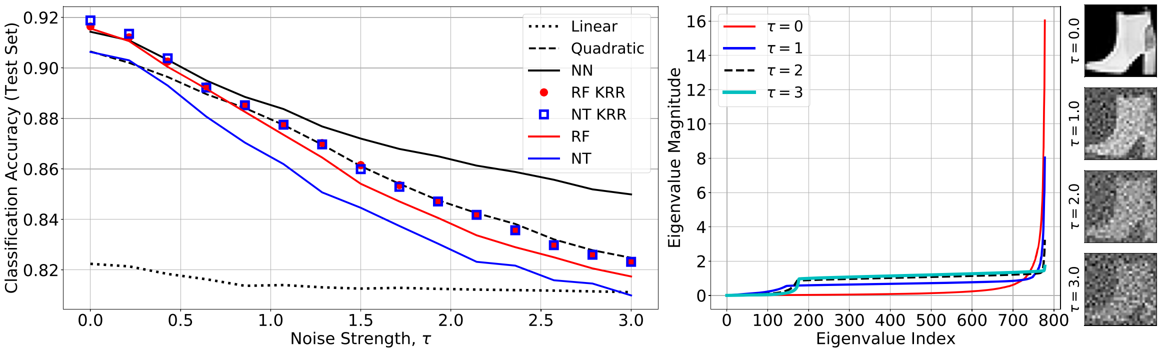

**Figure 1.** Test accuracy on Fashion MNIST (FMNIST) images perturbed by adding noise to the high-frequency Fourier components of the images (see examples on the right). (Left) Comparison of the accuracy of various methods as a function of the added noise. (Center) Eigenvalues of the empirical covariance of the images. As the noise increases, the images distribution becomes more isotropic.

In section 2, we introduce the spiked covariates model and characterize the performance of KRR, RF, NT, and NN models. Section 3 presents numerical experiments with real and synthetic data. Section 4 discusses our results in the context of earlier work.

## 2. Rigorous results for kernel methods and NT, RF NN expansions

### 2.1. The spiked covariates model

Let $d_0 = \lfloor d^\eta \rfloor$ for some $\eta \in (0,1)$. Let $\boldsymbol{U} \in \mathbb{R}^{d\times d_0}$ and $\boldsymbol{U}^\perp \in \mathbb{R}^{d\times(d-d_0)}$ be such that $[\boldsymbol{U}|\boldsymbol{U}^\perp]$ is an orthogonal matrix. We denote the subspace spanned by the columns of $\boldsymbol{U}$ by $\mathcal{V} \subseteq \mathbb{R}^d$ which we will refer to as the signal subspace, and the subspace spanned by the columns of $\boldsymbol{U}^\perp$ by $\mathcal{V}^\perp \subseteq \mathbb{R}^d$ which we will refer to as the noise subspace. In the case $\eta \in (0,1)$, the signal dimension $d_0 = \dim(\mathcal{V})$ is much smaller than the ambient dimension $d$. Our model for the covariate vector $\boldsymbol{x}_i$ is

$$\boldsymbol{x}_i = \boldsymbol{U}\boldsymbol{z}_{0,i} + \boldsymbol{U}^\perp \boldsymbol{z}_{1,i}, \qquad (\boldsymbol{z}_{0,i}, \boldsymbol{z}_{1,i}) \sim \mathrm{Unif}(\mathbb{S}^{d_0-1}(r\sqrt{d_0})) \otimes \mathrm{Unif}(\mathbb{S}^{d-d_0-1}(\sqrt{d-d_0})).$$

We call $\boldsymbol{z}_{0,i}$ the signal covariates, $\boldsymbol{z}_{1,i}$ the noise covariates, and $r$ the covariates signal-to-noise ratio (or covariates SNR). We will take $r > 1$, so that the variance of the signal covariates $\boldsymbol{z}_{0,i}$ is larger than that of the noise covariates $\boldsymbol{z}_{1,i}$. In high dimension, this model is—for many purposes—similar to an anisotropic Gaussian model $\boldsymbol{x}_i \sim \mathsf{N}(0, (r^2-1)\boldsymbol{U}\boldsymbol{U}^\mathsf{T} + \mathbf{I})$. As shown below, the effect of anisotropy on RKHS methods is significant

https://doi.org/10.1088/1742-5468/ac3a81

only if the covariate SNR $r$ is polynomially large in $d$. We shall therefore set $r = d^{\kappa/2}$ for a constant $\kappa > 0$.

We are given i.i.d. pairs $(y_i, \boldsymbol{x}_i)_{1\leqslant i\leqslant n}$, where $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, and $\varepsilon_i \sim N(0, \tau^2)$ is independent of $\boldsymbol{x}_i$. The function $f_*$ only depends on the projection of $\boldsymbol{x}_i$ onto the signal subspace $\mathcal{V}$ (i.e. on the signal covariates $\boldsymbol{z}_{0,i}$): $f_*(\boldsymbol{x}_i) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}_i)$, with $\varphi \in L^2(\mathbb{S}^{d_0-1}(r\sqrt{d_0}))$.

For the RF and NT models, we will assume that input layer weights to be i.i.d. $\boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$. For our purposes, this is essentially the same as $w_{ij} \sim \mathsf{N}(0, 1/d)$ independently, but slightly more convenient technically.

We will consider a more general model in appendix C, in which the distribution of $\boldsymbol{x}_i$ takes a more general product-of-uniforms form, and we assume a general $f_* \in L^2$.

### 2.2. A sharp characterization of RKHS methods

Given $h : [-1, 1] \to \mathbb{R}$, consider the rotationally invariant kernel $K_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$. This class includes the kernels that are obtained by taking the wide limit of the RF and NT models (here expectation is with respect to $(G_1, G_2) \sim \mathsf{N}(0, \mathbf{I}_2)$)

$$h_{\mathrm{RF}}(t) := \mathbb{E}\{\sigma(G_1)\sigma(tG_1 + \sqrt{1-t^2}G_2)\}, \qquad h_{\mathrm{NT}}(t) := t\mathbb{E}\{\sigma'(G_1)\sigma'(tG_1 + \sqrt{1-t^2}G_2)\}.$$

(These formulae correspond to $\boldsymbol{w}_i \sim \mathsf{N}(0, \mathbf{I}_d)$, but similar formulae hold for $\boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.) This correspondence holds beyond two-layers networks: under i.i.d. Gaussian initialization, the NT kernel for an arbitrary number of fully-connected layers is rotationally invariant (see the proof of proposition 2 of [2]), and hence is covered by the present analysis.

Any RKHS method with kernel $h$ outputs a model of the form $\hat{f}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i\leqslant n} a_i h(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle/d)$, with RKHS norm given by $\|\hat{f}(\cdot\,; \boldsymbol{a})\|_h^2 = \sum_{i,j\leqslant n} h(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d)a_i a_j$. We consider KRR on the dataset $\{(y_i, \boldsymbol{x}_i)\}_{i\leqslant n}$ with regularization parameter $\lambda$, namely:

$$\hat{\boldsymbol{a}}(\lambda) := \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x}_i; \boldsymbol{a})\right)^2 + \lambda\|\hat{f}(\cdot\ ; \boldsymbol{a})\|_h^2 \right\} = (\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{y},$$

where $\boldsymbol{H} = (H_{ij})_{ij\in[n]}$, with $H_{ij} = h(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d)$. We denote the prediction error of KRR by

$$R_{\mathrm{KRR}}(f_*, \lambda) = \mathbb{E}_{\boldsymbol{x}}\left[\left(f_*(\boldsymbol{x}) - \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{h}(\boldsymbol{x})\right)^2\right],$$

where $\boldsymbol{h}(\boldsymbol{x}) = (h(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d), \ldots, h(\langle \boldsymbol{x}, \boldsymbol{x}_n\rangle/d))^{\mathsf{T}}$.

Recall that we assume the target function $f_*(\boldsymbol{x}_i) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}_i)$. We denote $\mathsf{P}_{\leqslant k} : L^2 \to L^2$ to be the projection operator onto the space of degree $k$ orthogonal polynomials, and $\mathsf{P}_{>k} = \mathbf{I} - \mathsf{P}_{\leqslant k}$. Our next theorem shows that the impact of the low-dimensional latent structure on the generalization error of KRR is characterized by a certain *'effective dimension'*, $d_{\mathrm{eff}}$.

**Theorem 1.** *Let $h \in C^\infty([-1, 1])$. Let $\ell \in \mathbb{Z}_{\geqslant 0}$ be a fixed integer. We assume that $h^{(k)}(0) > 0$ for all $k \leqslant \ell$, and assume that there exists a $k > \ell$ such that $h^{(k)}(0) > 0$. (Recall that $h$ is positive semidefinite whence $h^{(k)}(0) \geqslant 0$ for all $k$.)*

https://doi.org/10.1088/1742-5468/ac3a81

*Define the effective dimension* $d_{\text{eff}} = \max\{d_0, d/r^2\} = d^{\max(1-\kappa,\eta)}$. *If* $\omega_d(d_{\text{eff}}^{\ell} \log(d_{\text{eff}})) \leqslant n \leqslant d_{\text{eff}}^{\ell+1-\delta}$ *for some* $\delta > 0$, *then for any regularization parameter* $\lambda = O_d(1)$, *the prediction error of KRR with kernel h is*

$$\left|R_{\text{KRR}}(f_*; \lambda) - \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2\right| \leqslant o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \tau^2). \tag{5}$$

Remarkably, the effective dimension $d_{\text{eff}} = d^{\max(1-\kappa,\eta)}$ depends both on the signal dimension $\dim(\mathcal{V}) = d^{\eta}$ and on the covariate SNR $r = d^{\kappa/2}$. Sample size $n = d_{\text{eff}}^{\ell}$ is necessary to learn a degree $\ell$ polynomial. If we fix $\eta \in (0, 1)$ and take $\kappa = 0+$, we get $d_{\text{eff}} \approx d$: this corresponds to almost isotropic $\boldsymbol{x}_i$. We thus recover theorem 4 in [21]. If instead $\kappa > 1 - \eta$, then most variance of $\boldsymbol{x}_i$ falls in the signal subspace $\mathcal{V}$, and we get $d_{\text{eff}} = d^{\eta} = \dim(\mathcal{V})$: the test error is effectively the same as if we had oracle knowledge of the signal subspace $\mathcal{V}$ and performed KRR on signal covariates $\boldsymbol{z}_{0,i} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}_i$. Theorem 1 describes the transition between these two regimes.

### 2.3. RF and NT models

How do the results of the previous section generalize to finite-width approximations of the RKHS? In particular, how do the RF and NT models behave at finite $N$? In order to simplify the picture, we focus here on the approximation error. Equivalently, we assume the sample size to be $n = \infty$ and consider the minimum population risk for $\mathsf{M} \in \{\text{RF}, \text{NT}\}$

$$R_{\mathsf{M},N}(f_*; \boldsymbol{W}) := \inf_{\hat{f} \in \mathcal{F}_{\mathsf{M}}^{N}(\boldsymbol{W})} \mathbb{E}\left\{\left[f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right]^2\right\}. \tag{6}$$

The next two theorems characterize the asymptotics of the approximation error for RF and NT models. We give generalizations of these statements to other settings and under weaker assumptions in appendix C.

**Theorem 2 (Approximation error for RF).** *Assume* $\sigma \in C^{\infty}(\mathbb{R})$, *with kth derivative* $\sigma^{(k)}(x)^2 \leqslant c_{0,k}\, \mathrm{e}^{c_{1,k}x^2/2}$ *for some* $c_{0,k} > 0$, $c_{1,k} < 1$, *and all* $x \in \mathbb{R}$ *and all k. Define its kth Hermite coefficient* $\mu_k(\sigma) := \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)\text{He}_k(G)]$. *Let* $\ell \in \mathbb{Z}_{\geqslant 0}$ *be a fixed integer, and assume* $\mu_k(\sigma) \neq 0$ *for all* $k \leqslant \ell$. *Define* $d_{\text{eff}} = d^{\max(1-\kappa,\eta)}$. *If* $d_{\text{eff}}^{\ell+\delta} \leqslant N \leqslant d_{\text{eff}}^{\ell+1-\delta}$ *for some* $\delta > 0$ *independent of N, d, then*

$$\left|R_{\text{RF},N}(f_*; \boldsymbol{W}) - \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2\right| \leqslant o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\ell} f_*\|_{L^2} \|f_*\|_{L^2}. \tag{7}$$

**Theorem 3 (Approximation error for NT).** *Assume* $\sigma \in C^{\infty}(\mathbb{R})$, *with kth derivative* $\sigma^{(k)}(x)^2 \leqslant c_{0,k}\, \mathrm{e}^{c_{1,k}x^2/2}$, *for some* $c_{0,k} > 0$, $c_{1,k} < 1$, *and all* $x \in \mathbb{R}$ *and all k. Let* $\ell \in \mathbb{Z}_{\geqslant 0}$, *and assume* $\mu_k(\sigma) \neq 0$ *for all* $k \leqslant \ell + 1$. *Further assume that, for all* $L \in \mathbb{Z}_{\geqslant 0}$, *there exist* $k_1, k_2$ *with* $L < k_1 < k_2$, *such that* $\mu_{k_1}(\sigma') \neq 0$, $\mu_{k_2}(\sigma') \neq 0$, *and* $\mu_{k_1}(x^2\sigma')/\mu_{k_1}(\sigma') \neq \mu_{k_2}(x^2\sigma')/\mu_{k_2}(\sigma')$. *Define* $d_{\text{eff}} = d^{\max(1-\kappa,\eta)}$. *If* $d_{\text{eff}}^{\ell+\delta} \leqslant N \leqslant d_{\text{eff}}^{\ell+1-\delta}$ *for some* $\delta > 0$ *independent of N, d, then*

$$\left|R_{\text{NT},N}(f_*; \boldsymbol{W}) - \|\mathsf{P}_{>\ell+1} f_*\|_{L^2}^2\right| \leqslant o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\ell+1} f_*\|_{L^2} \|f_*\|_{L^2}. \tag{8}$$

Here, the definitions of effective dimension $d_{\text{eff}}$ is the same as in theorem 1. While for the test error of KRR as in theorem 1, the effective dimension controls the sample complexity $n$ in learning a degree $\ell$ polynomial, in the present case it controls the number of neurons $N$ that is necessary to approximate a degree $\ell$ polynomial. In the case of RF, the latter happens as soon as $N \gg d_{\text{eff}}^{\ell}$, while for NT it happens as soon as $N \gg d_{\text{eff}}^{\ell-1}$. If we take $\eta \in (0, 1)$ and $\kappa = 0+$, the above theorems, again, recover theorems 1 and 2 of [21].

Notice that NT has higher approximation power than RF *in terms of the number of neurons*. This is expected, since NT models contain $Nd$ instead of $N$ parameters. On the other hand, NT has less power *in terms of number of parameters*: to fit a degree $\ell + 1$ polynomial, the parameter complexity for NT is $Nd = d_{\text{eff}}^{\ell} d$ while the parameter complexity for RF is $N = d_{\text{eff}}^{\ell+1} \ll d_{\text{eff}}^{\ell} d$. While the NT model has $p = Nd$ parameters, only $p_{\text{eff}}^{\text{NT}} = N d_{\text{eff}}$ of them appear to matter. We will refer to $p_{\text{eff}}^{\text{NT}} \equiv N d_{\text{eff}}$ as the *effective number of parameters* of NT models.

Finally, it is natural to ask what are the behaviors of RF and NT models at finite sample size. Denote by $R_{M,N,n}(f_*; \mathbf{W})$ the corresponding test error (assuming for instance ridge regression, with the optimal regularization $\lambda$). Of course the minimum population risk provides a lower bound: $R_{M,N,n}(f_*; \mathbf{W}) \geqslant R_{M,N}(f_*; \mathbf{W})$. Moreover, we conjecture that the risk is minimized at infinite $N$, $R_{M,N,n}(f_*; \mathbf{W}) \gtrsim R_n(f_*; h_M)$. Altogether this implies the lower bound $R_{M,N,n}(f_*; \mathbf{W}) \gtrsim \max(R_{M,N}(f_*; \mathbf{W}), R_n(f_*; h_M))$. We also conjecture that this lower bound is tight, up to terms vanishing as $N, n, d \to \infty$.

Namely (focusing on NT models), if $N d_{\text{eff}} \lesssim n$, and $d_{\text{eff}}^{\ell_1} \lesssim N d_{\text{eff}} \lesssim d_{\text{eff}}^{\ell_1+1}$ then the approximation error dominates and $R_{\mathsf{M},N,n}(f_*; \boldsymbol{W}) = \|\mathsf{P}_{>\ell_1} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\|f_*\|_{L^2}^2$. If on the other hand $N d_{\text{eff}} \gtrsim n$, and $d_{\text{eff}}^{\ell_2} \lesssim n \lesssim d_{\text{eff}}^{\ell_2+1}$ then the generalization error dominates and $R_{\mathsf{M},N,n}(f_*; \boldsymbol{W}) = \|\mathsf{P}_{>\ell_2} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\|f_*\|_{L^2}^2$.

### 2.4. Neural network models

Consider the approximation error for NNs

$$R_{\text{NN},N}(f_*) := \inf_{\hat{f} \in \mathcal{F}_{\text{NN}}^N} \mathbb{E}\left\{\left[f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right]^2\right\}. \tag{9}$$

Since $\varepsilon^{-1}[\sigma(\langle \boldsymbol{w}_i + \varepsilon \boldsymbol{a}_i, \boldsymbol{x}\rangle) - \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)] \overset{\varepsilon \to 0}{\to} \underline{\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle),}$ we have $\cup_{\boldsymbol{W}} \mathcal{F}_{\text{NT}}^{N/2}(\boldsymbol{W}) \subseteq \text{cl}(\mathcal{F}_{\text{NN}}^N)$, and $R_{\text{NN},N}(f_*) \leqslant \inf_{\mathbf{W}} R_{\text{NT},N/2}(f_*, \mathbf{W})$. By choosing $\overline{\boldsymbol{W}} = (\bar{\boldsymbol{w}}_i)_{i \leqslant N}$, with $\bar{\boldsymbol{w}}_i = \boldsymbol{U}\bar{\boldsymbol{v}}_i$ (see section 2.1 for definition of $\mathbf{U}$), we obtain that $\mathcal{F}_{\text{NT}}^N(\overline{\boldsymbol{W}})$ contains all functions of the form $\bar{f}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x})$, where $\bar{f}$ is in the class of functions $\mathcal{F}_{\text{NT}}^N(\overline{\boldsymbol{V}})$ on $\mathbb{R}^{d_0}$. Hence if $f_*(\boldsymbol{x}) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x})$, $R_{\text{NN},N}(f_*)$ is at most the error of approximating $\varphi(\mathbf{z})$ on the small sphere $\boldsymbol{z} \sim \text{Unif}(\mathbb{S}^{d_0-1})$ within the class $\mathcal{F}_{\text{NT}}^N(\overline{\boldsymbol{V}})$. As a consequence, by theorem 3, if $d_0^{\ell+\delta} \leqslant N \leqslant d_0^{\ell+1-\delta}$ for some $\delta > 0$, then $R_{\text{NN},N}(f_*) \leqslant R_{\text{NT},N/2}(f_*, \overline{\boldsymbol{W}}) \leqslant (1 + o_{d,\mathbb{P}}(1)) \cdot \|\mathsf{P}_{>\ell+1} f_*\|_{L^2}^2$.

**Theorem 4 (Approximation error for NN).** *Assume that $\sigma \in C^{\infty}(\mathbb{R})$ satisfies the same assumptions as in theorem 3. Further assume that $\sup_{x \in \mathbb{R}}|\sigma''(x)| < \infty$. If $d_0^{\ell+\delta} \leqslant N \leqslant d_0^{\ell+1-\delta}$ for some $\delta > 0$ independent of $N, d$, then the approximation error of NN models (3) is*

$$R_{\text{NN},N}(f_*) \leqslant (1 + o_d(1)) \cdot \|\mathsf{P}_{>\ell+1} f_*\|_{L^2}^2. \tag{10}$$

*Moreover, the quantity $R_{\text{NN},N}(f_*)$ is independent of $\kappa \geqslant 0$.*

https://doi.org/10.1088/1742-5468/ac3a81

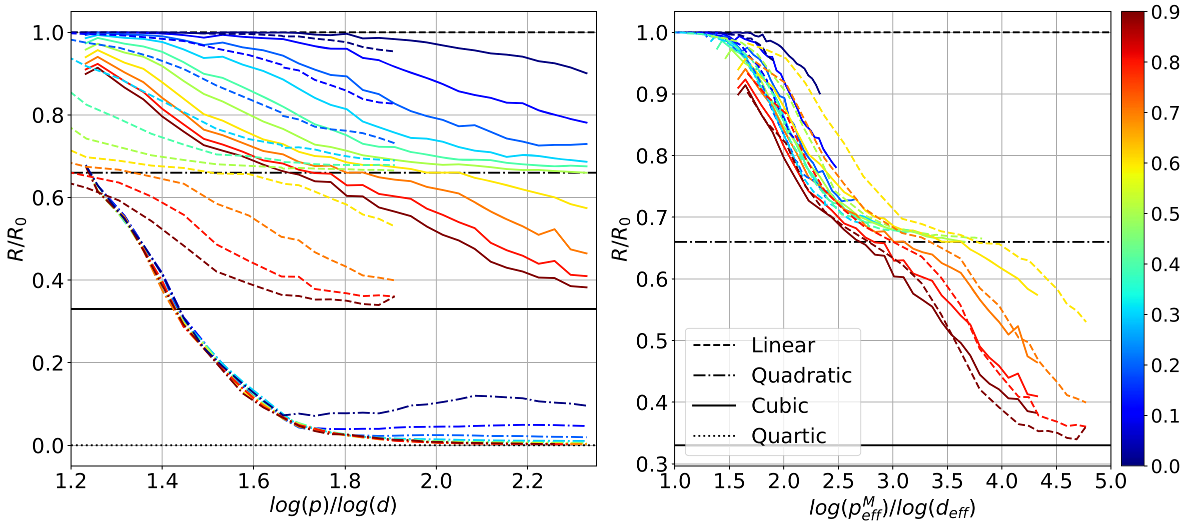

**Figure 2.** Finite-width two-layers NN and their linearizations RF and NT. Models are trained on $2^{20}$ training observations drawn i.i.d. from the distribution of section 2.1. Continuous lines: NT; dashed lines: RF; dot-dashed: NN. Various curves (colors) refer to values of the exponent $\kappa$ (larger $\kappa$ corresponds to stronger low-dimensional component). (Right) Curves for RF and NT as a function of the rescaled quantity $\log(p^{\mathsf{M}}_{\mathrm{eff}})/\log(d_{\mathrm{eff}})$.

As a consequence of theorems 3 and 4, there is a separation between NN and (uniformly sampled) NT models when $d_{\mathrm{eff}} \neq d_0$, i.e. $\kappa < 1 - \eta$. As $\kappa$ increases, the gap between NN and NT becomes smaller and smaller until $\kappa = 1 - \eta$.

## 3. Further numerical experiments

We carried out extensive numerical experiments on synthetic data to check our predictions for RF, NT, RKHS methods at finite sample size $n$, dimension $d$, and width $N$. We simulated two-layers fully-connected NN in the same context in order to compare their behavior to the behavior of the previous models. Finally, we carried out numerical experiments on FMNIST and CIFAR-10 data to test whether our qualitative predictions apply to image datasets. Throughout we use ReLU activations.

In figure 2 we investigate the approximation error of RF, NT, and NN models. We generate data $(y_i, \boldsymbol{x}_i)_{i\geqslant 1}$ according to the model of section 2.1, in $d = 1024$ dimensions, with a latent space dimension $d_0 = 16$, hence $\eta = 2/5$. The per-coordinate variance in the latent space is $r^2 = d^{\kappa}$, with $\kappa \in \{0.0, \ldots, 0.9\}$. Labels are obtained by $y_i = f_*(\boldsymbol{x}_i) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}_i)$ where $\varphi : \mathbb{R}^{d_0} \to \mathbb{R}$ is a degree-4 polynomial, without a linear component. Since

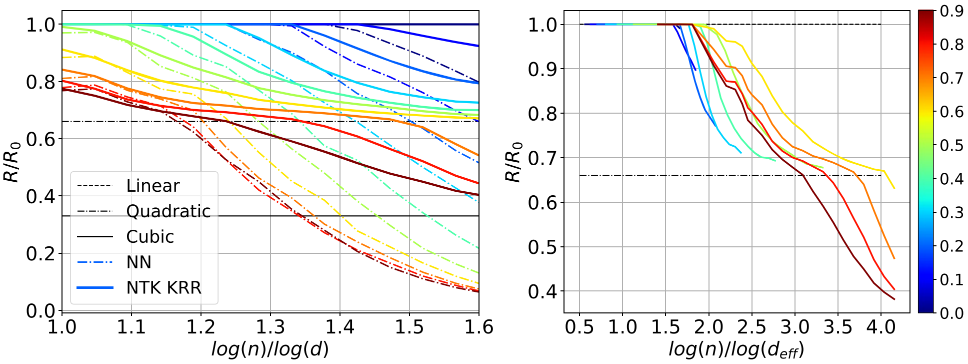

**Figure 3.** (Left) Comparison of the test error of NN (dot-dashed) and NTK KRR (solid) on the distribution of the section 2.1. Various curves (colors) refer to values of the exponent $\kappa$. (Right) KRR test error as a function of the number of observations adjusted by the effective dimension. Horizontal lines correspond to the best polynomial approximation.

we are interested in the minimum population risk, we use a large sample size $n = 2^{20}$: we expect the approximation error to dominate in this regime. (See appendix A for further details.)

We plot the normalized risk $R_{\mathrm{RF},N}(f_*, \mathbf{W})/R_0$, $R_{\mathrm{NT},N}(f_*, \mathbf{W})/R_0$, $R_{\mathrm{NN},N}(f_*)/R_0$, $R_0 := \|f_*\|_{L^2}^2$, for various widths $N$. These are compared with the error of the best polynomial approximation of degrees $\ell = 1$ to 3 (which correspond to $\|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 / \|f_*\|_{L^2}^2$). As expected, as the number of parameters increases, the approximation error of each function class decreases. NN provides much better approximations than any of the linear classes, and RF is superior to NT *given the same number of parameters*. This is captured by theorems 2 and 3: to fit a degree $\ell + 1$ polynomial, the parameter complexity for NT is $Nd = d_{\mathrm{eff}}^{\ell} d$ while for RF it is $N = d_{\mathrm{eff}}^{\ell+1} \ll d_{\mathrm{eff}}^{\ell} d$. We denote the effective number of parameters for NT by $p_{\mathrm{eff}}^{\mathrm{NT}} = N d_{\mathrm{eff}}$ and the effective number of parameter for RF by $p_{\mathrm{eff}}^{\mathrm{RF}} = N$. The right plot reports the same data, but we rescale the $x$-axis to be $\log(p_{\mathrm{eff}}^{\mathsf{M}}) / \log(d_{\mathrm{eff}})$. As predicted by the asymptotic theory of theorems 2 and 3, various curves for NT and RF tend to collapse on this scale. Finally, the approximation error of RF and NT depends strongly on $\kappa$: larger $\kappa$ leads to smaller effective dimension and hence smaller approximation error. In contrast, the error of NN, besides being smaller in absolute terms, is much less sensitive to $\kappa$.

In figure 3 we compare the test error of NN (with $N = 4096$) and KRR for the NT kernel (corresponding to the $N \to \infty$ limit in the lazy regime), for the same data distribution as in the previous figure. We observe that the test error of KRR is substantially larger than the one of NN, and deteriorates rapidly as $\kappa$ gets smaller (the effective dimension gets larger). In the right frame we plot the test error as a function of $\log(n)/\log(d_{\mathrm{eff}})$: we observe that the curves obtained for different $\kappa$ approximately collapse, confirming that $d_{\mathrm{eff}}$ is indeed the right dimension parameter controlling the sample complexity. Notice that also the error of NN deteriorates as $\kappa$ gets smaller, although not so rapidly: this behavior deserves further investigation. Notice also that the KRR error crosses the level of best degree-$\ell$ polynomial approximation roughly at $\log(n)/\log(d_{\mathrm{eff}}) \approx \ell$.

https://doi.org/10.1088/1742-5468/ac3a81

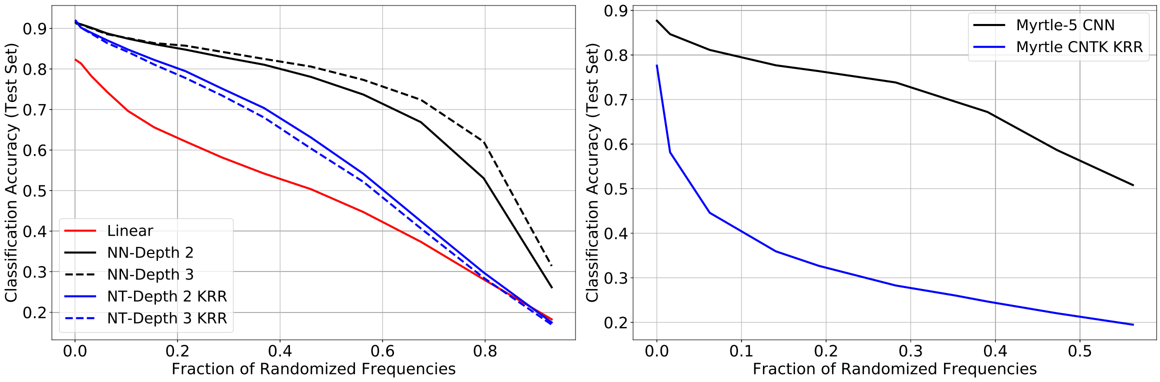

**Figure 4.** Comparison between multilayer NNs and the corresponding NT models under perturbations in frequency domain. (Left) Fully connected networks on FMNIST data. (Right) Comparison of CNN and convolutional neural tangent kernel (CNTK) KRR classification accuracy on CIFAR-10. We progressively replace the lowest frequencies of each image with Gaussian noise with matching covariance structure. Right: accuracy for FMNIST.

The basic qualitative insight of our work can be summarized as follows. Kernel methods are effective when a low-dimensional structure in the target function is aligned with a low-dimensional structure in the covariates. In image data, both the target function and the covariates are dominated by the low-frequency subspace. In figure 1 we tested this hypothesis by removing the low-dimensional structure of the covariate vectors: we simply added noise to the high-frequency part of the image. In figure 4 we try the opposite, by removing the component of the target function that is localized on low-frequency modes. We decompose each images into a low-frequency and a high-frequency part. We leave the high-frequency part unchanged, and replace the low-frequency part by Gaussian noise with the first two moments matching the empirical moments of the data.

In the left frame, we consider FMNIST data and compare fully-connected NNs with two or three layers (and $N = 4096$ nodes at each hidden layer) with the corresponding NT KRR model (infinite width). In the right frame, we use CIFAR-10 data and compare a Myrtle-5 network (a lightweight convolutional architecture [16, 32]) with the corresponding NT KRR. We observe the same behavior as in figure 1. While for the original data NT is comparable to NN, as the proportion of perturbed Fourier modes increases, the performance of NT deteriorates much more rapidly than the one of NN.

## 4. Discussion

The limitations of linear methods—such as KRR—in high dimension are well understood in the context of nonparametric function estimation. For instance, a basic result in this area establishes that estimating a Sobolev function $f_*$ in $d$ dimensions with mean square error $\varepsilon$ requires roughly $\varepsilon^{-2-d/\alpha}$ samples, with $\alpha$ the smoothness parameter [33].

This behavior is achieved by kernel smoothing and by KRR: however these methods are not expected to be adaptive when $f_*(\mathbf{x})$ only depends on a low-dimensional projection of $\mathbf{x}$, i.e. $f_*(\boldsymbol{x}) = \varphi(\boldsymbol{U}^\mathsf{T}\boldsymbol{x})$ for an unknown $\boldsymbol{U} \in \mathbb{R}^{d_0\times d}$, $d_0 \ll d$. On the contrary, fully-trained NN can overcome this problem [20].

However, these classical statistical results have some limitations. First, they focus on the low-dimensional regime: $d$ is fixed, while the sample size $n$ diverges. This is probably unrealistic for many machine learning applications, in which $d$ is at least of the order of a few hundreds. Second, classical lower bounds are typically established for the minimax risk, and hence they do not necessarily apply to specific functions.

To bridge these gaps, we developed a sharp characterization of the test error in the high-dimensional regime in which both $d$ and $n$ diverge, while being polynomially related. This characterization holds for any target function $f_*$, and expresses the limiting test error in terms of the polynomial decomposition. We also present analogous results for finite-width RF and NT models.

Our analysis is analogous and generalizes the recent results of [21]. However, while [21] assumed the covariates $\mathbf{x}_i$ to be uniformly distributed over the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, we introduced and analyzed a more general model in which the covariates mostly lie in the signal subspace with dimension $d_0 \ll d$, and the target function is also dependent on that subspace. In fact our results follow as special cases of a more general model discussed in appendix C.

Depending on the relation between signal dimension $d_0$, ambient dimension $d$, and the covariate signal-to-noise ratio $r$, the model presents a continuum of different behaviors. At one extreme, the covariates are fully $d$-dimensional, and RKHS methods are highly suboptimal compared to NN. At the other, covariates are close to $d_0$-dimensional and RKHS methods are instead more competitive with NN.

Finally, the Fourier decomposition of images is a simple proxy for the decomposition of the covariate vector $\mathbf{x}$ into its low-dimensional dominant component (low frequency) and high-dimensional component (high frequency) [34].

## Acknowledgments

This work was partially supported by the NSF Grants CCF-1714305, IIS-1741162, DMS-1418362, DMS-1407813 and by the ONR Grant N00014-18-1-2729.

## Data availability statement

The code used to produce our results can be accessed at https://github.com/bGhorbani/linearized_neural_networks.

## Appendix A. Details of numerical experiments

### A.1. General training details

All models studied in the paper are trained with squared loss and $\ell_2$ regularization. For multi-class datasets such as FMNIST, one-hot encoded labels are used for training. All models discussed in the paper use ReLU non-linearity. Fully-connected models are initialized according to mean-field parameterization [26, 35, 36]. All NN are optimized with SGD with 0.9 momentum. The learning-rate evolves according to the cosine rule

$$lr_t = lr_0 \max\left(\left(1 + \cos\left(\frac{t\pi}{T}\right)\right), \frac{1}{15}\right) \tag{11}$$

where $lr_0 = 10^{-3}$ and $T = 750$ is the total number of training epochs. To ensure the stability of the optimization for wide models, we use 15 linear warm-up epochs in the beginning.

When $N \gg 1$, training RF and NT with SGD is unstable (unless extremely small learning-rates are used). This makes the optimization prohibitively slow for large datasets. To avoid this issue, instead of SGD, we use conjugate gradient method (CG) for optimizing RF and NT. Since these two models are strongly convex[5], the optimizer is unique. Hence, using CG will not introduce any artifacts in the results.

In order to use CG, we first implement a function to perform Hessian-vector products in TensorFlow [37]. The function handle is then passed to scipy.sparse.cg for CG. Our Hessian-vector product code uses tensor manipulation utilities implemented by [38].

Unfortunately, scipy.sparse.cg does not support one-hot encoded labels. To avoid running CG for each class separately, when the labels are one-hot encoded, we use Adam optimizer [39] instead. When using Adam, the learning-rate still evolves as (11) with $lr_0 = 10^{-5}$. The batch-size is fixed at $10^4$ to encourage fast convergence to the minimum.

For NN, RF and NT, the training is primary done in TensorFlow (v1.12) [37]. For KRR, we generate the kernel matrix first and directly fit the model in regular python. The kernels associated with two-layer models are calculated analytically. For deeper models, the kernels are computed using neural-tangents library in JAX [40, 41].

### A.2. Synthetic data experiments

The synthetic data follows the distribution outlined in the main text. In particular,

$$\boldsymbol{x}_i = (\boldsymbol{u}_i, \boldsymbol{z}_i), \qquad y_i = \varphi(\boldsymbol{u}_i), \quad \boldsymbol{u}_i \in \mathbb{R}^{d_0}, \; \boldsymbol{z}_i \in \mathbb{R}^{d-d_0}, \tag{12}$$

[5] Note that all models are trained with $\ell_2$ regularization.

**Table A.1.** Hyper-parameter details for synthetic data experiments.

| Experiment | Model | $\ell_2$ regularization grid |
|---|---|---|
| Approximation error (figure 2) | NN | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-8,-4]$ |
| | NT | $\{10^{\alpha_i}\}_{i=1}^{10}$, $\alpha_i$ uniformly spaced in $[-4,2]$ |
| | RF | $\{10^{\alpha_i}\}_{i=1}^{10}$, $\alpha_i$ uniformly spaced in $[-5,2]$ |
| Generalization error (figure 3) | NN | $\{10^{\alpha_i}\}_{i=1}^{25}$, $\alpha_i$ uniformly spaced in $[-8,-2]$ |
| | NT KRR | $\{10^{\alpha_i}\}_{i=1}^{10}$, $\alpha_i$ uniformly spaced in $[0,6]$ |

where $\mathbf{u}_i$ and $\mathbf{z}_i$ are drawn i.i.d. from the hyper-spheres with radii $r\sqrt{d_0}$ and $\sqrt{d}$ respectively. We choose

$$r = d^{\kappa/2}, \qquad d_0 = d^{\eta}, \tag{13}$$

where $d$ is fixed to be 1024 and $\eta = \frac{2}{5}$. We change $\kappa$ in the interval $\{0,\ldots,0.9\}$. For each value of $\kappa$ we generate $2^{20}$ training and $10^4$ test observations[6].

The function $\varphi$ is the sum of three orthogonal components $\{\varphi_i\}_{i=1}^3$ with $\|\varphi_i\|_2 = 1$. To be more specific,

$$\varphi_i(\boldsymbol{x}) \propto \sum_{j=1}^{d_0-i} \alpha_j^{(i)} \prod_{k=j}^{j+i} \boldsymbol{x}_k, \qquad \alpha_j^{(i)} \overset{\text{i.i.d.}}{\sim} \exp(1). \tag{14}$$

This choice of $\varphi_i$ guarantees that each $\varphi_i$ is in the span of degree $i+1$ spherical harmonics.

In the experiments presented in figure 2, for NN and NT, the number of hidden units $N$ takes 30 geometrically spaced values in the interval $[5, 10^4]$. NN models are trained using SGD with momentum 0.9 (the learning-rate evolution is described above). We use batch-size of 512 for the warm-up epochs and batch-size of 1024 for the rest of the training. For RF, $N$ takes 24 geometrically spaced values in the interval $[100, 711\,680]$. The limit $N = 711\,680$ corresponds to the largest model size we are computationally able to train at this scale. All models are trained with $\ell_2$ regularization. The $\ell_2$ regularization grids used for these experiments are presented in table A.1. In all our experiments, we choose the $\ell_2$ regularization parameter that yields the best test performance[7]. In total, we train approximately 10 000 different models just for this subset of experiments.

In figure 3 of the main text, we compared the generalization performance of NTK KRR with NN. We use the same training and test data as above to perform this analysis. The number of training data points, $n$, takes 24 different values ranging from 50 to $10^5$. The number of test data points is always fixed at $10^4$.

[6] Strictly speaking, the model outlined in the main text requires $\mathbf{z}_i$ to be generated from the hyper-sphere of radius $\sqrt{d-d_0}$. In order to work with round numbers, in our experiments we use $\sqrt{d}$ instead of $\sqrt{d-d_0}$. The numerical difference between these two choices is negligible.

[7] Due to the large size of the test set, choosing these hyper-parameter based on the test set performance has a negligible over-fitting effect. In addition, in studying the approximation error overfitting is not relevant.

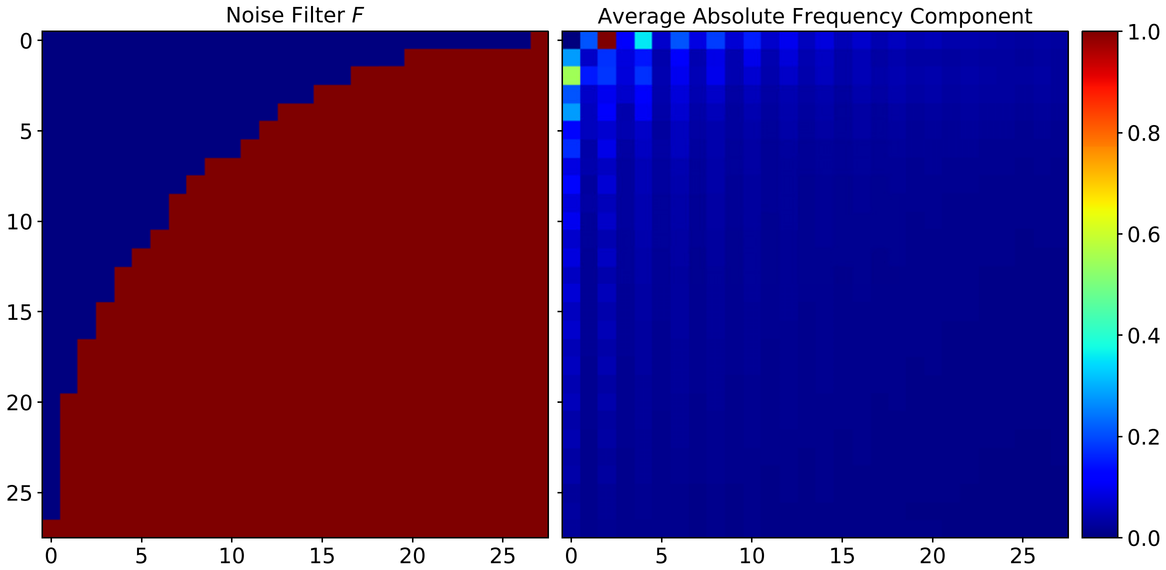

**Figure A.1.** (Left) The pictorial representation of the filter matrix $\mathbf{F}$ used for the FMNIST experiments. The matrix entries with value zero are represented by color blue while the entries with value one are represented by red. Coordinates on top left-hand side correspond to lower frequency components while coordinates closer to bottom right-hand side represent the high-frequency directions. (Right) The absolute value of the frequency components of FMNIST images averaged over the training data. The projection of the dataset into the low-frequency region chosen by the filter retains over 95% of the variation in the data.

### A.3. High-frequency noise experiment on FMNIST

In effort to make the distribution of the covariates more isotropic, in this experiment, we add high-frequency noise to both the training and test data.

Let $\boldsymbol{x} \in \mathbb{R}^{k\times k}$ be an image. We first remove the global average of the image and then add high-frequency Gaussian noise to $\mathbf{x}$ in the following manner:

(a) We convert $\boldsymbol{x}$ to frequency domain via discrete cosine transform (DCT II-orthogonal to be precise). We denote the representation of the image in the frequency domain $\tilde{\boldsymbol{x}} \in \mathbb{R}^{k\times k}$.

(b) We choose a filter $\boldsymbol{F} \in \{0,1\}^{k\times k}$. $\boldsymbol{F}$ determines on which frequencies the noise should be added. The noise matrix $\tilde{\boldsymbol{Z}}$ is defined as $\boldsymbol{Z} \odot \boldsymbol{F}$ where $\boldsymbol{Z} \in R^{k\times k}$ has i.i.d. $\mathsf{N}(0,1)$ entries.

(c) We define $\tilde{\boldsymbol{x}}_{\text{noisy}} = \tilde{\boldsymbol{x}} + \tau(\|\tilde{\boldsymbol{x}}\|/\|\tilde{\boldsymbol{Z}}\|)\tilde{\boldsymbol{Z}}$. The constant $\tau$ controls the noise magnitude.

(d) We perform inverse discrete cosine transform (DCT III-orthogonal) on $\tilde{\boldsymbol{x}}_{\text{noisy}}$ to convert the image to pixel domain. We denote the noisy image in the pixel domain as $\boldsymbol{x}_{\text{noisy}}$.

(e) Finally, we normalize the $\boldsymbol{x}_{\text{noisy}}$ so that it has norm $\sqrt{d}$.

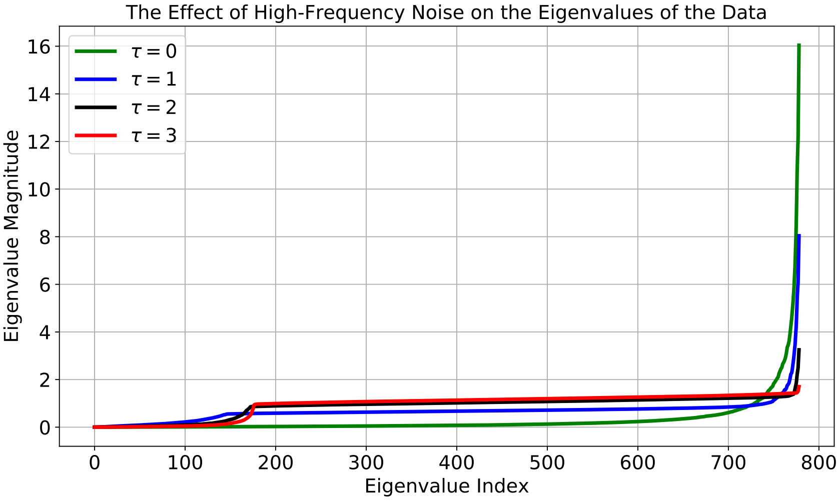

**Figure A.2.** The eigenvalues of the empirical covariance matrix of the FMNIST training data. As the noise intensity increases, the distribution of the eigenvalues becomes more isotropic. Note that due to the conservative choice of the filter $\mathbf{F}$, noise is not added to all of the low-variance directions. These left-out directions corresponds to the small eigenvalues appearing in the left-hand side of the plot.

In the frequency domain, a grayscale image is represented by a matrix $\tilde{\boldsymbol{x}} \in \mathbb{R}^{k\times k}$. Qualitatively speaking, elements $(\tilde{\boldsymbol{x}})_{i,j}$ with small values of $i$ and $j$ correspond to the low-frequency component of the image and elements with large indices correspond to high-frequency components. The matrix $\boldsymbol{F}$ is chosen such that no noise is added to low frequencies. Specifically, we choose

$$\boldsymbol{F}_{i,j} = \begin{cases} 1 & \text{if } (k-i)^2 + (k-j)^2 \leqslant (k-1)^2 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

This choice of $\boldsymbol{F}$ mirrors the average frequency domain representation of FMNIST images (see figure A.1 for a comparison). Figures A.2 and A.4 respectively show the eigenvalues of the empirical covariance of the dataset for various noise levels. As discussed in the main text, the distribution of the covariates becomes more isotropic as more and more high-frequency noise is added to the images.

Figure A.3 shows the normalized squared loss and the classification accuracy of the models as more and more high-frequency noise is added to the data. The normalization factor $R_0 = 0.9$ corresponds to the risk achievable by the (trivial) predictor $[\hat{y}_j(\boldsymbol{x})]_{1\leqslant j\leqslant 10} = 0.1$.

*A.3.1. Experiment hyper-parameters.* For NT and NN, the number of hidden units $N = 4096$. For RF, we fix $N = 321\,126$. These hyper-parameter choices ensure that the models have approximately the same number of trainable parameters. NN is trained

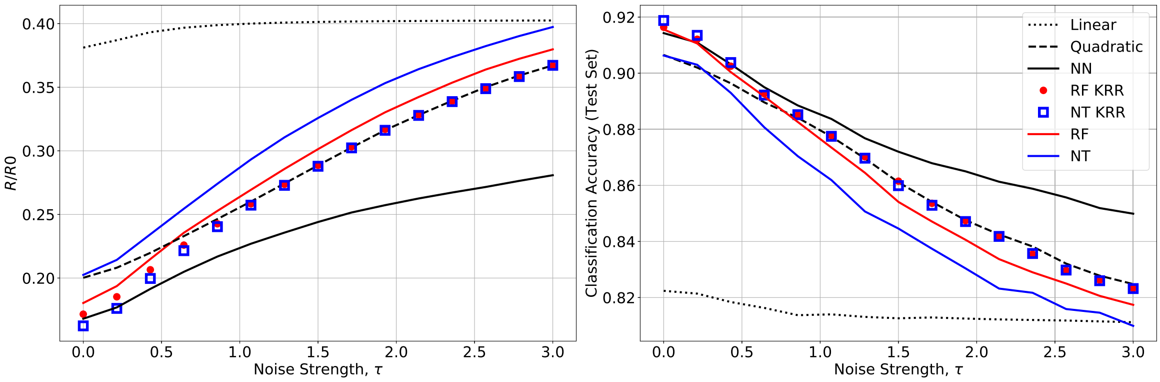

**Figure A.3.** The normalized test squared error (left) and the test accuracy (right) of the models trained and evaluated on FMNIST data with high-frequency noise.

**Table A.2.** Details of regularization parameters used for high-frequency noise experiments.

| Dataset | Model | $\ell_2$ regularization grid |
|---|---|---|
| FMNIST | NN | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-6,-2]$ |
| | NT | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-5,3]$ |
| | RF | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-5,3]$ |
| | NT KRR | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-1,5]$ |
| | RF KRR | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-1,5]$ |
| CIFAR-2 | NN | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-6,-2]$ |
| | NT | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-4,4]$ |
| | RF | $\{10^{\alpha_i}\}_{i=1}^{40}$, $\alpha_i$ uniformly spaced in $[-2,10]$ |
| | NT KRR | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-2,4]$ |
| | RF KRR | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-2,4]$ |

with SGD with 0.9 momentum and learning-rate described by (11). The batch-size for the warm-up epochs is 500. After the warm-up stage is over, we use batch-size of 1000 to train the network. Since CG is not available in this setting, NT and RF are optimized using Adam for $T = 750$ epochs with batch-size of $10^4$. The $\ell_2$ regularization grids used for training these models are listed in table A.2.

### A.4. High-frequency noise experiment on CIFAR-2

We perform a similar experiment on a subset of CIFAR-10. We choose two classes (airplane and cat) from the ten classes of CIFAR-10. This choice provides us with $10^4$ training and 2000 test data points. Given that the number of training observations is not very large, we reduce the covariate dimension by converting the images to grayscale. This transformation reduces the covariate dimension to $d = 1024$.

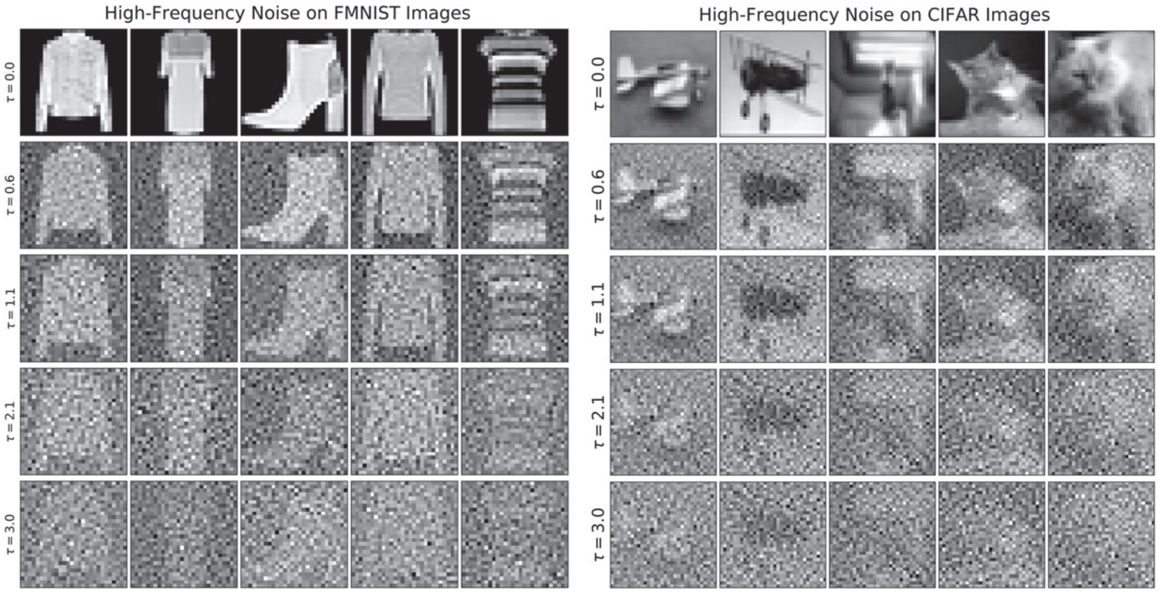
High-Frequency Noise on FMNIST Images
High-Frequency Noise on CIFAR Images
τ = 0.0
τ = 0.6
τ = 1.1
τ = 2.1
τ = 3.0

**Figure A.4.** (Left) FMNIST images with various high-frequency noise levels. (Right) CIFAR-2 images with various levels of high-frequency Gaussian noise. The images are converted to grayscale to make the covariate dimension manageable.

Figure A.5 demonstrates the evolution of the model performances as the noise intensity increases. In the noiseless regime ($\tau = 0$), all models have comparable performances. However, as the noise level increases, the performance gap between NN and RKHS methods widens. For reference, the accuracy gap between NN and NT KRR is only 0.6% at $\tau = 0$. However, at $\tau = 3$, this gap increases to 4.5%. The normalization factor $R_0 = 0.25$ corresponds to the risk achievable by the trivial estimator $\hat{y}(\boldsymbol{x}) = 0.5$.

*A.4.1. Experiment hyper-parameters.* For NT and NN, the number of hidden units $N = 4096$. For RF, we fix $N = 4.2 \times 10^6$. These hyper-parameter choices ensure that the models have approximately the same number of trainable parameters. NN is trained with SGD with 0.9 momentum and learning-rate described by (11). The batch-size is fixed at 250. NT is optimized via CG with 750 maximum iterations. The $\ell_2$ regularization grids used for training these models are listed in table A.2.

### A.5. Low-frequency noise experiments on FMNIST

To examine the ability of NN and RKHS methods in learning the information in low-variance components of the covariates, we replace the low-frequency components of the image with Gaussian noise. To be specific, we follow the following steps to generate the noisy datasets:

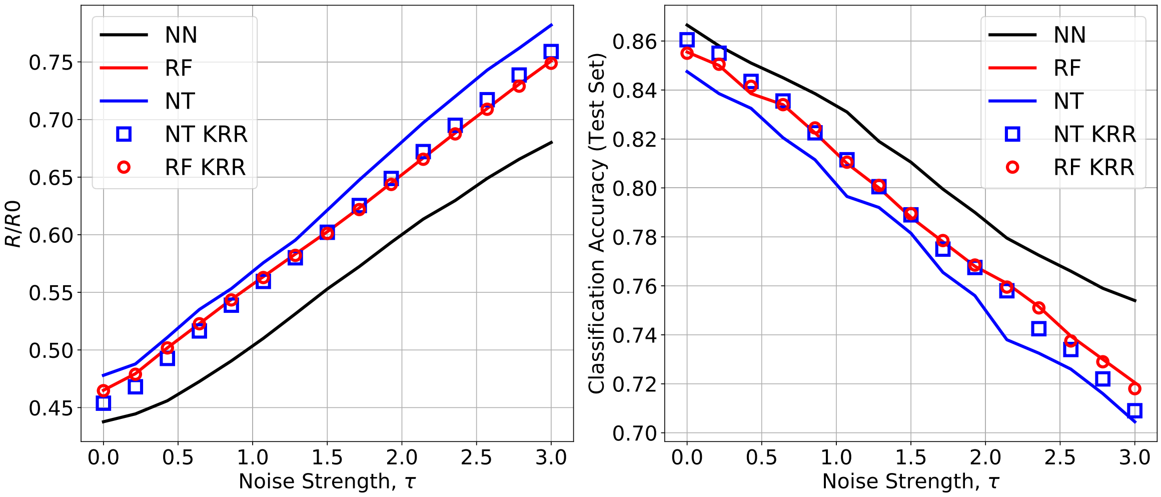

**Figure A.5.** Normalized test squared error (left) and test classification accuracy (right) of the models on noisy CIFAR-2. As the noisy intensity increases, the performance gap between NN and RKHS methods widens. For reference, the accuracy gap between NN and NT KRR is only 0.6% at $\tau = 0$. However, at $\tau = 3$, this gap increases to 4.5%. For finite-width models, $N$ is chosen such that the number of trainable parameters is approximately equal across the models. For NN and NT, $N = 4096$ and for RF, $N = 4.2 \times 10^6$. We use the noise filter described in (15).

(a) We normalize all images to have mean zero and norm $\sqrt{d}$.

(b) Let $\mathcal{D}_{\text{train}}$ denote the set of training images in the DCT-frequency domain. We compute the mean $\mu$ and the covariance $\Sigma$ of the elements of $\mathcal{D}_{\text{train}}$.

(c) We fix a threshold $\alpha \in \mathbb{N}$ where $1 \leqslant \alpha \leqslant k$.

(d) Let $\boldsymbol{x}$ be an image in the dataset (test or train). We denote the representation of $\boldsymbol{x}$ in the frequency domain with $\tilde{\boldsymbol{x}}$. For each image, we draw a noise matrix $\boldsymbol{z} \sim \mathcal{N}(\mu, \Sigma)$. We have

$$[\tilde{\boldsymbol{x}}_{\text{noisy}}]_{i,j} = \begin{cases} (\boldsymbol{z})_{i,j} & \text{if}\, i, j \leqslant \alpha \\ \tilde{\boldsymbol{x}}_{i,j} & \text{otherwise.} \end{cases}$$

(e) We perform IDCT on $\tilde{\boldsymbol{x}}_{\text{noisy}}$ to get the noisy image $\boldsymbol{x}_{\text{noisy}}$.

The fraction of the frequencies replaced by noise is $\alpha^2/k^2$. Figure A.8 shows several examples of noisy images for different thresholds $\alpha$.

*A.5.1. Experiment hyper-parameters.* For NN trained for this experiment, we fix the number of hidden units per-layer to $N = 4096$. This corresponds to approximately $3.2 \times 10^6$ trainable parameters for two-layer networks and $2 \times 10^7$ trainable parameters for three-layer networks. Both models are trained using SGD with momentum with learning

**Table A.3.** Details of regularization parameters used for low-frequency noise experiments.

| Dataset | Model | $\ell_2$ regularization grid |
|---|---|---|
| FMNIST | NN depth 2 | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-6,-2]$ |
| | NN depth 3 | $\{10^{\alpha_i}\}_{i=1}^{10}$, $\alpha_i$ uniformly spaced in $[-7,-5]$ |
| | NTK KRR depth 2 | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-1,5]$ |
| | NTK KRR depth 3 | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-4,3]$ |
| | Linear model | $\{10^{\alpha_i}\}_{i=1}^{30}$, $\alpha_i$ uniformly spaced in $[-1,5]$ |
| CIFAR-10 | Myrtle-5 | $\{10^{\alpha_i}\}_{i=1}^{10}$, $\alpha_i$ uniformly spaced in $[-5,-2]$ |
| | KRR (Myrtle-5 NTK) | $\{10^{\alpha_i}\}_{i=1}^{20}$, $\alpha_i$ uniformly spaced in $[-6,1]$ |

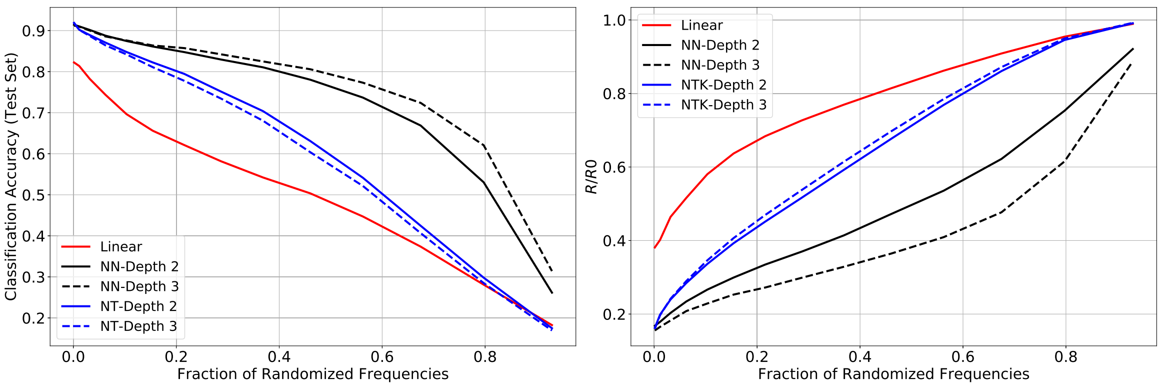

**Figure A.6.** Normalized test squared error (left) and test classification accuracy (right) of the models on FMNIST with low-frequency Gaussian noise.

rate described by (11) (with $lr_0 = 10^{-3}$). For the warm-up epochs, we use batch-size of 500. We increase the batch-size to 1000 after the warm-up stage. The regularization grids used for training our models are presented in table A.3 (figure A.6).

### A.6. Low-frequency noise experiments on CIFAR-10

To test whether our insights are valid for convolutional models, we repeat the same experiment for CNNs trained on CIFAR-10. The noisy data is generated as follows:

(a) Let $\mathcal{D}_{\text{train}}$ denote the set of training images in the DCT-frequency domain. Note that CIFAR-10 images have three channels. To convert the images to frequency domain, we apply two-dimensional discrete cosine transform (DCT-II orthogonal) to each channel separately. We compute the mean $\mu$ and the covariance $\Sigma$ of the elements of $\mathcal{D}_{\text{train}}$.

(b) We fix a threshold $\alpha \in \mathbb{N}$ where $1 \leqslant \alpha \leqslant 32$.

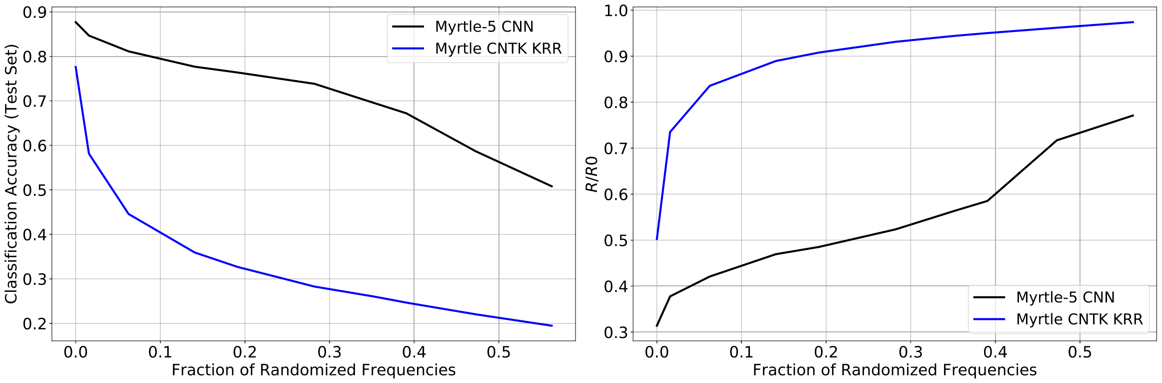

**Figure A.7.** Performance of Myrtle-5 and KRR with CNTK on noisy CIFAR-10. CNTK is generated from the Myrtle-5 architecture using neural-tangents JAX library. When no noise is present in the data, the CNN achieves 87.7% and the CNTK achieves 77.6% classification accuracy. After randomizing only 1.5% of the frequencies (corresponding to $\alpha = 4$) CNTK classification performance falls to 58.2% while the CNN retains 84.7% accuracy.

(c) Let $\boldsymbol{x} \in \mathbb{R}^{32\times32\times3}$ be an image in the dataset (test or train). We denote the representation of $\boldsymbol{x}$ in the DCT-frequency domain with $\tilde{\boldsymbol{x}} \in \mathbb{R}^{32\times32\times3}$. For each image, we draw a noise matrix $\boldsymbol{z} \sim \mathcal{N}(\mu, \Sigma)$. We have

$$[\tilde{\boldsymbol{x}}_{\text{noisy}}]_{i,j,k} = \begin{cases} (\boldsymbol{z})_{i,j,k} & \text{if}\, i, j \leqslant \alpha \\ \tilde{\boldsymbol{x}}_{i,j,k} & \text{otherwise.} \end{cases}$$

(d) We perform IDCT on $\tilde{\boldsymbol{x}}_{\text{noisy}}$ to get the noisy image $\boldsymbol{x}_{\text{noisy}}$.

(e) We normalize the noisy data to have zero per-channel mean and unit per-channel standard deviation. The normalization statistics are computed using only the training data.

We use Myrtle-5 architecture for our analysis. The Myrtle family is a collection of simple light-weight high-performance purely convolutional models. The simplicity of these models coupled with their good performance makes them a natural candidate for our analysis. The network only uses convolutions and average pooling. In particular, we do not use any batch-normalization [42] layers in this network (see [16] for details). We fix the number of channels in all convolutional layers to be $N = 512$. This corresponds to approximately $7 \times 10^6$ parameters. Similar to the fully-connected networks, our convolutional models are also optimized via SGD with 0.9 momentum (learning rate evolves as (11) with $lr_0 = 0.1$ and $T = 70$). We fix the batch-size to 128. To keep the experimental setting as simple as possible, we do not use any data augmentation for training the network. The results of the numerical experiments are reported in figure A.7.

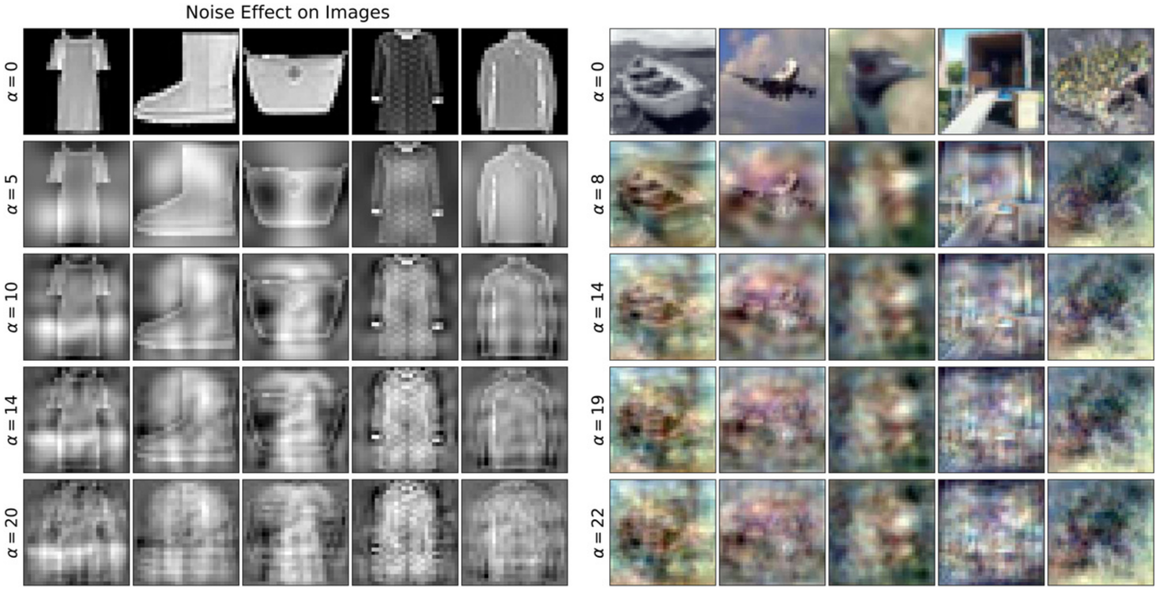

**Figure A.8.** The effect of low-frequency noise for various cut-off thresholds, $\alpha$. The left panel corresponds to the noisy FMNIST images and the right panel corresponds to CIFAR-10 images. In order to plot CIFAR-10 images, we rescale them to the interval $[0, 1]$.

## Appendix B. Technical background on function spaces on the sphere

### B.1. Functional spaces over the sphere

For $d \geqslant 1$, 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 $\mu_{d-1}$ the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in the following are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mu_{d-1})$, with scalar product and norm denoted as $\langle \cdot, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$:

$$\langle f, g \rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x}) g(\boldsymbol{x}) \mu_{d-1}(\mathrm{d}\boldsymbol{x}). \tag{16}$$

For $\ell \in \mathbb{Z}_{\geqslant 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree $\ell$ 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

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mu_{d-1}) = \bigotimes_{\ell=0}^{\infty} V_{d,\ell}. \tag{17}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(d,\ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{\ell-1}. \tag{18}$$

For each $\ell \in \mathbb{Z}_{\geqslant 0}$, the spherical harmonics $\{Y_{\ell,j}^{(d)}\}_{1\leqslant j\in\leqslant B(d,\ell)}$ 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 $\mathsf{P}_k$ the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mu_{d-1})$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{19}$$

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

### B.2. Gegenbauer polynomials

The $\ell$th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree $\ell$. 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 $\{Q_\ell^{(d)}\}_{\ell\geqslant 0}$ forms an orthogonal basis on $L^2([-d,d], \tilde{\mu}_{d-1}^1)$, where $\tilde{\mu}_{d-1}^1$ is the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle$ when $\boldsymbol{x} \sim\ \mu_{d-1}$, satisfying the normalization condition:

$$\begin{aligned}\int_{-d}^{d} Q_k^{(d)}(t) Q_j^{(d)}(t)\mathrm{d}\tilde{\mu}_{d-1}^1 &= \frac{w_{d-2}}{dw_{d-1}}\int_{-d}^{d} Q_k^{(d)}(t) Q_j^{(d)}(t)\left(1-\frac{t^2}{d^2}\right)^{(d-3)/2}\mathrm{d}t \\ &= \frac{1}{B(d,k)}\delta_{jk},\end{aligned} \tag{20}$$

where we denoted $w_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ the surface area of the sphere $\mathbb{S}^{d-1}(1)$. In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$.

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 $V_\ell$ 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.

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

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(d,k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{21}$$

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

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{22}$$

(c) Recurrence formula

$$\frac{t}{d} Q_k^{(d)}(t) = \frac{k}{2k+d-2} Q_{k-1}^{(d)}(t) + \frac{k+d-2}{2k+d-2} Q_{k+1}^{(d)}(t). \tag{23}$$

(d) Rodrigues formula

$$Q_k^{(d)}(t) = (-1/2)^k d^k \frac{\Gamma((d-1)/2)}{\Gamma(k+(d-1)/2)} \left(1 - \frac{t^2}{d^2}\right)^{(3-d)/2} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k \left(1 - \frac{t^2}{d^2}\right)^{k+(d-3)/2}. \tag{24}$$

Note in particular that property (b) implies 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

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) f(\boldsymbol{y}) \mu_{d-1}(\mathrm{d}\boldsymbol{y}). \tag{25}$$

### B.3. Hermite polynomials

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

$$\mathbb{E}\left\{\mathrm{He}_j(G)\mathrm{He}_k(G)\right\} = k!\delta_{jk}. \tag{26}$$

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

$$g(x) = \sum_{k=0}^{\infty} \frac{\mu_k(g)}{k!} \mathrm{He}_k(x), \qquad \mu_k(g) \equiv \mathbb{E}\left\{g(G)\mathrm{He}_k(G)\right\}. \tag{27}$$

Notice that for functions $g$ that are $k$-weakly differentiable with $g^{(k)}$ the $k$th weak derivative, we have

$$\mu_k(g) = \mathbb{E}_G[g^{(k)}(G)]. \tag{28}$$

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

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{\mathrm{d}}x)B(d,k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\mathrm{He}_k(x)\right\}. \tag{29}$$

Here and below, for $P$ a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of $P$.

### B.4. Tensor product of spherical harmonics

We will consider in this paper the product space

$$\mathrm{PS}^{\boldsymbol{d}} \equiv \prod_{q=1}^{Q} \mathbb{S}^{d_q-1}\left(\sqrt{d_q}\right), \tag{30}$$

and the uniform measure on $\mathrm{PS}^{\boldsymbol{d}}$, denoted $\mu_{\boldsymbol{d}} \equiv \mu_{d_1-1}\otimes\ldots\otimes\mu_{d_Q-1} = \bigotimes_{q\in[Q]}\mu_{d_q-1}$, where we recall $\mu_{d_q-1} \equiv \mathrm{Unif}(\mathbb{S}^{d_q-1}(\sqrt{d_q}))$. We consider the functional space of $L^2(\mathrm{PS}^{\boldsymbol{d}},\ \mu_{\boldsymbol{d}})$ with scalar product and norm denoted as $\langle\cdot,\cdot\rangle_{L^2}$ and $\|\cdot\|_{L^2}$:

$$\langle f, g\rangle_{L^2} \equiv \int_{\mathrm{PS}^{\boldsymbol{d}}} f(\overline{\boldsymbol{x}})g(\overline{\boldsymbol{x}})\mu_{\boldsymbol{d}}(\mathrm{d}\overline{\boldsymbol{x}}).$$

For $\boldsymbol{\ell} = (\ell_1,\ldots,\ell_Q) \in \mathbb{Z}^Q_{\geqslant 0}$, let $\tilde{V}^{\boldsymbol{d}}_{\boldsymbol{\ell}} \equiv \tilde{V}_{d_1,\ell_1}\otimes\ldots\otimes\tilde{V}_{d_Q,\ell_Q}$ be the span of tensor products of $Q$ homogeneous harmonic polynomials, respectively of degree $\ell_q$ on $\mathbb{R}^{d_q}$ in variable $\overline{\boldsymbol{x}}_q$. Denote by $V^{\boldsymbol{d}}_{\boldsymbol{\ell}}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}^{\boldsymbol{d}}_{\boldsymbol{\ell}}$ to $\mathrm{PS}^{\boldsymbol{d}}$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathrm{PS}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}}) = \bigoplus_{\boldsymbol{\ell}\in\mathbb{Z}^Q_{\geqslant 0}} V^{\boldsymbol{d}}_{\boldsymbol{\ell}}. \tag{31}$$

The dimension of each subspace is given by

$$B(\boldsymbol{d},\boldsymbol{\ell}) \equiv \dim(V^{\boldsymbol{d}}_{\boldsymbol{\ell}}) = \prod_{q=1}^{Q} B(d_q,\ell_q),$$

where we recall

$$B(d,\ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{\ell-1}.$$

We recall that for each $\ell\in\mathbb{Z}_{\geqslant 0}$, the spherical harmonics $\{Y^{(d)}_{\ell j}\}_{j\in[B(d,\ell)]}$ form an orthonormal basis of $V^{(d)}_\ell$ on $\mathbb{S}^{d-1}(\sqrt{d})$. Similarly, for each $\boldsymbol{\ell}\in\mathbb{Z}^Q_{\geqslant 0}$, the tensor product of spherical harmonics $\{Y^{\boldsymbol{d}}_{\boldsymbol{\ell},\boldsymbol{s}}\}_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{\ell})]}$ form an orthonormal basis of $V^{\boldsymbol{d}}_{\boldsymbol{\ell}}$, where $\boldsymbol{s} = (s_1,\ldots,s_Q) \in [B(\boldsymbol{d},\boldsymbol{\ell})]$ signify $s_q \in [B(d_q,\ell_q)]$ for $q = 1,\ldots,Q$ and

$$Y^{\boldsymbol{d}}_{\boldsymbol{\ell},\boldsymbol{s}} \equiv Y^{(d_1)}_{\ell_1,s_1} \otimes Y^{(d_2)}_{\ell_2,s_2} \otimes \ldots \otimes Y^{(d_Q)}_{\ell_Q,s_Q} = \bigotimes_{q=1}^{Q} Y^{(d_q)}_{\ell_q,s_q}.$$

We have the following orthonormalization property

$$\langle Y^{\boldsymbol{d}}_{\boldsymbol{\ell},\boldsymbol{s}}, Y^{\boldsymbol{d}}_{\boldsymbol{\ell}',\boldsymbol{s}'} \rangle_{L^2} = \prod_{q=1}^{Q} \left\langle Y^{(d_q)}_{\ell_q s_q}, Y^{(d_q)}_{\ell'_q s'_q} \right\rangle_{L^2\left(\mathbb{S}^{d_q-1}(\sqrt{d_q})\right)} = \prod_{q=1}^{Q} \delta_{\ell_q,\ell'_q}\delta_{s_q,s'_q} = \delta_{\boldsymbol{\ell},\boldsymbol{\ell}'}\delta_{\boldsymbol{s},\boldsymbol{s}'}.$$

We denote by $\mathsf{P}_{\boldsymbol{k}}$ the orthogonal projections on $V^{\boldsymbol{d}}_{\boldsymbol{k}}$ in $L^2(\mathrm{PS}^{\boldsymbol{d}},\ \mu_{\boldsymbol{d}})$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_{\boldsymbol{k}} f(\overline{\boldsymbol{x}}) \equiv \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \langle f, Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}} \rangle_{L^2} Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}). \tag{32}$$

We will denote for any $\mathcal{Q} \subset \mathbb{Z}^{Q}_{\geqslant 0}$, $\mathsf{P}_{\mathcal{Q}}$ the orthogonal projection on $\bigoplus_{\boldsymbol{k}\in\mathcal{Q}} V^{\boldsymbol{d}}_{\boldsymbol{k}}$, given by

$$\mathsf{P}_{\mathcal{Q}} = \sum_{\boldsymbol{k}\in\mathcal{Q}} \mathsf{P}_{\boldsymbol{k}}.$$

Similarly, the projection on $\mathcal{Q}^c$, the complementary of the set $\mathcal{Q}$ in $\mathbb{Z}^{Q}_{\geqslant 0}$, is given by

$$\mathsf{P}_{\mathcal{Q}^c} = \sum_{\boldsymbol{k}\notin\mathcal{Q}} \mathsf{P}_{\boldsymbol{k}}.$$

### B.5. Tensor product of Gegenbauer polynomials

We recall that $\tilde{\mu}^1_{d-1}$ denotes the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_d\rangle$ when $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. We consider similarly the projection of $\mathrm{PS}^{\boldsymbol{d}}$ on one coordinate per sphere. We define

$$\mathrm{ps}^{\boldsymbol{d}} \equiv \prod_{q=1}^{Q} [-d_q, d_q], \qquad \tilde{\mu}^1_{\boldsymbol{d}} \equiv \tilde{\mu}^1_{d_1-1} \otimes \ldots \otimes \tilde{\mu}^1_{d_Q-1} = \bigotimes_{q=1}^{Q} \tilde{\mu}^1_{d_q-1}, \tag{33}$$

and consider $L^2(\mathrm{ps}^{\boldsymbol{d}}, \tilde{\mu}^1_{\boldsymbol{d}})$.

Recall that the Gegenbauer polynomials $\{Q^{(d)}_k\}_{k\geqslant 0}$ form an orthogonal basis of $L^2([-d,d], \tilde{\mu}^1_{d-1})$.

Define for each $\boldsymbol{k} \in \mathbb{Z}^{Q}_{\geqslant 0}$, the tensor product of Gegenbauer polynomials

$$Q^{\boldsymbol{d}}_{\boldsymbol{k}} \equiv Q^{(d_1)}_{k_1} \otimes \ldots \otimes Q^{(d_q)}_{k_Q} = \bigotimes_{q=1}^{Q} Q^{(d_q)}_{k_q}. \tag{34}$$

We will use the following properties of the tensor product of Gegenbauer polynomials:

**Lemma 1 (Properties of products of Gegenbauer).** *Consider the tensor product of Gegenbauer polynomials* $\{Q^{\boldsymbol{d}}_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}}$ *defined in equation* (34)*. Then*

*(a) The set* $\{Q^{\boldsymbol{d}}_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}}$ *forms an orthogonal basis on* $L^2(\mathrm{ps}^{\boldsymbol{d}}, \tilde{\mu}^1_{\boldsymbol{d}})$*, satisfying the normalization condition: for any* $\boldsymbol{k}, \boldsymbol{k}' \in \mathbb{Z}^Q_{\geqslant 0}$,

$$\left\langle Q^{\boldsymbol{d}}_{\boldsymbol{k}}, Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\right\rangle_{L^2(\mathrm{ps}^{\boldsymbol{d}})} = \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k},\boldsymbol{k}'}. \tag{35}$$

*(b) For* $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(Q)})$ *and* $\overline{\boldsymbol{y}} = (\overline{\boldsymbol{y}}^{(1)}, \dots, \overline{\boldsymbol{y}}^{(Q)}) \in \mathrm{PS}^{\boldsymbol{d}}$*, and* $\boldsymbol{k}, \boldsymbol{k}' \in \mathbb{Z}^Q_{\geqslant 0}$,

$$\left\langle Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \cdot\rangle\}_{q\in[Q]}\right), Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{y}}^{(q)}, \cdot\rangle\}_{q\in[Q]}\right)\right\rangle_{L^2\left(\mathrm{PS}^{\boldsymbol{d}}\right)} = \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k},\boldsymbol{k}'}Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right). \tag{36}$$

*(c) For* $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(Q)})$ *and* $\overline{\boldsymbol{y}} = (\overline{\boldsymbol{y}}^{(1)}, \dots, \overline{\boldsymbol{y}}^{(Q)}) \in \mathrm{PS}^{\boldsymbol{d}}$*, and* $\boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0}$,

$$Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right) = \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}\sum_{\boldsymbol{s}\in B(\boldsymbol{d},\boldsymbol{k})} Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}})Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{y}}). \tag{37}$$

Notice that lemma 1(c) implies that $Q^{\boldsymbol{d}}_{\boldsymbol{k}}$ is (up to a constant) a representation of the projector onto the subspace $V^{\boldsymbol{d}}_{\boldsymbol{k}}$

$$[\mathsf{P}_{\boldsymbol{k}}f](\overline{\boldsymbol{x}}) = B(\boldsymbol{d},\boldsymbol{k})\int_{\mathrm{PS}^{\boldsymbol{d}}} Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right) f(\overline{\boldsymbol{y}})\mu_{\boldsymbol{d}}(\mathrm{d}\overline{\boldsymbol{y}}).$$

**Proof of lemma 1.** Part (a) comes from the normalization property (20) of Gegenbauer polynomials,

$$\begin{aligned}
\left\langle Q^{\boldsymbol{d}}_{\boldsymbol{k}}, Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\right\rangle_{L^2(\mathrm{ps}^{\boldsymbol{d}})} &= \left\langle Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\sqrt{d_q}\langle \boldsymbol{e}_q, \cdot\rangle\}_{q\in[Q]}\right), Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\sqrt{d_q}\langle \boldsymbol{e}_q, \cdot\rangle\}_{q\in[Q]}\right)\right\rangle_{L^2\left(\mathrm{PS}^{\boldsymbol{d}}\right)} \\
&= \prod_{q=1}^{Q}\left\langle Q^{(d_q)}_{k_q}\left(\sqrt{d_q}\langle \boldsymbol{e}_q, \cdot\rangle\right), Q^{(d_q)}_{k'_q}\left(\sqrt{d_q}\langle \boldsymbol{e}_q, \cdot\rangle\right)\right\rangle_{L^2\left(\mathbb{S}^{d_q-1}\left(\sqrt{d_q}\right)\right)} \\
&= \prod_{q=1}^{Q}\frac{1}{B(d_q,k_q)}\delta_{k_q,k'_q} \\
&= \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k},\boldsymbol{k}'},
\end{aligned}$$

where the $\{\boldsymbol{e}_q\}_{q\in[Q]}$ are unit vectors in $\mathbb{R}^{d_q}$ respectively.

Part (b) comes from equation (21),

$$\begin{aligned}
&\left\langle Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \cdot\rangle\}_{q\in[Q]}\right), Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{y}}^{(q)}, \cdot\rangle\}_{q\in[Q]}\right)\right\rangle_{L^2\left(\mathrm{PS}^{\boldsymbol{d}}\right)}\\
&\quad= \prod_{q=1}^{Q}\left\langle Q^{(d_q)}_{k_q}\left(\langle \overline{\boldsymbol{x}}^{(q)}, \cdot\rangle\right), Q^{(d_q)}_{k'_q}\left(\langle \overline{\boldsymbol{y}}^{(q)}, \cdot\rangle\right)\right\rangle_{L^2\left(\mathbb{S}^{d_q-1}\left(\sqrt{d_q}\right)\right)}\\
&\quad= \prod_{q=1}^{Q}\frac{1}{B(d_q,k_q)}\delta_{k_q,k'_q}Q^{(d_q)}_{k_q}\left(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\right)\\
&\quad= \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k},\boldsymbol{k}'}Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right),
\end{aligned}$$

while part $(c)$ is a direct consequence of equation (22). □

### B.6. Notations

Throughout the proofs, $O_d(\cdot)$ (resp. $o_d(\cdot)$) denotes the standard big-O (resp. little-o) notation, where the subscript $d$ emphasizes the asymptotic variable. We denote $O_{d,\mathbb{P}}(\cdot)$ (resp. $o_{d,\mathbb{P}}(\cdot)$) the big-O (resp. little-o) in probability notation: $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}(|h_1(d)/h_2(d)| > C_\varepsilon) \leqslant \varepsilon, \quad \forall\, d \geqslant d_\varepsilon,$$

and respectively: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability.

We will occasionally hide logarithmic factors using the $\tilde{O}_d(\cdot)$ notation (resp. $\tilde{o}_d(\cdot)$): $h_1(d) = \tilde{O}_d(h_2(d))$ if there exists a constant $C$ such that $h_1(d) \leqslant C(\log d)^C h_2(d)$. Similarly, we will denote $\tilde{O}_{d,\mathbb{P}}(\cdot)$ (resp. $\tilde{o}_{d,\mathbb{P}}(\cdot)$) when considering the big-O in probability notation up to a logarithmic factor.

Furthermore, $f = \omega_d(g)$ will denote $f(d)/g(d) \to \infty$.

## Appendix C. General framework and main theorems

In this section, define a more general model than the model considered in the main text. In the general model, we will assume the covariate vectors will follow a product of uniform distributions on the sphere, and assume a target function in $L^2$ space. We establish more general versions of theorems 1–3 on the two-spheres cases in the main text as theorems 5–7. We will prove theorems 5–7 in the following sections. At the end of this section, we will show that theorems 5–7 will imply theorems 1–3 in the main text.

## C.1. Setup on the product of spheres

Assume that the data $\boldsymbol{x}$ lies on the product of $Q$ spheres,

$$\boldsymbol{x} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(Q)}\right) \in \prod_{q\in[Q]} \mathbb{S}^{d_q-1}(r_q),$$

where $d_q = d^{\eta_q}$ and $r_q = d^{(\eta_q+\kappa_q)/2}$. Let $\boldsymbol{d} = (d_1, \ldots, d_q) = (d^{\eta_1}, \ldots, d^{\eta_q})$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_Q)$, where $\eta_q > 0$ and $\kappa_q \geqslant 0$ for $q = 1, \ldots, Q$. We will denote this space

$$\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}} = \prod_{q\in[Q]} \mathbb{S}^{d_q-1}(r_q). \tag{38}$$

Furthermore, assume that the data is generated following the uniform distribution on $\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}}$, i.e.

$$\boldsymbol{x} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}}) = \bigotimes_{q\in[Q]} \mathrm{Unif}\left(\mathbb{S}^{d_q-1}(r_q)\right) \equiv \mu_{\boldsymbol{d}}^{\boldsymbol{\kappa}}. \tag{39}$$

We have $\boldsymbol{x} \in \mathbb{R}^D$ and $\|\boldsymbol{x}\|_2 = R$ where $D = d^{\eta_1} + \cdots + d^{\eta_Q}$ and $R = (d^{\eta_1+\kappa_1} + \cdots + d^{\eta_Q+\kappa_Q})^{1/2}$.

We will make the following assumption that will simplify the proofs. Denote

$$\xi \equiv \max_{q\in[Q]}\{\eta_q + \kappa_q\}, \tag{40}$$

then $\xi$ is attained on only one of the sphere, whose coordinate will be denoted $q_\xi$, i.e. $\xi = \eta_{q_\xi} + \kappa_{q_\xi}$ and $\eta_q + \kappa_q < \xi$ for $q \neq q_\xi$.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function and $(\boldsymbol{w}_i)_{i\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{D-1})$ the weights. We introduce the random feature function class

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{W}) = \left\{ \hat{f}_{\mathrm{RF}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle \sqrt{D}/R) : a_i \in \mathbb{R}, \quad \forall\, i \in [N] \right\},$$

and the neural tangent function class

$$\mathcal{F}_{\mathrm{NT}}(\boldsymbol{W}) = \left\{ \hat{f}_{\mathrm{RF}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle \sqrt{D}/R) : \boldsymbol{a}_i \in \mathbb{R}^D, \quad \forall\, i \in [N] \right\}.$$

We will denote $\boldsymbol{\theta}_i = \sqrt{D}\boldsymbol{w}_i$. Notice that the normalization in the definition of the function class insures that the scalar product $\langle \boldsymbol{x}, \boldsymbol{\theta}_i\rangle/R$ is of order 1. This corresponds to normalizing the data.

We consider the approximation of $f$ by functions in function classes $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ and $\mathcal{F}_{\mathrm{NT}}(\boldsymbol{\Theta})$.

## C.2. Reparametrization

Recall $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ independently. We decompose $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^{(1)}, \ldots, \boldsymbol{\theta}_i^{(Q)})$ into $Q$ sections corresponding to the $d_q$ coordinates associated to the $q$th sphere. Let us consider the following reparametrization of $(\boldsymbol{\theta}_i)_{i\in[N]} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$:

$$\left(\overline{\boldsymbol{\theta}}_i^{(1)}, \dots, \overline{\boldsymbol{\theta}}_i^{(Q)}, \tau_i^{(1)}, \dots, \tau_i^{(Q)}\right),$$

where

$$\overline{\boldsymbol{\theta}}_i^{(q)} \equiv \sqrt{d_q}\boldsymbol{\theta}_i^{(q)}/\|\boldsymbol{\theta}_i^{(q)}\|_2, \qquad \tau_i^{(q)} \equiv \|\boldsymbol{\theta}_i^{(q)}\|_2/\sqrt{d_q}, \quad \text{for } q = 1, \dots, Q.$$

Hence

$$\boldsymbol{\theta}_i = \left(\tau_i^{(1)} \cdot \overline{\boldsymbol{\theta}}_i^{(1)}, \dots, \tau_i^{(Q)} \cdot \overline{\boldsymbol{\theta}}_i^{(Q)}\right).$$

It is easy to check that the variables $(\overline{\boldsymbol{\theta}}^{(1)}, \dots, \overline{\boldsymbol{\theta}}^{(Q)})$ are independent and independent of $(\tau_i^{(1)}, \dots, \tau_i^{(Q)})$, and verify

$$\overline{\boldsymbol{\theta}}_i^{(q)} \sim \text{Unif}(\mathbb{S}^{d_q-1}(\sqrt{d_q})), \qquad \tau_i^{(q)} \sim d_q^{-1/2}\sqrt{\text{Beta}\left(\frac{d_q}{2}, \frac{D-d_q}{2}\right)}, \quad \text{for } q = 1, \dots, Q.$$

We will denote $\overline{\boldsymbol{\theta}}_i \equiv (\overline{\boldsymbol{\theta}}_i^{(1)}, \dots, \overline{\boldsymbol{\theta}}_i^{(Q)})$ and $\boldsymbol{\tau}_i \equiv (\tau_i^{(1)}, \dots, \tau_i^{(Q)})$. With these notations, we have

$$\overline{\boldsymbol{\theta}}_i \in \prod_{q\in[Q]} \mathbb{S}^{d_q-1}(\sqrt{d_q}) \equiv \text{PS}^{\boldsymbol{d}},$$

where $\text{PS}^{\boldsymbol{d}}$ is the ‘normalized space of product of spheres’, and

$$(\overline{\boldsymbol{\theta}}_i)_{i\in[N]} \overset{\text{i.i.d.}}{\sim} \bigotimes_{q\in[Q]} \text{Unif}(\mathbb{S}^{d_q-1}(\sqrt{d_q})) \equiv \mu_{\boldsymbol{d}}.$$

Similarly, we will denote the rescaled data $\overline{\boldsymbol{x}} \in \text{PS}^{\boldsymbol{d}}$,

$$\overline{\boldsymbol{x}} = \left(\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(Q)}\right) \sim \bigotimes_{q\in[Q]} \text{Unif}(\mathbb{S}^{d_q-1}(\sqrt{d_q})),$$

obtained by taking $\overline{\boldsymbol{x}}^{(q)} = \sqrt{d_q}\boldsymbol{x}^{(q)}/r_q = d^{-\kappa_q/2}\boldsymbol{x}^{(q)}$ for each $q \in [Q]$.

The proof will proceed as follows: first, noticing that $\tau^{(q)}$ concentrates around 1 for every $q = 1, \dots, Q$, we will restrict ourselves without loss of generality to the following high probability event

$$\mathcal{P}_{d,N,\varepsilon} \equiv \left\{\boldsymbol{\Theta}|\tau_i^{(q)} \in [1-\varepsilon, 1+\varepsilon], \quad \forall\, i \in [N],\ \forall\, q \in [Q]\right\} \subset \mathbb{S}^{D-1}(\sqrt{D})^N,$$

where $\varepsilon > 0$ will be chosen sufficiently small. Then, we rewrite the activation function

$$\sigma(\langle\cdot,\cdot\rangle/R) : \mathbb{S}^{D-1}(\sqrt{D}) \times \text{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}} \to \mathbb{R},$$

as a function, for a random $\boldsymbol{\tau}$ (but close to $(1, \dots, 1)$)

$$\sigma_{\boldsymbol{d},\boldsymbol{\tau}} : \text{PS}^{\boldsymbol{d}} \times \text{PS}^{\boldsymbol{d}} \to \mathbb{R},$$

given for $\boldsymbol{\theta} = (\overline{\boldsymbol{\theta}}, \boldsymbol{\tau})$ by

$$\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) = \sigma\left(\sum_{q\in[Q]} \frac{\tau^{(q)} r_q}{R} \cdot \frac{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle}{\sqrt{d_q}}\right).$$

We can therefore apply the algebra of tensor product of spherical harmonics and use the machinery developed in [21].

### C.3. Notations

Recall the definitions $\boldsymbol{d} = (d_1, \ldots, d_q)$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_Q)$, $d_q = d^{\eta_q}$, $r_q = d^{(\eta_q+\kappa_q)/2}$, $D = d^{\eta_1} + \cdots + d^{\eta_Q}$ and $R = (d^{\eta_1+\kappa_1} + \cdots + d^{\eta_Q+\kappa_Q})^{1/2}$. Let us denote $\xi = \max_{q\in[Q]}\{\eta_q + \kappa_q\}$ and $q_\xi = \mathrm{argmin}_{q\in[Q]}\{\eta_q + \kappa_q\}$.

Recall that $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ independently. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$. We denote $\mathbb{E}_{\boldsymbol{\theta}}$ to be the expectation operator with respect to $\boldsymbol{\theta} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ and $\mathbb{E}_{\boldsymbol{\Theta}}$ the expectation operator with respect to $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))^{\otimes N}$.

We will denote $\mathbb{E}_{\overline{\boldsymbol{\theta}}}$ the expectation operator with respect to $\overline{\boldsymbol{\theta}} \equiv (\overline{\boldsymbol{\theta}}^{(1)}, \ldots, \overline{\boldsymbol{\theta}}^{(Q)}) \sim \mu_{\boldsymbol{d}}$, $\mathbb{E}_{\overline{\boldsymbol{\Theta}}}$ the expectation operator with respect to $\overline{\boldsymbol{\Theta}} = (\overline{\boldsymbol{\theta}}_1, \ldots, \overline{\boldsymbol{\theta}}_N)$, and $\mathbb{E}_{\boldsymbol{\tau}}$ the expectation operator with respect to $\boldsymbol{\tau}$ (we recall $\boldsymbol{\tau} \equiv (\tau^{(1)}, \ldots, \tau^{(Q)})$) or $(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_N)$ (where the $\boldsymbol{\tau}_i$ are independent) depending on the context. In particular, notice that $\mathbb{E}_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\tau}}\mathbb{E}_{\overline{\boldsymbol{\theta}}}$ and $\mathbb{E}_{\boldsymbol{\Theta}} = \mathbb{E}_{\boldsymbol{\tau}}\mathbb{E}_{\overline{\boldsymbol{\Theta}}}$.

We will denote $\mathbb{E}_{\boldsymbol{\Theta}_\varepsilon}$ the expectation operator with respect to $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ restricted to $\mathcal{P}_{d,N,\varepsilon}$ and $\mathbb{E}_{\boldsymbol{\tau}_\varepsilon}$ the expectation operator with respect to $\boldsymbol{\tau}$ restricted to $[1-\varepsilon, 1+\varepsilon]^Q$. Notice that $\mathbb{E}_{\boldsymbol{\Theta}_\varepsilon} = \mathbb{E}_{\boldsymbol{\tau}_\varepsilon}\mathbb{E}_{\overline{\boldsymbol{\Theta}}}$.

Let $\mathbb{E}_{\boldsymbol{x}}$ to be the expectation operator with respect to $\boldsymbol{x} \sim \mu_{\boldsymbol{\kappa}}^{\boldsymbol{d}}$, and $\mathbb{E}_{\overline{\boldsymbol{x}}}$ the expectation operator with respect to $\overline{\boldsymbol{x}} \sim \mu_{\boldsymbol{d}}$.

### C.4. Generalization error of kernel ridge regression

We consider the KRR solution $\hat{a}_i$, namely

$$\hat{\boldsymbol{a}} = (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{y},$$

where the kernel matrix $\boldsymbol{H} = (H_{ij})_{ij\in[n]}$ is assumed to be given by

$$H_{ij} = \bar{h}_d(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/R^2) = \mathbb{E}_{\overline{\boldsymbol{\theta}}\sim\mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})}[\sigma(\langle \overline{\boldsymbol{\theta}}, \boldsymbol{x}\rangle/R)\sigma(\langle \overline{\boldsymbol{\theta}}, \boldsymbol{y}\rangle/R)],$$

and $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, with

$$\boldsymbol{f} = (f_d(\boldsymbol{x}_1), \ldots, f_d(\boldsymbol{x}_n))^\mathsf{T},$$
$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}.$$

The prediction function at location $\boldsymbol{x}$ gives

$$\hat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{y}^\mathsf{T}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{h}(\boldsymbol{x}),$$

with

$$\boldsymbol{h}(\boldsymbol{x}) = [\bar{h}_d(\langle \boldsymbol{x}, \boldsymbol{x}_1 \rangle / R^2), \dots, \bar{h}_d(\langle \boldsymbol{x}, \boldsymbol{x}_n \rangle / R^2)]^{\mathsf{T}}.$$

The test error of empirical KRR is defined as

$$R_{\mathrm{KRR}}(f_d, \boldsymbol{X}, \lambda) \equiv \mathbb{E}_{\boldsymbol{x}} \left[ \left( f_d(\boldsymbol{x}) - \boldsymbol{y}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{h}(\boldsymbol{x}) \right)^2 \right].$$

We define the set $\overline{\mathcal{Q}}_{\mathrm{KRR}}(\gamma) \subseteq \mathbb{Z}_{\geqslant 0}^Q$ as follows (recall that $\xi \equiv \max_{q \in [Q]} (\eta_q + \kappa_q)$):

$$\overline{\mathcal{Q}}_{\mathrm{KRR}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q \,\middle|\, \sum_{q=1}^{Q} (\xi - \kappa_q) k_q \leqslant \gamma \right\}, \tag{41}$$

and the function $m : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ which at $\gamma$ associates

$$m(\gamma) = \min_{\boldsymbol{k} \notin \overline{\mathcal{Q}}_{\mathrm{KRR}}(\gamma)} \sum_{q \in [Q]} (\xi - \kappa_q) k_q.$$

Notice that by definition $m(\gamma) > \gamma$.

We consider sequences of problems indexed by the integer $d$, and we view the problem parameters (in particular, the dimensions $d_q$, the radii $r_q$, the kernel $h_d$, and so on) as functions of $d$.

**Assumption 1.** Let $\{h_d\}_{d \geqslant 1}$ be a sequence of functions $h_d : [-1, 1] \to \mathbb{R}$ such that $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d)$ is a positive semidefinite kernel.

(a) For $\gamma > 0$ (which is specified in the theorem), we denote $L = \max_{q \in [Q]} \lceil \gamma / \eta_q \rceil$. We assume that $h_d$ is $L$-weakly differentiable. We assume that for $0 \leqslant k \leqslant L$, the $k$th weak derivative verifies almost surely $h_d^{(k)}(u) \leqslant C$ for some constants $C > 0$ independent of $d$. Furthermore, we assume there exists $k > L$ such that $h_d^{(k)}(0) \geqslant c > 0$ with $c$ independent of $d$.

(b) For $\gamma > 0$ (which is specified in the theorem), we define

$$\overline{K} = \max_{\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{KRR}}(\gamma)} |\boldsymbol{k}|.$$

We assume that $\sigma$ verifies for $k \leqslant \overline{K}$, $h_d^{(k)}(0) \geqslant c$, with $c > 0$ independent of $d$.

**Theorem 5 (Risk of the KRR model).** *Let $\{f_d \in L^2(\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}}^{\boldsymbol{\kappa}})\}_{d \geqslant 1}$ be a sequence of functions. Assume $w_d(d^\gamma \log d) \leqslant n \leqslant O_d(d^{m(\gamma) - \delta})$ for some $\gamma > 0$ and $\delta > 0$. Let $\{h_d\}_{d \geqslant 1}$ be a sequence of functions that satisfies assumption 1 at level $\gamma$. Let $\boldsymbol{X} = (\boldsymbol{x}_i)_{i \in [n]}$ with $(\boldsymbol{x}_i)_{i \in [n]} \sim \mathrm{Unif}(\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}})$ independently, and $y_i = f_d(\mathbf{x}_i) + \varepsilon_i$ and $\varepsilon_i \sim_{iid} N(0, \tau^2)$ for some $\tau^2 \geqslant 0$. Then for any $\varepsilon > 0$, and for any $\lambda = O_d(1)$, with high probability we have*

$$\left| R_{\mathrm{KRR}}(f_d, \boldsymbol{X}, \lambda) - \| \mathsf{P}_{\mathcal{Q}^c} f_d \|_{L^2}^2 \right| \leqslant \varepsilon (\| f_d \|_{L^2}^2 + \tau^2). \tag{42}$$

See appendix D for the proof of this theorem.

### C.5. Approximation error of the random features model

We consider the minimum population error for the random features model

$$R_{\mathrm{RF}}(f_d, \boldsymbol{W}) = \inf_{f \in \mathcal{F}_{\mathrm{RF}}(\boldsymbol{W})} \mathbb{E}\left[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

Let us define the sets:

$$\mathcal{Q}_{\mathrm{RF}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q} \middle| \sum_{q=1}^{Q} (\xi - \kappa_q) k_q < \gamma \right\}, \tag{43}$$

$$\overline{\mathcal{Q}}_{\mathrm{RF}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q} \middle| \sum_{q=1}^{Q} (\xi - \kappa_q) k_q \leqslant \gamma \right\}. \tag{44}$$

**Assumption 2.** Let $\sigma$ be an activation function.

(a) There exists constants $c_0, c_1$, with $c_0 > 0$ and $c_1 < 1$ such that the activation function $\sigma$ verifies $\sigma(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ almost surely for $u \in \mathbb{R}$.

(b) For $\gamma > 0$ (which is specified in the theorem), we denote $L = \max_{q\in[Q]} \lceil \gamma/\eta_q \rceil$. We assume that $\sigma$ is $L$-weakly differentiable. Define

$$K = \min_{\boldsymbol{k} \in \mathcal{Q}_{\mathrm{RF}}(\gamma)^c} |\boldsymbol{k}|.$$

We assume that for $K \leqslant k \leqslant L$, the $k$th weak derivative verifies almost surely $\sigma^{(k)}(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.

Furthermore we will assume that $\sigma$ is not a degree-$\lfloor \gamma/\eta_{q_\xi} \rfloor$ polynomial where we recall that $q_\xi$ corresponds to the unique $\mathrm{argmin}_{q\in[Q]}\{\eta_q + \kappa_q\}$.

(c) For $\gamma > 0$ (which is specified in the theorem), we define

$$\overline{K} = \max_{\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{RF}}(\gamma)} |\boldsymbol{k}|.$$

We assume that $\sigma$ verifies for $k \leqslant \overline{K}$, $\mu_k(\sigma) \neq 0$. Furthermore we assume that for $k \leqslant \overline{K}$, the $k$th weak derivative verifies almost surely $\sigma^{(k)}(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.

Assumption 2(a) implies that $\sigma \in L^2(\mathbb{R}, \gamma)$ where $\gamma(\mathrm{d}x) = \mathrm{e}^{-x^2/2}\,\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure. We recall the Hermite decomposition of $\sigma$,

$$\sigma(x) = \sum_{k=0}^{\infty} \frac{\mu_k(\sigma)}{k!} \mathrm{He}_k(x), \qquad \mu_k(\sigma) \equiv \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)\mathrm{He}_k(G)]. \tag{45}$$

**Theorem 6 (Risk of the RF model).** *Let $\{f_d \in L^2(\mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}}^{\boldsymbol{\kappa}})\}_{d\geqslant 1}$ be a sequence of functions. Let $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\in[N]}$ with $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ independently. We have the following results.*

*(a) Assume $N \leqslant o_d(d^\gamma)$ for a fixed $\gamma > 0$. Let $\sigma$ satisfy assumptions 2(a) and (b) at level $\gamma$. Then, for any $\varepsilon > 0$, the following holds with high probability:*

$$\left| R_{\mathrm{RF}}(f_d, \boldsymbol{W}) - R_{\mathrm{RF}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) - \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2 \right| \leqslant \varepsilon \|f_d\|_{L^2} \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}, \tag{46}$$

*where $\mathcal{Q} \equiv \mathcal{Q}_{\mathrm{RF}}(\gamma)$ is defined in equation (43).*

*(b) Assume $N \geqslant w_d(d^\gamma)$ for some positive constant $\gamma > 0$, and $\sigma$ satisfy assumptions 2(a) and (c) at level $\gamma$. Then for any $\varepsilon > 0$, the following holds with high probability:*

$$0 \leqslant R_{\mathrm{RF}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) \leqslant \varepsilon \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2, \tag{47}$$

*where $\mathcal{Q} \equiv \overline{\mathcal{Q}}_{\mathrm{RF}}(\gamma)$ is defined in equation (44).*

See appendix E for the proof of the lower bound (46), and appendix F for the proof of the upper bound (47).

**Remark 1.** This theorems shows that for each $\gamma \notin (\xi - \kappa_1)\mathbb{Z}_{\geqslant 0} + \cdots + (\xi - \kappa_Q)\mathbb{Z}_{\geqslant 0}$, we can decompose our functional space as

$$L^2(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}, \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}}) = \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) \oplus \mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma),$$

where

$$\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) = \bigoplus_{\boldsymbol{k} \in \mathcal{Q}_{\mathrm{RF}}(\gamma)} \boldsymbol{V}^{\boldsymbol{d}}_{\boldsymbol{k}},$$

$$\mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) = \bigoplus_{\boldsymbol{k} \notin \mathcal{Q}_{\mathrm{RF}}(\gamma)} \boldsymbol{V}^{\boldsymbol{d}}_{\boldsymbol{k}},$$

such that for $N = d^\gamma$, RF model fits the subspace of low degree polynomials $\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma)$ and cannot fit $\mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma)$, i.e.

$$R_{\mathrm{RF}}(f_d, \boldsymbol{W}) \approx \|\mathsf{P}_{\mathcal{Q}_{\mathrm{RF}}(\gamma)^c} f_d\|_{L^2}^2.$$

**Remark 2.** In other words, we can fit a polynomial of degree $\boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0}$, if and only if

$$d^{(\xi-\kappa_1)k_1} \cdot \ldots \cdot d^{(\xi-\kappa_Q)k_Q} = d_{1,\mathrm{eff}}^{k_1} \ldots d_{Q,\mathrm{eff}}^{k_Q} = o_d(N).$$

Each subspace has therefore an effective dimension $d_{q,\mathrm{eff}} \equiv d^{\xi-\kappa_q} = d_q^{(\xi-\kappa_q)/\eta_q} \asymp D^{(\xi-\kappa_q)/\max_{q\in[Q]}\eta_q}$. This can be understood intuitively as follows,

$$\sigma\left(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle / R\right) = \sigma\left(\sum_{q\in[Q]} \langle \boldsymbol{\theta}^{(q)}, \boldsymbol{x}^{(q)}\rangle / R\right).$$

The term $q_\xi$ (recall that $q_\xi = \mathrm{argmax}_q(\eta_q + \kappa_q)$ and $\xi = \eta_{q_\xi} + \kappa_{q_\xi}$) verifies $\langle \boldsymbol{\theta}^{(q_\xi)}, \boldsymbol{x}^{(q_\xi)}\rangle / R = \Theta_d(1)$ and has the same effective dimension $d_{q_\xi,\mathrm{eff}} = d^{\eta_{q_\xi}}$ has in the uniform case restricted to the sphere $\mathbb{S}^{d^{\eta_q}-1}(\sqrt{d^{\eta_q}})$ (the scaling of the sphere do not matter because of the global normalization factor $R^{-1}$). However, for $\eta_q + \kappa_q < \xi$, we have $\langle \boldsymbol{\theta}^{(q)}, \boldsymbol{x}^{(q)} \rangle / R = \Theta_d(d^{(\eta_q+\kappa_q-\xi)/2})$ and we will need $d^{\xi-\kappa_q-\eta_q}$ more neurons to capture the dependency on the $q$th sphere coordinates. The effective dimension is therefore given by $d_{q,\mathrm{eff}} = d_q \cdot d^{\xi-\kappa_q-\eta_q} = d^{\xi-\kappa_q}$.

### C.6. Approximation error of the neural tangent model

We consider the minimum population error for the random features model

$$R_{\mathrm{NT}}(f_d, \boldsymbol{W}) = \inf_{f \in \mathcal{F}_{\mathrm{NT}}(\boldsymbol{W})} \mathbb{E}\left[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

For $\boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0}$, we denote by $S(\boldsymbol{k}) \subseteq [Q]$ the subset of indices $q \in [Q]$ such that $k_q > 0$.

We define the sets

$$\mathcal{Q}_{\mathrm{NT}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0} \middle| \sum_{q=1}^{Q} (\xi - \kappa_q) k_q < \gamma + \left( \xi - \min_{q \in S(\boldsymbol{k})} \kappa_q \right) \right\}, \tag{48}$$

$$\overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0} \middle| \sum_{q=1}^{Q} (\xi - \kappa_q) k_q \leqslant \gamma + \left( \xi - \min_{q \in S(\boldsymbol{k})} \kappa_q \right) \right\}. \tag{49}$$

**Assumption 3.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function.

(a) The activation function $\sigma$ is weakly differentiable with weak derivative $\sigma'$. There exists constants $c_0, c_1$, with $c_0 > 0$ and $c_1 < 1$ such that the activation function $\sigma$ verifies $\sigma'(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ almost surely for $u \in \mathbb{R}$.

(b) For $\gamma > 0$ (which is specified in the theorem), we denote $L = \max_{q \in [Q]} \lceil \gamma / \eta_q \rceil$. We assume that $\sigma'$ is $L$-weakly differentiable. Define

$$K = \min_{\boldsymbol{k} \in \mathcal{Q}_{\mathrm{NT}}(\gamma)^c} |\boldsymbol{k}|.$$

We assume that for $K - 1 \leqslant k \leqslant L$, the $k$th weak derivative verifies almost surely $\sigma^{(k+1)}(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.

Furthermore, we assume that $\sigma'$ verifies a non-degeneracy condition. Recall that $\mu_k(h) \equiv \mathbb{E}_{G \sim \mathsf{N}(0,1)}[h(G)\mathrm{He}_k(G)]$ denote the $k$th coefficient of the Hermite expansion of $h \in L_2(\mathbb{R}, \gamma)$ (with $\gamma$ the standard Gaussian measure). Then there exists $k_1, k_2 \geqslant 2L + 7[\max_{q \in [Q]} \xi / \eta_q]$ such that $\mu_{k_1}(\sigma'), \mu_{k_2}(\sigma') \neq 0$ and

$$\frac{\mu_{k_1}(x^2 \sigma')}{\mu_{k_1}(\sigma')} \neq \frac{\mu_{k_2}(x^2 \sigma')}{\mu_{k_2}(\sigma')}. \tag{50}$$

(c) For $\gamma > 0$ (which is specified in the theorem), we define

$$\overline{K} = \max_{\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma)} |\boldsymbol{k}|.$$

We assume that $\sigma$ verifies for $k \leqslant \overline{K} + 1$, $\mu_k(\sigma') = \mu_{k+1}(\sigma) \neq 0$. Furthermore we assume that for $k \leqslant \overline{K} + 1$, the $k$th weak derivative verifies almost surely $\sigma^{(k+1)}(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$.

Assumption 3(a) implies that $\sigma' \in L^2(\mathbb{R}, \gamma)$ where $\gamma(\mathrm{d}x) = \mathrm{e}^{-x^2/2}\,\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure. We recall the Hermite decomposition of $\sigma'$:

$$\sigma'(x) = \sum_{k=0}^{\infty} \frac{\mu_k(\sigma')}{k!}\mathrm{He}_k(x), \qquad \mu_k(\sigma') \equiv \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma'(G)\mathrm{He}_k(G)]. \tag{51}$$

In assumption 3(b), it is useful to notice that the Hermite coefficients of $x^2\sigma'(x)$ can be computed from the ones of $\sigma'(x)$ using the relation $\mu_k(x^2\sigma') = \mu_{k+2}(\sigma') + [1 + 2k]\mu_k(\sigma') + k(k-1)\mu_{k-2}(\sigma')$.

**Theorem 7 (Risk of the NT model).** *Let $\{f_d \in L^2(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}, \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}})\}_{d\geqslant 1}$ be a sequence of functions. Let $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\in[N]}$ with $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ independently. We have the following results.*

*(a) Assume $N \leqslant o_d(d^\gamma)$ for a fixed $\gamma > 0$. Let $\sigma$ satisfy assumptions 3(a) and (b) at level $\gamma$. Then, for any $\varepsilon > 0$, the following holds with high probability:*

$$\left| R_{\mathrm{NT}}(f_d, \boldsymbol{W}) - R_{\mathrm{NT}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) - \|\mathsf{P}_{\mathcal{Q}^c} f_d\|^2_{L^2} \right| \leqslant \varepsilon \|f_d\|_{L^2} \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}, \tag{52}$$

*where $\mathcal{Q} \equiv \mathcal{Q}_{\mathrm{NT}}(\gamma)$ is defined in equation (48).*

*(b) Assume $N \geqslant w_d(d^\gamma)$ for some positive constant $\gamma > 0$, and $\sigma$ satisfy assumptions 3(a) and (c) at level $\gamma$. Then for any $\varepsilon > 0$, the following holds with high probability:*

$$0 \leqslant R_{\mathrm{NT}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) \leqslant \varepsilon \|\mathsf{P}_{\mathcal{Q}} f_d\|^2_{L^2}, \tag{53}$$

*where $\mathcal{Q} \equiv \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma)$ is defined in equation (49).*

See appendix G for the proof of lower bound, and appendix H for the proof of upper bound.

**Remark 3.** This theorems shows that each for each $\gamma > 0$ such that $\mathcal{Q}_{\mathrm{NT}}(\gamma)^c \cap \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma) = \emptyset$, we can decompose our functional space as

$$L^2(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}, \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}}) = \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) \oplus \mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma),$$

where

$$\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) = \bigoplus_{\boldsymbol{k}\in\mathcal{Q}_{\mathrm{NT}}(\gamma)} \boldsymbol{V}^{\boldsymbol{d}}_{\boldsymbol{k}},$$

$$\mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma) = \bigoplus_{\boldsymbol{k}\notin\mathcal{Q}_{\mathrm{NT}}(\gamma)} \boldsymbol{V}^{\boldsymbol{d}}_{\boldsymbol{k}},$$

such that for $N = d^\gamma$, NT model fits the subspace of low degree polynomials $\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma)$ and cannot fit $\mathcal{F}^c(\boldsymbol{\beta}, \boldsymbol{\kappa}, \gamma)$ at all, i.e.

$$R_{\mathrm{NT}}(f_d, \boldsymbol{W}) \approx \|\mathsf{P}_{\mathcal{Q}_{\mathrm{NT}}(\gamma)^c} f_d\|^2_{L^2}.$$

**Remark 4.** In other words, we can fit a polynomial of degree $\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q}$, if and only if

$$d^{(\xi-\kappa_1)k_1} \cdot \ldots \cdot d^{(\xi-\kappa_Q)k_Q} = d_{1,\text{eff}}^{k_1} \ldots d_{Q,\text{eff}}^{k_Q} = o_d(d^{\beta} N),$$

where $\beta = \xi - \ \min_{q \in S(\boldsymbol{k})} \kappa_q$.

### C.7. Connecting to the theorems in the main text

Let us connect the above general results to the two-spheres setting described in the main text. We consider two spheres with $\eta_1 = \eta$, $\kappa_1 = \kappa$ for the first sphere, and $\eta_2 = 1$, $\kappa_2 = 0$ for the second sphere. We have $\xi = \ \max(\eta + \kappa, 1)$.

Let $w_d(d^{\gamma} \log d) \leqslant n \leqslant O_d(d^{\gamma+\delta})$ with $\delta > 0$ constant sufficiently small, then by theorem 5 the function subspace learned by KRR is given by the polynomials of degree $k_1$ in the first sphere coordinates and $k_2$ in the second sphere with

$$\max(\eta, 1-\kappa)k_1 + \max(\eta + \kappa, 1)k_2 < \gamma.$$

We consider functions that only depend on the first sphere, i.e. $k_2 = 0$ and denote $d_{\text{eff}} = d^{\max(\eta,1-\kappa)}$. Then the subspace of approximation is given by the $k$ polynomials in the first sphere such that $d_{\text{eff}}^{k} \leqslant d^{\gamma}$. Furthermore, one can check that the assumptions listed in theorem 1 in the main text verifies assumption 1.

Similarly, for $w_d(d^{\gamma}) \leqslant N \leqslant O_d(d^{\gamma+\delta})$ with $\delta > 0$ constant sufficiently small, theorem 6 implies that the RF models can only approximate $k$ polynomials in the first sphere such that $d_{\text{eff}}^{k} \leqslant d^{\gamma}$. Furthermore, assumptions listed in theorem 2 in the main text verifies assumption 2.

In the case of NT, we only consider $\boldsymbol{k} = (k_1, 0)$ and $S(\boldsymbol{k}) = \{1\}$. We get $\min_{q \in S(\boldsymbol{k})} \kappa_q = \kappa$. The subspace approximated is given by the $k$ polynomials in the first sphere such that $d_{\text{eff}}^{k} \leqslant d^{\gamma} d_{\text{eff}}$. Furthermore, assumptions listed in theorem 3 in the main text verifies assumption 3.

## Appendix D. Proof of theorem 5

The proof follows closely the proof of theorem 4 in [21].

### D.1. Preliminaries

Let us rewrite the kernel functions $\{h_d\}_{d\geqslant 1}$ as functions on the product of normalized spheres: for $\boldsymbol{x} = \{\boldsymbol{x}^{(q)}\}_{q\in[Q]}$ and $\boldsymbol{y} = \{\boldsymbol{y}^{(q)}\}_{q\in[Q]} \in \mathrm{PS}_{\boldsymbol{\kappa}}^{\boldsymbol{d}}$:

$$\begin{aligned} h_d(\langle \boldsymbol{y}, \boldsymbol{x}\rangle / R^2) &= h_d\left( \sum_{q\in[Q]} (r_q^2 / R^2 \sqrt{d_q}) \cdot \langle \overline{\boldsymbol{y}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle / \sqrt{d_q} \right) \\ &\equiv h_{\boldsymbol{d}}\left( \{ \langle \overline{\boldsymbol{y}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle / \sqrt{d_q} \}_{q\in[Q]} \right). \end{aligned} \tag{54}$$

Consider the expansion of $h_{\boldsymbol{d}}$ in terms of tensor product of Gegenbauer polynomials. We have

$$h_d(\langle \boldsymbol{y}, \boldsymbol{x} \rangle / R^2) = \sum_{\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) B(\boldsymbol{d}, \boldsymbol{k}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ \langle \overline{\boldsymbol{y}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle \}_{q \in [Q]} \right),$$

where

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) = \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ h_{\boldsymbol{d}} \left( \overline{x}_1^{(1)}, \ldots, \overline{x}_1^{(Q)} \right) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \sqrt{d_1} \overline{x}_1^{(1)}, \ldots, \sqrt{d_Q} \overline{x}_1^{(Q)} \right) \right],$$

where the expectation is taken over $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)}, \ldots, \overline{\boldsymbol{x}}^{(Q)}) \sim \mu_{\boldsymbol{d}}$.

**Lemma 2.** *Let $\{h_d\}_{d \geqslant 1}$ be a sequence of kernel functions that satisfies assumption 1. Assume $w_d(d^\gamma) \leqslant n \leqslant o_d(d^{m(\gamma)})$ for some $\gamma > 0$. Consider $\mathcal{Q} = \overline{\mathcal{Q}}_{\text{KRR}}(\gamma)$ as defined in equation (41). Then there exists constants $c, C > 0$ such that for $d$ large enough,*

$$\max_{\boldsymbol{k} \notin \mathcal{Q}} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) \leqslant C d^{-m(\gamma)},$$

$$\min_{\boldsymbol{k} \in \mathcal{Q}} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) \geqslant c d^{-\gamma}.$$

**Proof of lemma 2.** Notice that by lemma 18,

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) = \left( \prod_{q \in [Q]} \alpha_q^{k_q} \right) \cdot R(\boldsymbol{d}, \boldsymbol{k}) \cdot \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ \left( \prod_{q \in [Q]} \left( 1 - \frac{(\overline{x}_1^{(q)})^2}{d_q} \right)^{k_q} \right) \cdot h_d^{(|\boldsymbol{k}|)} \left( \sum_{q \in [Q]} \alpha_q \overline{x}_1^{(q)} \right) \right],$$

where $\alpha_q = d_q^{-1/2} r_q^2 / R^2 = (1 + o_d(1)) d^{\eta_q/2 + \kappa_q - \xi}$. By assumption 1(a), we have

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) B(\boldsymbol{d}, \boldsymbol{k}) \leqslant C \prod_{q \in [Q]} d^{(\kappa_q - \xi) k_q}.$$

Furthermore, by assumption 1(b) and dominated convergence,

$$\mathbb{E}_{\overline{\boldsymbol{x}}} \left[ \left( \prod_{q \in [Q]} \left( 1 - \frac{(\overline{x}_1^{(q)})^2}{d_q} \right)^{k_q} \right) \cdot h_d^{(|\boldsymbol{k}|)} \left( \sum_{q \in [Q]} \alpha_q \overline{x}_1^{(q)} \right) \right] \to h_d^{(|\boldsymbol{k}|)}(0) \geqslant c > 0,$$

for $k \geqslant \overline{K}$. The lemma then follows from the same proof as in lemmas 9 and 10, where we adapt the proofs of lemmas 19 and 20 to $h_{\boldsymbol{d}}$. □

### D.2. Proof of theorem 5

**Step 1. Rewrite the $\boldsymbol{y}$, $\boldsymbol{E}$, $\boldsymbol{H}$, $\boldsymbol{M}$ matrices.**

The test error of empirical KRR gives

$$\begin{aligned} R_{\text{KRR}}(f_d, \boldsymbol{X}, \lambda) &\equiv \mathbb{E}_{\boldsymbol{x}} \left[ \left( f_d(\boldsymbol{x}) - \boldsymbol{y}^\mathsf{T} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{h}(\boldsymbol{x}) \right)^2 \right] \\ &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - 2 \boldsymbol{y}^\mathsf{T} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{E} + \boldsymbol{y}^\mathsf{T} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{E} = (E_1, \ldots, E_n)^{\mathsf{T}}$ and $\boldsymbol{M} = (M_{ij})_{ij\in[n]}$, with

$$\begin{aligned} E_i &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})h_d(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle/d)], \\ M_{ij} &= \mathbb{E}_{\boldsymbol{x}}[h_d(\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle/d)h_d(\langle \boldsymbol{x}_j, \boldsymbol{x}\rangle/d)]. \end{aligned}$$

Let $B = \sum_{\boldsymbol{k}\in\mathcal{Q}} B(\boldsymbol{d}, \boldsymbol{k})$. Define for any $\boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0}$,

$$\begin{aligned} \boldsymbol{D}_{\boldsymbol{k}} &= \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})\mathbf{I}_{B(\boldsymbol{d},\boldsymbol{k})}, \\ \boldsymbol{Y}_{\boldsymbol{k}} &= (Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}_i))_{i\in[n],\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{n\times B(\boldsymbol{d},\boldsymbol{k})}, \\ \boldsymbol{\lambda}_{\boldsymbol{k}} &= (\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(f_d))^{\mathsf{T}}_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{B(\boldsymbol{d},\boldsymbol{k})}, \\ \boldsymbol{D}_{\mathcal{Q}} &= \operatorname{diag}\left(\left(\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})\mathbf{I}_{B(\boldsymbol{d},\boldsymbol{k})}\right)_{\boldsymbol{k}\in\mathcal{Q}}\right) \in \mathbb{R}^{B\times B} \\ \boldsymbol{Y}_{\mathcal{Q}} &= (\,\boldsymbol{Y}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{Q}} \in \mathbb{R}^{n\times B}, \\ \boldsymbol{\lambda}_{\mathcal{Q}} &= \left((\boldsymbol{\lambda}^{\mathsf{T}}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{Q}}\right)^{\mathsf{T}} \in \mathbb{R}^{B}. \end{aligned}$$

Let the spherical harmonics decomposition of $f_d$ be

$$f_d(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(f_d)Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}),$$

and the Gegenbauer decomposition of $h_{\boldsymbol{d}}$ be

$$h_{\boldsymbol{d}}\left(\overline{x}^{(1)}_1, \ldots, \overline{x}^{(Q)}_1\right) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})B(\boldsymbol{d}, \boldsymbol{k})Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\overline{x}^{(1)}_1, \ldots, \sqrt{d_Q}\overline{x}^{(Q)}_1\right).$$

We write the decompositions of vectors $\boldsymbol{f}$, $\boldsymbol{E}$, $\boldsymbol{H}$, and $\boldsymbol{M}$. We have

$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}} + \sum_{\boldsymbol{k}\in\mathcal{Q}^c} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{\lambda}_{\boldsymbol{k}}, \\ \boldsymbol{E} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}} + \sum_{\boldsymbol{k}\in\mathcal{Q}^c} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{\lambda}_{\boldsymbol{k}}, \\ \boldsymbol{H} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}^{\mathsf{T}}_{\mathcal{Q}} + \sum_{\boldsymbol{k}\in\mathcal{Q}^c} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{Y}^{\mathsf{T}}_{\boldsymbol{k}}, \\ \boldsymbol{M} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}^2_{\mathcal{Q}}\boldsymbol{Y}^{\mathsf{T}}_{\mathcal{Q}} + \sum_{\boldsymbol{k}\in\mathcal{Q}^c} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}^2_{\boldsymbol{k}}\boldsymbol{Y}^{\mathsf{T}}_{\boldsymbol{k}}. \end{aligned}$$

From lemma 4, we can rewrite

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}^{\mathsf{T}}_{\mathcal{Q}} + \kappa_h(\mathbf{I}_n + \boldsymbol{\Delta}_h), \\ \boldsymbol{M} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}^2_{\mathcal{Q}}\boldsymbol{Y}^{\mathsf{T}}_{\mathcal{Q}} + \kappa_u\boldsymbol{\Delta}_u, \end{aligned}$$

where $\kappa_h = \Theta_d(1)$, $\kappa_u = O_d(d^{-m(\gamma)})$, $\|\boldsymbol{\Delta}_h\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\|\boldsymbol{\Delta}_u\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$.

**Step 2. Decompose the risk**

The rest of the proof follows closely from theorem 4 in [21]. We decompose the risk as follows

$$R_{\mathrm{KRR}}(f_d, \boldsymbol{X}, \lambda) = \|f_d\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5,$$

where

$$\begin{aligned}
T_1 &= \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{E},\\
T_2 &= \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{f},\\
T_3 &= \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{\varepsilon},\\
T_4 &= \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{E},\\
T_5 &= \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{f}.
\end{aligned}$$

Further, we denote $\boldsymbol{f}_{\mathcal{Q}}$, $\boldsymbol{f}_{\mathcal{Q}^c}$, $\boldsymbol{E}_{\mathcal{Q}}$, and $\boldsymbol{E}_{\mathcal{Q}^c}$,

$$\begin{aligned}
\boldsymbol{f}_{\mathcal{Q}} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}, & \boldsymbol{E}_{\mathcal{Q}} &= \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}},\\
\boldsymbol{f}_{\mathcal{Q}^c} &= \sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{\lambda}_{\boldsymbol{k}}, & \boldsymbol{E}_{\mathcal{Q}^c} &= \sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{\lambda}_{\boldsymbol{k}}.
\end{aligned}$$

**Step 3. Term $T_2$**

Note we have

$$T_2 = T_{21} + T_{22} + T_{23},$$

where

$$\begin{aligned}
T_{21} &= \boldsymbol{f}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}},\\
T_{22} &= 2\boldsymbol{f}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}^c},\\
T_{23} &= \boldsymbol{f}_{\mathcal{Q}^c}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}^c}.
\end{aligned}$$

By lemma 6, we have

$$\|n(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} - \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1), \tag{55}$$

hence

$$\begin{aligned}
T_{21} &= \boldsymbol{\lambda}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}\\
&= \boldsymbol{\lambda}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}/n^2 + [\|\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2/n] \cdot o_{d,\mathbb{P}}(1).
\end{aligned}$$

By lemma 3, we have (with $\|\boldsymbol{\Delta}\|_2 = o_{d,\mathbb{P}}(1)$)

$$\boldsymbol{\lambda}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}/n^2 = \boldsymbol{\lambda}_{\mathcal{Q}}^{\mathsf{T}}(\mathbf{I}_B + \boldsymbol{\Delta})^2\boldsymbol{\lambda}_{\mathcal{Q}} = \|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2(1 + o_{d,\mathbb{P}}(1)).$$

Moreover, we have

$$\|\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2/n = \boldsymbol{\lambda}_{\mathcal{Q}}^{\mathsf{T}}(\mathbf{I}_B + \boldsymbol{\Delta})\boldsymbol{\lambda}_{\mathcal{Q}} = \|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2(1 + o_{d,\mathbb{P}}(1)).$$

As a result, we have

$$T_{21} = \|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2(1 + o_{d,\mathbb{P}}(1)) = \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2(1 + o_{d,\mathbb{P}}(1)). \tag{56}$$

By equation (55) again, we have

$$\begin{aligned} T_{23} &= \left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{\lambda}_k^{\mathsf{T}} \boldsymbol{Y}_k^{\mathsf{T}}\right) (\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1} \left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{Y}_k\boldsymbol{\lambda}_k\right) \\ &= \left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{\lambda}_k^{\mathsf{T}} \boldsymbol{Y}_k^{\mathsf{T}}\right) \boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} \left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{Y}_k\boldsymbol{\lambda}_k\right) /n^2 + \left[\left\|\sum_{k\in\mathcal{Q}^c} \boldsymbol{Y}_k\boldsymbol{\lambda}_k\right\|_2^2 /n\right] \cdot o_{d,\mathbb{P}}(1). \end{aligned}$$

By lemma 5, we have

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{\lambda}_k^{\mathsf{T}} \boldsymbol{Y}_k^{\mathsf{T}}\right) \boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} \left(\sum_{k\in\mathcal{Q}^c} \boldsymbol{Y}_k\boldsymbol{\lambda}_k\right)\right] /n^2 &= \sum_{u,v\in\mathcal{Q}^c} \boldsymbol{\lambda}_u^{\mathsf{T}} \{\mathbb{E}[(\boldsymbol{Y}_u^{\mathsf{T}} \boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} \boldsymbol{Y}_v)]/n^2\} \boldsymbol{\lambda}_v \\ &= \frac{B}{n} \sum_{k\in\mathcal{Q}^c} \|\boldsymbol{\lambda}_k\|_2^2. \end{aligned}$$

Moreover

$$\mathbb{E}\left[\left\|\sum_{k\in\mathcal{Q}^c} \boldsymbol{Y}_k\boldsymbol{\lambda}_k\right\|_2^2 /n\right] = \sum_{k\in\mathcal{Q}^c} \|\boldsymbol{\lambda}_k\|_2^2 = \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2.$$

This gives

$$T_{23} = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2. \tag{57}$$

Using Cauchy Schwarz inequality for $T_{22}$, we get

$$T_{22} \leqslant 2(T_{21}T_{23})^{1/2} = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2} \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}. \tag{58}$$

As a result, combining equations (56)–(58), we have

$$T_2 = \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^2}^2. \tag{59}$$

**Step 4. Term** $T_1$. Note we have

$$T_1 = T_{11} + T_{12} + T_{13},$$

where

$$
\begin{aligned}
T_{11} &= \boldsymbol{f}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{E}_{\mathcal{Q}},\\
T_{12} &= \boldsymbol{f}_{\mathcal{Q}^c}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{E}_{\mathcal{Q}},\\
T_{13} &= \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{E}_{\mathcal{Q}^c}.
\end{aligned}
$$

By lemma 7, we have

$$
\|\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}-\mathbf{I}_B\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),
$$

so that

$$
T_{11} = \boldsymbol{\lambda}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}} = \|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2^2(1+o_{d,\mathbb{P}}(1)) = \|\mathsf{P}_{\mathcal{Q}}f_d\|_2^2(1+o_{d,\mathbb{P}}(1)). \tag{60}
$$

Using Cauchy Schwarz inequality for $T_{12}$, and by the expression of $\boldsymbol{M} = \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}^2\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + \kappa_u\boldsymbol{\Delta}_u$ with $\|\boldsymbol{\Delta}_u\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ and $\kappa_u = O_d(d^{-m(\lambda)})$, we get with high probability

$$
\begin{aligned}
|T_{12}| &= \left|\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{\lambda}_{\mathcal{Q}}\right|\\
&\leqslant \left\|\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\right\|_2\|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2\\
&= \left[\left(\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\right)(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}^2\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\left(\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\right)\right]^{1/2}\|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2\\
&\leqslant \left[\left(\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\right)(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\left(\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\boldsymbol{\lambda}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\right)\right]^{1/2}\|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2\\
&= T_{23}^{1/2}\|\boldsymbol{\lambda}_{\mathcal{Q}}\|_2 = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}.
\end{aligned} \tag{61}
$$

For term $T_{13}$, we have

$$
|T_{13}| = |\boldsymbol{f}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{E}_{\mathcal{Q}^c}| \leqslant \|\boldsymbol{f}\|_2\|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}}\|\boldsymbol{E}_{\mathcal{Q}^c}\|_2.
$$

Note we have $\mathbb{E}[\|\boldsymbol{f}\|_2^2] = n\|f_d\|_{L^2}^2$, and $\|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}} \leqslant 2/(\kappa_h+\lambda)$ with high probability, and

$$
\mathbb{E}[\|\boldsymbol{E}_{\mathcal{Q}^c}\|_2^2] = n\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2\|\mathsf{P}_{\boldsymbol{k}}f_d\|_{L^2}^2 \leqslant n\left[\max_{\boldsymbol{k}\in\mathcal{Q}^c}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2\right]\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}^2.
$$

As a result, we have

$$
\begin{aligned}
|T_{13}| &\leqslant O_d(1)\cdot\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}\|f_d\|_{L^2}\left[n^2\max_{\boldsymbol{k}\in\mathcal{Q}^c}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2\right]^{1/2}/(\kappa_h+\lambda)\\
&= o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}\|f_d\|_{L^2},
\end{aligned} \tag{62}
$$

https://doi.org/10.1088/1742-5468/ac3a81

where the last equality used the fact that $n \leqslant O_d(d^{m(\gamma)-\delta})$ and lemma 2. Combining equations (60)–(62), we get

$$T_1 = \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^2}^2. \tag{63}$$

**Step 5. Terms** $T_3, T_4$ and $T_5$. By lemma 6 again, we have

$$\mathbb{E}_{\varepsilon}[T_3]/\tau^2 = \mathrm{tr}((\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}) = \mathrm{tr}(\boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n^2) + o_{d,\mathbb{P}}(1).$$

By lemma 3, we have

$$\mathrm{tr}(\boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n^2) = \mathrm{tr}(\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} \boldsymbol{Y}_{\mathcal{Q}})/n^2 = nB/n^2 + o_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1).$$

This gives

$$T_3 = o_{d,\mathbb{P}}(1) \cdot \tau^2. \tag{64}$$

Let us consider $T_4$ term:

$$\begin{aligned}
\mathbb{E}_{\varepsilon}[T_4^2]/\tau^2 &= \mathbb{E}_{\varepsilon}[\boldsymbol{\varepsilon}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{E} \boldsymbol{E}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\varepsilon}]/\tau^2 \\
&= \boldsymbol{E}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{E}.
\end{aligned}$$

For any integer $L$, denote $\mathcal{L} \equiv [0, L]^Q \cap \mathbb{Z}_{\geqslant 0}^Q$, and $\boldsymbol{Y}_{\mathcal{L}} = (\boldsymbol{Y}_{\boldsymbol{k}})_{\boldsymbol{k} \in \mathcal{L}}$ and $\boldsymbol{D}_{\mathcal{L}} = (\boldsymbol{D}_{\boldsymbol{k}})_{\boldsymbol{k} \in \mathcal{L}}$. Then notice that by lemmas 3, 6 and the definition of $\boldsymbol{M}$, we get

$$\begin{aligned}
\|\boldsymbol{D}_{\mathcal{L}} \boldsymbol{Y}_{\mathcal{L}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{Y}_{\mathcal{L}} \boldsymbol{D}_{\mathcal{L}}\|_{\mathrm{op}} &= \|(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{Y}_{\mathcal{L}} \boldsymbol{D}_{\mathcal{L}}^2 \boldsymbol{Y}_{\mathcal{L}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\|_{\mathrm{op}} \\
&\leqslant \|(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\|_{\mathrm{op}}. \\
&\leqslant \|\boldsymbol{Y}_{\mathcal{Q}} \boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n\|_{\mathrm{op}}/n + o_{\mathbb{P},d}(1) \cdot /n \\
&= o_{d,\mathbb{P}}(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\boldsymbol{E}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{E} &= \lim_{L \to \infty} \boldsymbol{E}_{\mathcal{L}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{E}_{\mathcal{L}} \\
&= \lim_{L \to \infty} \boldsymbol{\lambda}_{\mathcal{L}}^{\mathsf{T}} [\boldsymbol{D}_{\mathcal{L}} \boldsymbol{Y}_{\mathcal{L}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{Y}_{\mathcal{L}} \boldsymbol{D}_{\mathcal{L}}] \boldsymbol{\lambda}_{\mathcal{L}} \\
&\leqslant \|(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\|_{\mathrm{op}} \cdot \lim_{L \to \infty} \|\boldsymbol{\lambda}_{\mathcal{L}}\|_2^2 \\
&\leqslant o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^2}^2,
\end{aligned}$$

which gives

$$T_4 = o_{d,\mathbb{P}}(1) \cdot \tau \|f_d\|_{L^2} = o_{d,\mathbb{P}}(1) \cdot (\tau^2 + \|f_d\|_{L^2}^2). \tag{65}$$

We decompose $T_5$ using $\boldsymbol{f} = \boldsymbol{f}_{\mathcal{Q}} + \boldsymbol{f}_{\mathcal{Q}^c}$,

$$T_5 = T_{51} + T_{52},$$

where

$$\begin{aligned} T_{51} &= \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}}, \\ T_{52} &= \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}^c}. \end{aligned}$$

First notice that

$$\|\boldsymbol{M}^{1/2}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-2}\boldsymbol{M}^{1/2}\|_{\mathrm{op}} = \|(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Then by lemma 6, we get

$$\begin{aligned} \mathbb{E}_{\varepsilon}[T_{51}^2]/\tau^2 &= \mathbb{E}_{\varepsilon}[\boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}}\boldsymbol{f}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{\varepsilon}]/\tau^2 \\ &= \boldsymbol{f}_{\mathcal{Q}}^{\mathsf{T}}[(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}]^2\boldsymbol{f}_{\mathcal{Q}} \\ &\leqslant \|\boldsymbol{M}^{1/2}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-2}\boldsymbol{M}^{1/2}\|_{\mathrm{op}}\|\boldsymbol{M}^{1/2}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{f}_{\mathcal{Q}}\|_2^2 \\ &= o_{d,\mathbb{P}}(1)\cdot T_{21} \\ &= o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}^2. \end{aligned}$$

Similarly, we get

$$\mathbb{E}_{\varepsilon}[T_{52}^2]/\tau^2 = o_{d,\mathbb{P}}(1)\cdot T_{23} = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}^2.$$

By Markov's inequality, we deduce that

$$T_5 = o_{d,\mathbb{P}}(1)\cdot\tau(\|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2} + \|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}) = o_{d,\mathbb{P}}(1)\cdot(\tau^2 + \|f_d\|_{L^2}^2). \tag{66}$$

**Step 6. Finish the proof.**
Combining equations (59) and (63)–(66), we have

$$\begin{aligned} R_{\mathrm{KRR}}(f_d, \boldsymbol{X}, \lambda) &= \|f_d\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5 \\ &= \|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^2}^2 + \tau^2), \end{aligned}$$

which concludes the proof.

### D.3. Auxiliary results

**Lemma 3.** *Let* $\{Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}\}_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^{Q},\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}$ *be the collection of tensor product of spherical harmonics on* $\mathrm{PS}^{\mathbf{d}}$. *Let* $(\overline{\boldsymbol{x}}_i)_{i\in[n]}\sim_{iid}\mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$. *Denote*

$$\boldsymbol{Y}_{\boldsymbol{k}} = (Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i))_{i\in[n],\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{n\times B(\boldsymbol{d},\boldsymbol{k})}.$$

*Assume that* $n \geqslant w_d(d^{\gamma}\log d)$ *and consider*

$$\mathcal{R} = \left\{ \boldsymbol{k} \in \mathbb{Z}^{Q}_{\geqslant 0} \,\middle|\, \sum_{q\in[Q]} \eta_q k_q < m(\gamma) \right\}.$$

*Denote* $A = \sum_{\boldsymbol{k}\in\mathcal{R}} B(\boldsymbol{d}, \boldsymbol{k})$ *and*

$$\boldsymbol{Y}_{\mathcal{R}} = (\,\boldsymbol{Y}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{R}} \in \mathbb{R}^{n\times A}.$$

*Then we have*

$$\boldsymbol{Y}_{\mathcal{R}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{R}}/n = \mathbf{I}_A + \boldsymbol{\Delta},$$

*with* $\boldsymbol{\Delta} \in \mathbb{R}^{A\times A}$ *and* $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] = o_d(1)$.

**Proof of lemma 3.** Let $\boldsymbol{\Psi} = \boldsymbol{Y}_{\mathcal{R}}^{\mathsf{T}}\boldsymbol{Y}_{\mathcal{R}}/n \in \mathbb{R}^{A\times A}$. We can rewrite $\boldsymbol{\Psi}$ as

$$\boldsymbol{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i \boldsymbol{h}_i^{\mathsf{T}},$$

where $\boldsymbol{h}_i = (Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}_i))_{\boldsymbol{k}\in\mathcal{R},\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{A}$. We use matrix Bernstein inequality. Denote $\boldsymbol{X}_i = \boldsymbol{h}_i\boldsymbol{h}_i - \mathbf{I}_A \in \mathbb{R}^{A\times A}$. Then we have $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$, and

$$\begin{aligned}\|\boldsymbol{X}_i\|_{\mathrm{op}} \leqslant \|\boldsymbol{h}_i\|_2^2 + 1 &= \sum_{\boldsymbol{k}\in\mathcal{R}} \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}_i)^2 + 1 \\ &= \sum_{\boldsymbol{k}\in\mathcal{R}} B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}_i^{(q)}, \overline{\boldsymbol{x}}_i^{(q)}\rangle\}_{q\in[Q]}\right) + 1 = A + 1,\end{aligned}$$

where we use formula (22) and the normalization $Q^{\boldsymbol{d}}_{\boldsymbol{k}}(d_1, \ldots, d_Q) = 1$. Denote $V = \|\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{X}_i^2]\|_{\mathrm{op}}$. Then we have

$$\begin{aligned}V &= n\|\mathbb{E}[(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - \mathbf{I}_A)^2]\|_{\mathrm{op}} = n\|\mathbb{E}[\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - 2\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} + \mathbf{I}_A]\|_{\mathrm{op}} \\ &= n\|(A-1)\mathbf{I}_A\|_{\mathrm{op}} = n(A-1),\end{aligned}$$

where we used $\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{h}_i = \|\boldsymbol{h}_i\|_2^2 = A$ and $\mathbb{E}[\boldsymbol{h}_i(\overline{\boldsymbol{x}}_i)\boldsymbol{h}_i^{\mathsf{T}}(\overline{\boldsymbol{x}}_i)] = (\mathbb{E}[Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}_i)Y^{\boldsymbol{d}}_{\boldsymbol{k}',\boldsymbol{s}'}(\overline{\boldsymbol{x}}_i)])_{\boldsymbol{k}\boldsymbol{s},\boldsymbol{k}'\boldsymbol{s}'} = \mathbf{I}_A$. As a result, we have for any $t > 0$,

$$\begin{aligned}\mathbb{P}(\|\boldsymbol{\Psi} - \mathbf{I}_A\|_{\mathrm{op}} \geqslant t) &\leqslant A\,\exp\{-n^2t^2/[2n(A-1) + 2(A+1)nt/3]\} \\ &\leqslant \exp\{-(n/A)t^2/[10(1+t)] + \log A\}. \qquad (67)\end{aligned}$$

Notice that there exists $C > 0$ such that $A \leqslant C\max_{\boldsymbol{k}\in\mathcal{R}}\prod_{q\in[Q]} d^{\eta_q k_q} \leqslant Cd^{\gamma}$ (by definition of $m(\gamma)$ and $\mathcal{R}$) and therefore $n \geqslant w_d(A\log A)$. Integrating the tail bound (67) proves the lemma. □

**Lemma 4.** *Let $\sigma$ be an activation function satisfying assumption 1. Let $w_d(d^{\gamma}\log d) \leqslant n \leqslant O_d(d^{m(\gamma)-\delta})$ for some $\gamma > 0$ and $\delta > 0$. Then there exists sequences $\kappa_h$ and $\kappa_u$ such that*

$$\boldsymbol{H} = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}} = \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + \kappa_h(\mathbf{I}_n + \boldsymbol{\Delta}_h), \tag{68}$$

$$\boldsymbol{M} = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}^2\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}} = \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}^2\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + \kappa_m\boldsymbol{\Delta}_m, \tag{69}$$

*where $\kappa_h = \Theta_d(1)$, $\kappa_m = O_d(d^{-m(\gamma)})$, $\|\boldsymbol{\Delta}_h\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\|\boldsymbol{\Delta}_m\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$.*

**Proof of lemma 4.** Define

$$\mathcal{R} = \left\{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}\,\middle|\,\sum_{q\in[Q]}\eta_q k_q < m(\gamma)\right\},$$

$$\mathcal{S} = \left\{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}\,\middle|\,\sum_{q\in[Q]}\eta_q k_q \geqslant m(\gamma)\right\},$$

such that $\mathcal{R}\cup\mathcal{S} = \mathbb{Z}^Q_{\geqslant 0}$. The proof comes from bounding the eigenvalues of the matrix $\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}$ for $\boldsymbol{k}\in\mathcal{R}$ and $\boldsymbol{k}\in\mathcal{S}$ separately. From corollary 1, we have

$$\sup_{\boldsymbol{k}\in\mathcal{S}}\|\,\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}/B(\boldsymbol{d},\boldsymbol{k}) - \mathbf{I}_n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Hence, we can write

$$\sum_{\boldsymbol{k}\in\mathcal{S}}\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}} = \kappa_h(\mathbf{I}_n + \boldsymbol{\Delta}_{h,1}), \tag{70}$$

with $\kappa_h = \sum_{\boldsymbol{k}\in\mathcal{S}}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})B(\boldsymbol{d},\boldsymbol{k}) = O_d(1)$. From assumption 1(b) and a proof similar to lemma 20, there exists $\boldsymbol{k} = (0,\ldots,k,\ldots,0)$ (for $k > L$ at position $q_\xi$) such that $\liminf_{d\to\infty}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})B(\boldsymbol{d},\boldsymbol{k}) > 0$. Hence, $\kappa_h = \Theta_d(1)$.

From lemma 3 we have for $\boldsymbol{k}\in\mathcal{R}\cap\mathcal{Q}^c$,

$$\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{k}}/n = \mathbf{I}_{B(\boldsymbol{d},\boldsymbol{k})} + \boldsymbol{\Delta},$$

with $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. We deduce that $\|\,\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n)$. Hence,

$$\|\,\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})) = o_{d,\mathbb{P}}(1),$$

where we used lemma 2. We deduce that

$$\sum_{\boldsymbol{k}\in\mathcal{R}\cap\mathcal{Q}^c}\boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}} = \kappa_h\boldsymbol{\Delta}_{h,2}, \tag{71}$$

with $\|\boldsymbol{\Delta}_{h,2}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ where we used $\kappa_h^{-1} = O_d(1)$. Combining equations (70) and (71) yields equation (68).

Similarly, we get

$$\sum_{\boldsymbol{k}\in\mathcal{Q}^c} \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{D}_{\boldsymbol{k}}^2\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}} = \sum_{\boldsymbol{k}\in\mathcal{R}\cap\mathcal{Q}^c} [\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 n]\, \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}/n + \sum_{\boldsymbol{k}\in\mathcal{S}} [\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 B(\boldsymbol{d},\boldsymbol{k})]\, \boldsymbol{Y}_{\boldsymbol{k}}\boldsymbol{Y}_{\boldsymbol{k}}^{\mathsf{T}}/B(\boldsymbol{d},\boldsymbol{k}).$$

Using lemma 2, we have $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 n \leqslant C d^{-2m(\gamma)} n = O_{d,\mathbb{P}}(d^{-m(\gamma)})$ and $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 B(\boldsymbol{d},\boldsymbol{k}) \leqslant C\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}}) \leqslant C' d^{-m(\gamma)}$. Hence equation (69) is verified with

$$\kappa_m = \sum_{\boldsymbol{k}\in\mathcal{R}\cap\mathcal{Q}^c} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 n + \sum_{\boldsymbol{k}\in\mathcal{S}} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(h_{\boldsymbol{d}})^2 B(\boldsymbol{d},\boldsymbol{k}).$$

□

**Lemma 5.** *Let $\{Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}\}_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q,\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}$ be the collection of product of spherical harmonics on $L^2(\mathrm{PS}^{\mathbf{d}}, \mu_{\mathbf{d}})$. Let $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$. Denote*

$$\boldsymbol{Y}_{\boldsymbol{k}} = (Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i))_{i\in[n],\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{n\times B(\boldsymbol{d},\boldsymbol{k})}.$$

*Then for $\boldsymbol{u},\boldsymbol{v},\boldsymbol{t}\in\mathbb{Z}_{\geqslant 0}^Q$ and $\mathbf{u}\neq\boldsymbol{v}$, we have*

$$\mathbb{E}[\,\boldsymbol{Y}_{\boldsymbol{u}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{t}}\boldsymbol{Y}_{\boldsymbol{t}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{v}}] = \mathbf{0}.$$

*For $\boldsymbol{u},\boldsymbol{t}\in\mathbb{Z}_{\geqslant 0}^Q$, we have*

$$\mathbb{E}[\,\boldsymbol{Y}_{\boldsymbol{u}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{t}}\boldsymbol{Y}_{\boldsymbol{t}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{u}}] = [B(\boldsymbol{d},\boldsymbol{t})n + n(n-1)\delta_{\boldsymbol{u},\boldsymbol{t}}]\mathbf{I}_{B(\boldsymbol{d},\boldsymbol{u})}.$$

**Proof.** We have

$$\begin{aligned}
\mathbb{E}[\,\boldsymbol{Y}_{\boldsymbol{u}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{t}}\boldsymbol{Y}_{\boldsymbol{t}}^{\mathsf{T}}\boldsymbol{Y}_{\boldsymbol{v}}] &= \sum_{i,j\in[n]}\sum_{\boldsymbol{m}\in[B(\boldsymbol{d},\boldsymbol{t})]} \left(\mathbb{E}[Y_{\boldsymbol{u},\boldsymbol{p}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)\left(Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_j)\right)Y_{\boldsymbol{v},\boldsymbol{q}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_j)]\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}\\
&= \sum_{i\in[n]}\left(\mathbb{E}\left[Y_{\boldsymbol{u},\boldsymbol{p}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)\left(\sum_{\boldsymbol{m}\in[B(\boldsymbol{d},\boldsymbol{t})]} Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)\right)Y_{\boldsymbol{v},\boldsymbol{q}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)\right]\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}\\
&\quad + \sum_{i\neq j\in[n]}\sum_{\boldsymbol{m}\in[B(\boldsymbol{d},\boldsymbol{t})]} \left(\mathbb{E}[Y_{\boldsymbol{u},\boldsymbol{p}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)Y_{\boldsymbol{t},\boldsymbol{m}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_j)Y_{\boldsymbol{v},\boldsymbol{q}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_j)]\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}\\
&= B(\boldsymbol{d},\boldsymbol{t})\sum_{i\in[n]}\left(\mathbb{E}[Y_{\boldsymbol{u},\boldsymbol{p}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)Y_{\boldsymbol{v},\boldsymbol{q}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_i)]\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}\\
&\quad + \sum_{i\neq j\in[n]}\sum_{\boldsymbol{m}\in[B(\boldsymbol{d},\boldsymbol{t})]} \left(\delta_{\boldsymbol{u},\boldsymbol{t}}\delta_{\boldsymbol{p},\boldsymbol{m}}\delta_{\boldsymbol{t},\boldsymbol{v}}\delta_{\boldsymbol{q},\boldsymbol{m}}\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}\\
&= \left(B(\boldsymbol{d},\boldsymbol{t})n\delta_{\boldsymbol{u},\boldsymbol{v}}\delta_{\boldsymbol{p},\boldsymbol{q}} + n(n-1)\delta_{\boldsymbol{u},\boldsymbol{t}}\delta_{\boldsymbol{t},\boldsymbol{v}}\delta_{\boldsymbol{p},\boldsymbol{q}}\right)_{\boldsymbol{p}\in[B(\boldsymbol{d},\boldsymbol{u})],\boldsymbol{q}\in[B(\boldsymbol{d},\boldsymbol{v})]}.
\end{aligned} \tag{72}$$

This proves the lemma. □

**Lemma 6.** *Let $\sigma$ be an activation function satisfying assumption 1. Assume $\omega_d(d^\gamma \log d) \leqslant n \leqslant O_d(d^{m(\gamma)-\delta})$ for some $\gamma > 0$ and $\delta > 0$. We have*

$$\|n(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1} - \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

**Proof of lemma 6.** Denote

$$\boldsymbol{Y}_{\boldsymbol{k}} = (Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}}_i))_{i\in[n],\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \in \mathbb{R}^{n\times B(\boldsymbol{d},\boldsymbol{k})}. \tag{73}$$

Denote $B = \sum_{\boldsymbol{k}\in\mathcal{Q}} B(\boldsymbol{d},\boldsymbol{k})$, and

$$\boldsymbol{Y}_{\mathcal{Q}} = (\boldsymbol{Y}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{Q}} \in \mathbb{R}^{n\times B},$$

and

$$\boldsymbol{D}_{\mathcal{Q}} = \mathrm{diag}((\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})\mathbf{I}_{B(\boldsymbol{d},\boldsymbol{k})})_{\boldsymbol{k}\in\mathcal{Q}}) \in \mathbb{R}^{B\times B}.$$

From lemma 4, we have

$$\begin{aligned}
&n(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\\
&\quad = n(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1}(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}^2\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + \kappa_m\boldsymbol{\Delta}_m)\\
&\qquad \times (\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1}\\
&\quad = T_1 + T_2,
\end{aligned}$$

where $\|\boldsymbol{\Delta}_h\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, $\|\boldsymbol{\Delta}_u\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ and $\kappa_m = O_d(d^{-m(\gamma)})$, and

$$T_1 = n\kappa_m(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1}\boldsymbol{\Delta}_m(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1},$$

$$T_2 = n(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}^2\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}} + (\kappa_h+\lambda)\mathbf{I}_n + \kappa_h\boldsymbol{\Delta}_h)^{-1}.$$

Then, we can use the same proof as in lemma 13 in [21] to bound $\|T_1\|_{\mathrm{op}}$ (recall $n = O_d(d^{m(\gamma)-\delta})$)

$$\|T_1\|_{\mathrm{op}} \leqslant 2n\kappa_m/(\kappa_h+\lambda)^2\|\boldsymbol{\Delta}_m\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

and $\|T_2 - \boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}/n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, where we only need to check that

$$\lambda_{\min}(\boldsymbol{D}_{\mathcal{Q}}/[(\kappa_h+\lambda)/n]) = \min_{\boldsymbol{k}\in\mathcal{Q}}[n\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})]/(\kappa_h+\lambda) = w_d(1),$$

which directly follows from lemma 2. □

**Lemma 7.** *Let $\sigma$ be an activation function satisfying assumption 1. Assume $\omega_d(d^\gamma \log d) \leqslant n \leqslant O_d(d^{m(\gamma)-\delta})$ for some $\gamma > 0$ and $\delta > 0$. We have*

$$\|\boldsymbol{Y}_{\mathcal{Q}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{Y}_{\mathcal{Q}}\boldsymbol{D}_{\mathcal{Q}} - \mathbf{I}_B\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

**Proof of lemma 7.** This lemma can be deduced directly from lemma 14 in [21], by noticing that

$$\lambda_{\min}\,(\boldsymbol{D}_{\mathcal{Q}}/[(\kappa_h+\lambda)/n] = \min_{\boldsymbol{k}\in\mathcal{Q}}[n\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(h_{\boldsymbol{d}})]/(\kappa_h+\lambda) = \omega_d(1),$$

from lemma 2. □

## Appendix E. Proof of theorem 6(a): lower bound for the RF model

### E.1. Preliminaries

In the theorems, we show our results in high probability with respect to $\boldsymbol{\Theta}$. Hence, in the proof we will restrict the sample space to the high probability event $\mathcal{P}_\varepsilon \equiv \mathcal{P}_{d,N,\varepsilon}$ for $\varepsilon > 0$ small enough, where

$$\mathcal{P}_{d,N,\varepsilon} \equiv \left\{\boldsymbol{\Theta}\,\middle|\,\tau_i^{(q)} \in [1-\varepsilon, 1+\varepsilon], \quad \forall\, i \in [N],\ \forall\, q \in [Q]\right\} \subset \left(\mathbb{S}^{D-1}(\sqrt{D})\right)^{\otimes N}. \tag{74}$$

We will denote $\mathbb{E}_{\boldsymbol{\tau}_\varepsilon}$ the expectation over $\boldsymbol{\tau}$ restricted to $\tau^{(q)} \in [1-\varepsilon, 1+\varepsilon]$ for all $q \in [Q]$, and $\mathbb{E}_{\boldsymbol{\Theta}_\varepsilon}$ the expectation over $\boldsymbol{\Theta}$ restricted to the event $\mathcal{P}_\varepsilon$.

**Lemma 8.** *Assume $N = o(d^\gamma)$ for some $\gamma > 0$. We have for any fixed $\varepsilon > 0$,*

$$\mathbb{P}(\mathcal{P}^c_\varepsilon) = o_d(1).$$

**Proof of lemma 8.** The tail inequality in lemma 16 and the assumption $N = o(d^\gamma)$ imply that there exists some constants $C, c > 0$ such that

$$\mathbb{P}(\mathcal{P}^c_{d,N,\varepsilon}) \leqslant \sum_{q\in[Q]} N\mathbb{P}(|\tau^{(q)} - 1| > \varepsilon) \leqslant \sum_{q\in[Q]} C\,\exp(\gamma\,\log(d) - cd^{\eta_q}\varepsilon) = o_d(1).$$

□

We consider the activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Let $\boldsymbol{\theta} \sim \mathbb{S}^{D-1}(\sqrt{D})$ and $\boldsymbol{x} = \{\boldsymbol{x}^{(q)}\}_{q\in[Q]} \in \mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}$. We introduce the function $\sigma_{\boldsymbol{d},\boldsymbol{\tau}} : \mathrm{ps}^{\boldsymbol{d}} \to \mathbb{R}$ such that

$$\begin{aligned}\sigma(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R) &= \sigma\left(\sum_{q\in[Q]} \tau^{(q)}\cdot(r_q/R)\cdot\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\right)\\ &\equiv \sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right).\end{aligned} \tag{75}$$

Consider the expansion of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of tensor product of Gegenbauer polynomials. We have

$$\sigma(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right), \tag{76}$$

where

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)}\right) Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\dots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right],$$

where the expectation is taken over $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)},\dots,\overline{\boldsymbol{x}}^{(Q)}) \sim \mu_{\boldsymbol{d}}$.

**Lemma 9.** *Let $\sigma$ be an activation function that satisfies assumptions 2(a) and (b). Consider $N \leqslant o_d(d^{\gamma})$ and $\mathcal{Q} = \mathcal{Q}_{\mathrm{RF}}(\gamma)$ as defined in theorem 6(a). Then there exists $\varepsilon_0 > 0$ and $d_0$ and a constant $C > 0$ such that for $d \geqslant d_0$ and $\boldsymbol{\tau} \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$,*

$$\max_{\boldsymbol{k}\notin\mathcal{Q}} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant Cd^{-\gamma}.$$

**Proof of lemma 9.** Notice that by assumption 2(b) we can apply lemma 19 to any $\boldsymbol{k} \in \mathcal{Q}^c$ such that $|\boldsymbol{k}| = k_1 + \dots + k_Q \leqslant L$. In particular, there exists $C > 0$, $\varepsilon_0' > 0$ and $d_0'$ such that for any $\boldsymbol{k} \in \mathcal{Q}^c$ with $|\boldsymbol{k}| \leqslant L$, $d \geqslant d_0'$ and $\boldsymbol{\tau} \in [1-\varepsilon_0', 1+\varepsilon_0']^Q$,

$$\left(\prod_{q\in[Q]} d^{(\xi-\eta_q-\kappa_q)k_q}\right) B(\boldsymbol{d},\boldsymbol{k})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant C < \infty.$$

Furthermore, using that $B(\boldsymbol{d},\boldsymbol{k}) = \Theta(d_1^{k_1}d_2^{k_2}\dots d_Q^{k_Q})$, there exists $C' > 0$ such that for $\boldsymbol{k} \in \mathcal{Q}^c$ with $|\boldsymbol{k}| \leqslant L$,

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant C' \prod_{q\in[Q]} d^{(\eta_q+\kappa_q-\xi)k_q} d_q^{-k_q} = C' \prod_{q\in[Q]} d^{(\kappa_q-\xi)k_q} \leqslant C'd^{-\gamma}, \tag{77}$$

where we used in the last inequality $\boldsymbol{k} \notin \mathcal{Q}_{\mathrm{RF}}(\gamma)$ implies $(\xi-\kappa_1)k_1 + \dots + (\xi-\kappa_Q)k_Q \geqslant \gamma$ by definition.

Furthermore, from assumption 2 and lemma 17(b), there exists $\varepsilon_0'' > 0$, $d_0''$ and $C < \infty$, such that

$$\sup_{d\geqslant d_0''}\ \sup_{\boldsymbol{\tau}\in[1-\varepsilon_0'',1+\varepsilon_0'']^Q} \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] < C.$$

From the Gegenbauer decomposition (76), this implies that for any $\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q$, $d \leqslant d_0''$ and $\boldsymbol{\tau} \in [1-\varepsilon_0'', 1+\varepsilon_0'']^Q$,

$$B(\boldsymbol{d},\boldsymbol{k})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant C.$$

In particular, for $|\boldsymbol{k}| = k_1 + \dots + k_Q > L = \max_{q\in[Q]}\lceil \gamma/\eta_q \rceil$, we have

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant \frac{C}{B(\boldsymbol{d},\boldsymbol{k})} \leqslant C' \prod_{q\in[Q]} d^{-\eta_q k_q} \leqslant C' \prod_{q\in[Q]} d^{-\gamma k_q/L} \leqslant C'd^{-\gamma}. \tag{78}$$

Combining equations (77) and (78) yields the result. □

### E.2. Proof of theorem 6(a): outline

Let $\mathcal{Q} \equiv \mathcal{Q}_{\mathrm{RF}}(\gamma)$ as defined in theorem 6(a) and $\mathbf{\Theta} = \sqrt{D}\,\boldsymbol{W}$ such that $\boldsymbol{\theta}_i = \sqrt{D}\boldsymbol{w}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$.

Define the random vectors $\boldsymbol{V} = (V_1, \ldots, V_N)^{\mathsf{T}}$, $\boldsymbol{V}_{\mathcal{Q}} = (V_{1,\mathcal{Q}}, \ldots, V_{N,\mathcal{Q}})^{\mathsf{T}}$, $\boldsymbol{V}_{\mathcal{Q}^c} = (V_{1,\mathcal{Q}^c}, \ldots, V_{N,\mathcal{Q}^c})^{\mathsf{T}}$, with

$$V_{i,\mathcal{Q}} \equiv \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}} f_d](\boldsymbol{x})\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)], \tag{79}$$

$$V_{i,\mathcal{Q}^c} \equiv \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x})\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)], \tag{80}$$

$$V_i \equiv \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)] = V_{i,\mathcal{Q}} + V_{i,\mathcal{Q}^c}. \tag{81}$$

Define the random matrix $\boldsymbol{U} = (U_{ij})_{i,j\in[N]}$, with

$$U_{ij} = \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_i\rangle/R)\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle/R)]. \tag{82}$$

In what follows, we write $R_{\mathrm{RF}}(f_d) = R_{\mathrm{RF}}(f_d, \boldsymbol{W}) = R_{\mathrm{RF}}(f_d, \mathbf{\Theta}/\sqrt{D})$ for the random features risk, omitting the dependence on the weights $\boldsymbol{W} = \mathbf{\Theta}/\sqrt{D}$. By the definition and a simple calculation, we have

$$\begin{aligned}
R_{\mathrm{RF}}(f_d) &= \min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{\mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - 2\langle \boldsymbol{a}, \boldsymbol{V}\rangle + \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\right\}\\
&= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V},\\
R_{\mathrm{RF}}(\mathsf{P}_{\mathcal{Q}} f_d) &= \min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{\mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_{\mathcal{Q}} f_d(\boldsymbol{x})^2] - 2\langle \boldsymbol{a}, \boldsymbol{V}_{\leqslant \ell}\rangle + \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\right\}\\
&= \mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_{\mathcal{Q}} f_d(\boldsymbol{x})^2] - \boldsymbol{V}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}}.
\end{aligned}$$

By orthogonality, we have

$$\mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] = \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}} f_d](\boldsymbol{x})^2] + \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x})^2],$$

which gives

$$\begin{aligned}
&\left|R_{\mathrm{RF}}(f_d) - R_{\mathrm{RF}}(\mathsf{P}_{\mathcal{Q}} f_d) - \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x})^2]\right|\\
&\quad = \left|\boldsymbol{V}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}} - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}\right|\\
&\quad = \left|\boldsymbol{V}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}} - (\boldsymbol{V}_{\mathcal{Q}} + \boldsymbol{V}_{\mathcal{Q}^c})^{\mathsf{T}}\boldsymbol{U}^{-1}(\boldsymbol{V}_{\mathcal{Q}} + \boldsymbol{V}_{\mathcal{Q}^c})\right|\\
&\quad = \left|2\boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}^c} - \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}^c}\right|\\
&\quad \leqslant 2\|\boldsymbol{U}^{-1/2}\boldsymbol{V}_{\mathcal{Q}^c}\|_2\|\boldsymbol{U}^{-1/2}\boldsymbol{V}\|_2 + \|\boldsymbol{U}^{-1}\|_{\mathrm{op}}\|\boldsymbol{V}_{\mathcal{Q}^c}\|_2^2\\
&\quad \leqslant 2\|\boldsymbol{U}^{-1/2}\|_{\mathrm{op}}\|\boldsymbol{V}_{\mathcal{Q}^c}\|_2\|f_d\|_{L^2} + \|\boldsymbol{U}^{-1}\|_{\mathrm{op}}\|\boldsymbol{V}_{\mathcal{Q}^c}\|_2^2,
\end{aligned} \tag{83}$$

where the last inequality used the fact that

$$0 \leqslant R_{\mathrm{RF}}(f_d) = \|f_d\|_{L^2}^2 - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V},$$

so that

$$\|\boldsymbol{U}^{-1/2}\boldsymbol{V}\|_2^2 = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V} \leqslant \|f_d\|_{L^2}^2.$$

The theorem follows from the following two claims

$$\Vert \boldsymbol{V}_{\mathcal{Q}^c}\Vert_2/\Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2} = o_{d,\mathbb{P}}(1), \tag{84}$$

$$\Vert \boldsymbol{U}^{-1}\Vert_{\mathrm{op}} = O_{d,\mathbb{P}}(1). \tag{85}$$

This is achieved by the propositions 1 and 2 stated below.

**Proposition 1 (Expected norm of V).** *Let $\sigma$ be an activation function satisfying assumptions 2(a) and (b) for a fixed $\gamma > 0$. Denote $\mathcal{Q} = \mathcal{Q}_{\mathrm{RF}}(\gamma)$. Let $\varepsilon > 0$ and define $\mathcal{E}_{\mathcal{Q}^c,\varepsilon}$ by*

$$\mathcal{E}_{\mathcal{Q}^c,\varepsilon} \equiv \mathbb{E}_{\boldsymbol{\theta}_\varepsilon}[\langle \mathsf{P}_{\mathcal{Q}^c,0}f_d, \sigma(\langle \boldsymbol{\theta}, \cdot\rangle/R)\rangle_{L^2}^2],$$

*where we recall that $\mathbb{E}_{\boldsymbol{\theta}_\varepsilon} = \mathbb{E}_{\boldsymbol{\tau}_\varepsilon}\mathbb{E}_{\overline{\boldsymbol{\theta}}}$ the expectation with respect to $\boldsymbol{\tau}$ restricted to $[1-\varepsilon, 1+\varepsilon]^Q$ and $\overline{\boldsymbol{\theta}} \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$.*

*Then there exists a constant $C > 0$ and $\varepsilon_0 > 0$ (depending only on the constants of assumptions 2(a) and (b)) such that for $d$ sufficiently large,*

$$\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0} \leqslant Cd^{-\gamma} \cdot \Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2}^2.$$

**Proposition 2 (Lower bound on the kernel matrix).** *Assume $N = o_d(d^\gamma)$ for a fixed integer $\gamma > 0$. Let $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ independently, and $\sigma$ be an activation function satisfying assumption 2(a). Let $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ be the kernel matrix defined by equation (82). Then there exists a constant $\varepsilon > 0$ that depends on the activation function $\sigma$, such that*

$$\lambda_{\min}(\boldsymbol{U}) \geqslant \varepsilon,$$

*with high probability as $d \to \infty$.*

The proofs of these two propositions are provided in the next sections.

Proposition 1 shows that there exists $\varepsilon_0 > 0$ such that

$$\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[\Vert \boldsymbol{V}_{\mathcal{Q}^c}\Vert_2^2] = N\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0} \leqslant CNd^{-\gamma}\Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2}^2.$$

Hence, by Markov's inequality, we get for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(\Vert \boldsymbol{V}_{\mathcal{Q}^c}\Vert_2 \geqslant \varepsilon\cdot\Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2}) &\leqslant \mathbb{P}(\{\Vert \boldsymbol{V}_{\mathcal{Q}^c}\Vert_2 \geqslant \varepsilon\cdot\Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2}\}\cap\mathcal{P}_{\varepsilon_0}) + \mathbb{P}(\mathcal{P}_{\varepsilon_0}^c)\\ &\leqslant \frac{N\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0}}{\varepsilon^2\Vert \mathsf{P}_{\mathcal{Q}^c}f_d\Vert_{L^2}^2} + o_d(1)\\ &\leqslant C'Nd^{-\gamma} + o_d(1),\end{aligned}$$

where we used lemma 8. By assumption, we have $N = o_d(d^\gamma)$, hence equation (84) is verified. Furthermore equation (85) follows simply from proposition 2. This proves the theorem.

### E.3. Proof of proposition 1

We will denote:

$$\overline{f}_d(\overline{\boldsymbol{x}}) = f_d(\boldsymbol{x}),$$

such that $\overline{f}$ is a function on the normalized product of spheres $\mathrm{PS}^{\mathbf{d}}$. (Note that we defined $\mathsf{P}_{\boldsymbol{k}} f_d(\boldsymbol{x}) \equiv \mathsf{P}_{\boldsymbol{k}} \overline{f}_d(\overline{\boldsymbol{x}})$ the unambiguous polynomial approximation of $f_d$ with polynomial of degree $\mathbf{k}$.) We have

$$\begin{aligned} V_{i,\mathcal{Q}^c} &= \mathbb{E}_{\boldsymbol{x}}\left[ [\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x}) \sigma \left( \sum_{q\in[Q]} \langle \boldsymbol{x}^{(q)}, \boldsymbol{\theta}_i^{(q)} \rangle / R \right) \right] \\ &= \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ [\mathsf{P}_{\mathcal{Q}^c} \overline{f}_d](\overline{\boldsymbol{x}}) \sigma_{\boldsymbol{d},\boldsymbol{\tau}_i} \left( \{ \langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{\theta}}_i^{(q)} \rangle / \sqrt{d_q} \}_{q\in[Q]} \right) \right]. \end{aligned}$$

We recall the expansion of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of tensor product of Gegenbauer polynomials

$$\sigma(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle / R) = \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle \}_{q\in[Q]} \right),$$

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ \sigma_{\boldsymbol{d},\boldsymbol{\tau}} \left( \overline{x}_1^{(1)}, \ldots, \overline{x}_1^{(Q)} \right) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \sqrt{d_1} \overline{x}_1^{(1)}, \ldots, \sqrt{d_Q} \overline{x}_1^{(Q)} \right) \right].$$

For any $\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q$, the spherical harmonics expansion of $P_{\boldsymbol{k}} \overline{f}_d$ gives

$$P_{\boldsymbol{k}} \overline{f}_d(\overline{\boldsymbol{x}}) = \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \lambda_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{f}_d) Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}).$$

Using equation (37) to get the following property

$$\begin{aligned} \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ Q_{\boldsymbol{k}'}^{\boldsymbol{d}} \left( \{ \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle \}_{q\in[Q]} \right) Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}) \right] &= \frac{1}{B(\boldsymbol{d},\boldsymbol{k}')} \sum_{\boldsymbol{s}'\in[B(\boldsymbol{d},\boldsymbol{k})]} Y_{\boldsymbol{k}',\boldsymbol{s}'}^{\boldsymbol{d}}(\overline{\boldsymbol{\theta}}) \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ Y_{\boldsymbol{k}',\boldsymbol{s}'}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}) Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}) \right] \\ &= \frac{1}{B(\boldsymbol{d},\boldsymbol{k})} Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{\theta}}) \delta_{\boldsymbol{k},\boldsymbol{k}'}, \end{aligned} \tag{86}$$

we get

$$\begin{aligned} &\mathbb{E}_{\overline{\boldsymbol{x}}} \left[ [\mathsf{P}_{\boldsymbol{k}} \overline{f}_d](\overline{\boldsymbol{x}}) \sigma_{\boldsymbol{d},\boldsymbol{\tau}} \left( \{ \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle / \sqrt{d_q} \}_{q\in[Q]} \right) \right] \\ &\quad = \sum_{\boldsymbol{k}'\geqslant \mathbf{0}} \lambda_{\boldsymbol{k}'}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}') \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \lambda_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{f}_d) \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle \}_{q\in[Q]} \right) \right] \\ &\quad = \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \lambda_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{f}_d) \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{\theta}}). \end{aligned}$$

Let $\varepsilon_0 > 0$ be a constant as specified in lemma 9. We consider

$$\begin{aligned}
\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0} &= \mathbb{E}_{\overline{\boldsymbol{\theta}},\boldsymbol{\tau}_{\varepsilon_0}}\left[\mathbb{E}_{\overline{\boldsymbol{x}}}\left[[\mathsf{P}_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{x}})\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]^2\right] \\
&= \sum_{\boldsymbol{k},\boldsymbol{k}'\in\mathcal{Q}^c}\mathbb{E}_{\overline{\boldsymbol{\theta}},\boldsymbol{\tau}_{\varepsilon_0}}\left[\mathbb{E}_{\overline{\boldsymbol{x}}}\left[[\mathsf{P}_{\boldsymbol{k}}\overline{f}_d](\overline{\boldsymbol{x}})\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]\right. \\
&\quad\left.\times\,\mathbb{E}_{\overline{\boldsymbol{y}}}\left[[\mathsf{P}_{\boldsymbol{k}'}\overline{f}_d](\overline{\boldsymbol{y}})\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]\right] \\
&= \sum_{\boldsymbol{k},\boldsymbol{k}'\in\mathcal{Q}^c}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})\lambda^{\boldsymbol{d}}_{\boldsymbol{k}'}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})\right] \\
&\quad\times\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\sum_{\boldsymbol{s}'\in[B(\boldsymbol{d},\boldsymbol{k}')]}\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)\lambda^{\boldsymbol{d}}_{\boldsymbol{k}',\boldsymbol{s}'}(\overline{f}_d)\mathbb{E}_{\overline{\boldsymbol{\theta}}}[Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}})Y^{\boldsymbol{d}}_{\boldsymbol{k}',\boldsymbol{s}'}(\overline{\boldsymbol{\theta}})] \\
&= \sum_{\boldsymbol{k}\in\mathcal{Q}^c}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2]\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)^2 \\
&\leqslant \left[\max_{\boldsymbol{k}\in\mathcal{Q}^c}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2]\right]\cdot\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)^2 \\
&= \left[\max_{\boldsymbol{k}\in\mathcal{Q}^c}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2]\right]\cdot\|\mathsf{P}_{\mathcal{Q}^c}\overline{f}_d\|_{L^2}. \qquad (87)
\end{aligned}$$

From lemma 9, there exists a constant $C > 0$ such that for $d$ sufficiently large, we have for any $\boldsymbol{k}\in\mathcal{Q}^c$,

$$\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2]\leqslant\sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2\leqslant Cd^{-\gamma}. \qquad (88)$$

Combining equation (87) and (88) yields

$$\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0}\leqslant Cd^{-\gamma}\cdot\|\mathsf{P}_{\mathcal{Q}^c}\overline{f}_d\|_{L^2}.$$

### E.4. Proof of proposition 2

**Step 1. Construction of the activation functions $\hat{\sigma}$, $\bar{\sigma}$.**

Without loss of generality, we will assume that $q_\xi = 1$. From assumption 2(b), $\sigma$ is not a degree $\lfloor\gamma/\eta_1\rfloor$-polynomial. This is equivalent to having $m\geqslant\lfloor\gamma/\eta_1\rfloor+1$ such that $\mu_m(\sigma)\neq 0$. Let us denote

$$m=\inf\{k\geqslant\lfloor\gamma/\eta_1\rfloor+1|\mu_m(\sigma)\neq 0\}.$$

Recall the expansion of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of product of Gegenbauer polynomials

$$\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)=\sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right),$$

where

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)}\right)Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\ldots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right].$$

Denoting $\boldsymbol{m} = (m, 0, \ldots, 0) \in \mathbb{Z}_{\geqslant 0}^Q$ and using the Gegenbauer coefficients of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$, we define an activation function $\bar{\sigma}_{\boldsymbol{d},\boldsymbol{\tau}}$ which is a degree $m$ polynomial in $\overline{\boldsymbol{x}}^{(1)}$ and do not depend on $\overline{\boldsymbol{x}}^{(q)}$ for $q \geqslant 2$.

$$\begin{aligned}\bar{\sigma}_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) &= \lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{m})Q_{\boldsymbol{m}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\\ &= \lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(d_1,m)Q_m^{(d_1)}(\sqrt{d_1}\overline{x}_1^{(1)}),\end{aligned}$$

and an activation function

$$\hat{\sigma}_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) = \sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\mathbb{Z}_{\geqslant 0}^Q}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right).$$

**Step 2. The kernel functions $u_d$, $\hat{u}_d$ and $\bar{u}_d$.**
Let $u_d$, $\hat{u}_d$ and $\bar{u}_d$ be defined by

$$\begin{aligned}&u_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\\ &\quad= \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle\boldsymbol{\theta}_1, \boldsymbol{x}\rangle/R)\sigma(\langle\boldsymbol{\theta}_2, \boldsymbol{x}\rangle/R)]\\ &\quad= \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_1})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_2})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\end{aligned}\tag{89}$$

and

$$\begin{aligned}&\hat{u}_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\\ &\quad= \mathbb{E}_{\boldsymbol{x}}[\hat{\sigma}(\langle\boldsymbol{\theta}_1, \boldsymbol{x}\rangle/R)\hat{\sigma}(\langle\boldsymbol{\theta}_2, \boldsymbol{x}\rangle/R)]\\ &\quad= \sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\mathbb{Z}_{\geqslant 0}^Q}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_1})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_2})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\end{aligned}\tag{90}$$

and

$$\begin{aligned}\bar{u}_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) &= \mathbb{E}_{\boldsymbol{x}}[\bar{\sigma}(\langle\boldsymbol{\theta}_1, \boldsymbol{x}\rangle/R)\bar{\sigma}(\langle\boldsymbol{\theta}_2, \boldsymbol{x}\rangle/R)]\\ &= \lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_1})\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\tau_2})B(d_1,m)Q_m^{(d_1)}(\langle\overline{\boldsymbol{\theta}}_1^{(1)}, \overline{\boldsymbol{\theta}}_2^{(1)}\rangle).\end{aligned}\tag{91}$$

We immediately have $u_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} = \hat{u}_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} + \bar{u}_{\boldsymbol{d}}^{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}$. Note that all three correspond to positive semi-definite kernels.

**Step 3. Analyzing the kernel matrix.**
Let $\boldsymbol{U}$, $\hat{\boldsymbol{U}}$, $\bar{\boldsymbol{U}} \in \mathbb{R}^{N\times N}$ with

$$\begin{aligned}\boldsymbol{U}_{ij} &= u_{\boldsymbol{d}}^{\tau_i,\tau_j}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right),\\ \hat{\boldsymbol{U}}_{ij} &= \hat{u}_{\boldsymbol{d}}^{\tau_i,\tau_j}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right),\\ \bar{\boldsymbol{U}}_{ij} &= \bar{u}_{\boldsymbol{d}}^{\tau_i,\tau_j}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right).\end{aligned}$$

Since $\hat{\boldsymbol{U}} = \boldsymbol{U} - \bar{\boldsymbol{U}} \succeq 0$, we immediately have $\boldsymbol{U} \succeq \bar{\boldsymbol{U}}$. In the following, we will lower bound $\bar{\boldsymbol{U}}$.

By the decomposition of $\bar{\boldsymbol{U}}$ in terms of Gegenbauer polynomials (91), we have

$$\bar{\boldsymbol{U}} = B(d_1, m)\mathrm{diag}\left(\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}_i})\right)\cdot \boldsymbol{W}_m \cdot \mathrm{diag}\left(\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}_i})\right),$$

where $\boldsymbol{W}_m \in \mathbb{R}^{N\times N}$ with $W_{m,ij} = Q_m^{(d_1)}(\langle \overline{\boldsymbol{\theta}}_i^{(1)}, \overline{\boldsymbol{\theta}}_j^{(1)}\rangle)$. From proposition 6 (recalling that by definition of $m > \gamma/\eta_1$, i.e. $\gamma < m\eta_1$, we have $N < d^{\eta_1 m-\delta} = d_1^{m-\delta'}$ for some $\delta > 0$), we have

$$\|\boldsymbol{W}_m - \mathbf{I}_N\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Hence we get

$$\left\|\overline{\boldsymbol{U}} - B(d_1, m)\mathrm{diag}\left(\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}_i})^2\right)\right\|_{\mathrm{op}} = \max_{i\in[N]}\left\{B(d_1, m)\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}_i})^2\right\}\cdot o_{d,\mathbb{P}}(1). \tag{92}$$

From assumption 2(a) and lemma 20 applied to coefficient $\mathbf{m}$, as well as the assumption that $\mu_m(\sigma) \neq 0$, there exists $\varepsilon_0 > 0$ and $C, c > 0$ such that for $d$ large enough,

$$\begin{aligned}\sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q} B(d_1, m)\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \leqslant C < \infty,\\ \inf_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q} B(d_1, m)\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \geqslant c > 0.\end{aligned} \tag{93}$$

We restrict ourselves to the event $\mathcal{P}_{\varepsilon_0}$ defined in equation (74), which happens with high probability (lemma 8). Hence from equations (92) and (93), we deduce that with high probability

$$\overline{\boldsymbol{U}} = B(d_1, m)\mathrm{diag}\left(\lambda_{\boldsymbol{m}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}_i})^2\right) + o_{d,\mathbb{P}}(1) \succeq \frac{c}{2}\mathbf{I}_N.$$

We conclude that with high probability

$$\boldsymbol{U} = \bar{\boldsymbol{U}} + \hat{\boldsymbol{U}} \succeq \overline{\boldsymbol{U}} \succeq \frac{c}{2}\mathbf{I}_N.$$

## Appendix F. Proof of theorem 6(b): upper bound for RF model

### F.1. Preliminaries

**Lemma 10.** *Let $\sigma$ be an activation function that satisfies assumptions 2(a) and (b). Let $\|\boldsymbol{w}^{(q)}\|_2 = 1$ be unit vectors of $\mathbb{R}^{d_q}$, for $q = 1, \dots, Q$. Fix $\gamma > 0$ and denote $\mathcal{Q} = \overline{\mathcal{Q}}_{\mathrm{RF}}(\gamma)$. Then there exists $\varepsilon_0 > 0$ and $d_0$ and constants $C, c > 0$ such that for $d \geqslant d_0$ and $\boldsymbol{\tau} \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$,*

$$\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] \leqslant C < \infty, \tag{94}$$

$$\min_{\boldsymbol{k}\in\mathcal{Q}} \lambda^{\boldsymbol{d}}_{k,0}(\sigma_{d,\tau})^2 \geqslant cd^{-\gamma} > 0. \tag{95}$$

**Proof of lemma 10.** The first inequality comes simply from assumption 2(a) and lemma 17(b). For the second inequality, notice that by assumption 2(c) we can apply lemma 19 to any $\boldsymbol{k} \in \mathcal{Q}$. Hence (using that $\mu_k(\sigma)^2 > 0$ and we can choose $\delta$ sufficiently small), we deduce that there exists $c > 0$, $\varepsilon_0 > 0$ and $d_0$ such that for any $d \geqslant d_0$, $\boldsymbol{\tau} \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$ and $\boldsymbol{k} \in \mathcal{Q}$,

$$\left(\prod_{q\in[Q]} d^{(\xi-\eta_q-\kappa_q)k_q}\right) B(\boldsymbol{d},\boldsymbol{k})\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \geqslant c > 0.$$

Furthermore, using that $B(\boldsymbol{d},\boldsymbol{k}) = \Theta(d_1^{k_1} d_2^{k_2} \dots d_Q^{k_Q})$, there exists $c' > 0$ such that for any $\boldsymbol{k} \in \mathcal{Q}$,

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \geqslant c' \prod_{q\in[Q]} d^{(\eta_q+\kappa_q-\xi)k_q} d_q^{k_q} = c' \prod_{q\in[Q]} d^{(\kappa_q-\xi)k_q} \geqslant cd^{-\gamma},$$

where we used in the last inequality $\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{RF}}(\gamma)$ implies $(\xi-\kappa_1)k_1 + \dots + (\xi-\kappa_Q)k_Q \leqslant \gamma$ by definition. □

### F.2. Properties of the limiting kernel

Similarly to the proof of theorem 1(b) in [21], we construct a limiting kernel which is used as a proxy to upper bound the RF risk.

We recall the definition of $\mathrm{PS}^{\boldsymbol{d}} = \prod_{q\in[Q]} \mathbb{S}^{d_q-1}(\sqrt{d_q})$ and $\mu_{\boldsymbol{d}} = \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$. Let us denote $\mathcal{L} = L^2(\mathrm{PS}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}})$. Fix $\boldsymbol{\tau} \in \mathbb{R}^Q_{>0}$ and recall the definition for a given $\boldsymbol{\theta} = (\overline{\boldsymbol{\theta}}, \boldsymbol{\tau})$ of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \cdot\rangle/\sqrt{d_q}\}) \in \mathcal{L}$,

$$\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}_q, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) = \sigma\left(\sum_{q\in[Q]} \tau^{(q)}(r_q/R)\langle \overline{\boldsymbol{\theta}}_q, \overline{\boldsymbol{x}}^{(q)}\rangle\right).$$

Define the operator $\mathbb{T}_{\boldsymbol{\tau}} : \mathcal{L} \to \mathcal{L}$, such that for any $g \in \mathcal{L}$,

$$\mathbb{T}_{\boldsymbol{\tau}} g(\overline{\boldsymbol{\theta}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle / \sqrt{d_q}\}_{q\in[Q]}\right) g(\overline{\boldsymbol{x}})\right].$$

It is easy to check that the adjoint operator $\mathbb{T}^*_{\boldsymbol{\tau}} : \mathcal{L} \to \mathcal{L}$ verifies $\mathbb{T}^* = \mathbb{T}$ with variables $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{\theta}}$ exchanged.

We define the operator $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} : \mathcal{L} \to \mathcal{L}$ as $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} \equiv \mathbb{T}_{\boldsymbol{\tau}} \mathbb{T}^*_{\boldsymbol{\tau}'}$. For $g \in \mathcal{L}$, we can write

$$\mathbb{K}_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} g(\overline{\boldsymbol{\theta}}_1) = \mathbb{E}_{\overline{\boldsymbol{\theta}}_2}[K_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2) g(\overline{\boldsymbol{\theta}}_2)],$$

where

$$K_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}_1}\left(\{\langle \overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle / \sqrt{d_q}\}_{q\in[Q]}\right) \sigma_{\boldsymbol{d},\boldsymbol{\tau}_2}\left(\{\langle \overline{\boldsymbol{\theta}}_2^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle / \sqrt{d_q}\}_{q\in[Q]}\right)\right].$$

We recall the decomposition of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of tensor product of Gegenbauer polynomials

$$\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\overline{x}_1^{(q)}\}_{q\in[Q]}\right) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\overline{x}_1^{(q)}\}_{q\in[Q]}\right),$$

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\overline{x}_1^{(q)}\}_{q\in[Q]}\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\sqrt{d_q}\,\overline{x}_1^{(q)}\}_{q\in[Q]}\right)\right].$$

Recall that $\{Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}\}_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0},\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{s})]}$ forms an orthonormal basis of $\mathcal{L}$. From equation (86), we have for any $\boldsymbol{k} \geqslant 0$ and $\boldsymbol{s} \in [B(\boldsymbol{d},\boldsymbol{k})]$

$$\begin{aligned}\mathbb{T}_{\boldsymbol{\tau}} Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}) &= \sum_{\boldsymbol{k}'\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}'}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}') \mathbb{E}_{\overline{\boldsymbol{x}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}})\right]\\ &= \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}),\end{aligned}$$

where we used

$$\mathbb{E}_{\overline{\boldsymbol{x}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{x}})\right] = \frac{\delta_{\boldsymbol{k},\boldsymbol{k}'}}{B(\boldsymbol{d},\boldsymbol{k})} \boldsymbol{Y}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}).$$

The same equation holds for $\mathbb{T}^*_{\boldsymbol{\tau}}$. Therefore, we directly deduce that

$$\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}) = (\mathbb{T}_{\boldsymbol{\tau}} \mathbb{T}^*_{\boldsymbol{\tau}'}) Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}) = \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}'}) Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{\theta}}).$$

We deduce that $\{Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}\}_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0},\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{s})]}$ is an orthonormal basis that diagonalizes the operator $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'}$.

Let $\varepsilon_0 > 0$ be defined as in lemma 10. We will consider $\boldsymbol{\tau}, \boldsymbol{\tau}' \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$ and restrict ourselves to the subspace $V^{\boldsymbol{d}}_{\mathcal{Q}}$. From the choice of $\varepsilon_0$ and for $d$ large enough, the eigenvalues $\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}'}) \neq 0$ for any $\boldsymbol{k} \in \mathcal{Q}$. Hence, the operator $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}}$ is invertible.

### F.3. Proof of theorem 6(b)

Without loss of generality, let us assume that $\{f_d\}$ are polynomials contained in $V^{\boldsymbol{d}}_{\mathcal{Q}}$, i.e. $\overline{f}_d = \mathsf{P}_{\mathcal{Q}}\overline{f}_d$.

Consider

$$\hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle / R).$$

Define $\alpha_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}) \equiv \mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d(\overline{\boldsymbol{\theta}})$ and choose $a^*_i = N^{-1}\alpha_{\boldsymbol{\tau}_i}(\overline{\boldsymbol{\theta}}_i)$, where we denoted $\overline{\boldsymbol{\theta}}_i = (\overline{\boldsymbol{\theta}}^{(q)}_i)_{q\in[Q]}$ with $\overline{\boldsymbol{\theta}}^{(q)}_i = \boldsymbol{\theta}^{(q)}_i/\tau^{(q)}_i \in \mathbb{S}^{d_q-1}(\sqrt{d_q})$ and $\tau^{(q)}_i = \|\boldsymbol{\theta}^{(q)}_i\|_2/\sqrt{d_q}$ independent of $\overline{\boldsymbol{\theta}}^{(q)}_i$.

Let $\varepsilon_0 > 0$ be defined as in lemma 10 and consider the expectation over $\mathcal{P}_{\varepsilon_0}$ of the RF risk (in particular, $\boldsymbol{a}^* = (a^*_1, \ldots, a^*_N)$ are well defined):

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[R_{\mathrm{RF}}(f_d, \boldsymbol{\Theta})] &= \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\inf_{\boldsymbol{a}\in\mathbb{R}^N} \mathbb{E}_{\boldsymbol{x}}[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}))^2]\right]\\ &\leqslant \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}^*(\boldsymbol{\Theta})))^2\right]\right].\end{aligned}$$

We can expand the squared loss at $\boldsymbol{a}^*$ as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}^*))^2] &= \|f_d\|^2_{L^2} - 2\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x}}[a^*_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R) f_d(\boldsymbol{x})]\\ &\quad + \sum_{i,j=1}^{N} \mathbb{E}_{\boldsymbol{x}}[a^*_i a^*_j \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)]. \qquad (96)\end{aligned}$$

The second term of the expansion (96) around $\boldsymbol{a}^*$ verifies

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x}}\left[a^*_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R) f_d(\boldsymbol{x})\right]\right]\\ &\quad = \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\alpha_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}})\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\overline{f}_d(\overline{\boldsymbol{x}})\right]\right]\right]\\ &\quad = \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\langle \mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d, \mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\rangle_{L^2}\right]\\ &\quad = \|f_d\|^2_{L^2}, \qquad (97)\end{aligned}$$

where we used that for each $\boldsymbol{\tau} \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$, we have $\mathbb{T}^*_{\boldsymbol{\tau}}\mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}} = \mathbf{I}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}}$.

Let us consider the third term in the expansion (96) around $\boldsymbol{a}^*$: the non diagonal term verifies

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\neq j}\mathbb{E}_{\boldsymbol{x}}\left[a_i^* a_j^* \sigma(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\sigma(\langle\boldsymbol{\theta}_j,\boldsymbol{x}\rangle/R)\right]\right]\\
&\quad=\left(1-N^{-1}\right)\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0},\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2}\Big[\alpha_{\boldsymbol{\tau}^1}(\overline{\boldsymbol{\theta}}_1)\alpha_{\boldsymbol{\tau}^2}(\overline{\boldsymbol{\theta}}_2)\\
&\qquad\times E_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}^1}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\sigma_{\boldsymbol{d},\boldsymbol{\tau}^2}\left(\{\langle\overline{\boldsymbol{\theta}}_2^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]\Big]\\
&\quad=\left(1-N^{-1}\right)\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0},\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2}\left[\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\overline{f}_d(\overline{\boldsymbol{\theta}}_1)\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}(\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2)\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}T_{\boldsymbol{\tau}^2}\overline{f}_d(\overline{\boldsymbol{\theta}}_2)\right]\\
&\quad=\left(1-N^{-1}\right)\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0}}\left[\langle\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\overline{f}_d,\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}\overline{f}_d\rangle_{L^2}\right].
\end{aligned}
$$

For $\boldsymbol{k}\in\mathcal{Q}$ and $\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]$ and $\boldsymbol{\tau}^1,\boldsymbol{\tau}^2\in[1-\varepsilon_0,1+\varepsilon_0]^Q$, we have (for $d$ large enough)

$$
\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}=\left(\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\right)\cdot\left(\mathbb{T}^*_{\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}\right)\cdot Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}=Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}.
$$

Hence for any $\boldsymbol{\tau}^1,\boldsymbol{\tau}^2\in[1-\varepsilon_0,1+\varepsilon_0]^Q$, $\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}}=\mathbf{I}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}}$. Hence

$$
\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\neq j}\mathbb{E}_{\boldsymbol{x}}\left[a_i^* a_j^* \sigma(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\sigma(\langle\boldsymbol{\theta}_j,\boldsymbol{x}\rangle/R)\right]\right]=\left(1-N^{-1}\right)\|f_d\|^2_{L^2}. \tag{98}
$$

The diagonal term verifies

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\in[N]}\mathbb{E}_{\boldsymbol{x}}\left[(a_i^*)^2\sigma(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)^2\right]\right]\\
&\quad=N^{-1}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0},\overline{\boldsymbol{\theta}}}\left[\alpha_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}\,)^2K_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\right]\\
&\quad\leqslant N^{-1}\left[\max_{\overline{\boldsymbol{\theta}},\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q}K_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\right]\cdot\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\|\mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|^2_{L^2}].
\end{aligned}
$$

We have by definition of $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}$

$$
\sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q}K_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})=\sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q}\|\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\|^2_{L^2}\leqslant C,
$$

for $d$ large enough (using lemma 10). Furthermore

$$
\begin{aligned}
\|\mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|^2_{L^2}&=\sum_{\boldsymbol{k}\in\mathcal{Q}}\frac{1}{\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2}\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)^2\\
&\leqslant\left[\max_{\boldsymbol{k}\in\mathcal{Q}}\frac{1}{\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2}\right]\cdot\|\mathsf{P}_{\mathcal{Q}}f_d\|^2_{L^2}.
\end{aligned}
$$

From lemma 10, we get

$$\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\|\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|_{L^2}^2] \leqslant Cd^{\gamma} \cdot \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2.$$

Hence,

$$\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\in[N]}\mathbb{E}_{\boldsymbol{x}}\left[(a_i^*)^2\sigma(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)^2\right]\right] \leqslant C\frac{d^{\gamma}}{N}\|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2. \tag{99}$$

Combining equations (97)–(99), we get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[R_{\mathrm{RF}}(f_d,\boldsymbol{\Theta})]\\
&\quad\leqslant \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\left(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x};\boldsymbol{\Theta},\boldsymbol{a}^*(\boldsymbol{\Theta}))\right)^2\right]\right]\\
&\quad= \|f_d\|_{L^2}^2 - 2\|f_d\|_{L^2}^2 + (1-N^{-1})\|f_d\|_{L^2}^2 + N^{-1}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0},\overline{\boldsymbol{\theta}}}\left[(\alpha_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}\,))^2 K_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\right]\\
&\quad\leqslant C\frac{d^{\gamma}}{N}\|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2.
\end{aligned}$$

By Markov's inequality, we get for any $\varepsilon > 0$ and $d$ large enough,

$$\begin{aligned}
\mathbb{P}(R_{\mathrm{RF}}(f_d,\boldsymbol{\Theta}) > \varepsilon\cdot\|f_d\|_{L^2}^2) &\leqslant \mathbb{P}(\{R_{\mathrm{RF}}(f_d,\boldsymbol{\Theta}) > \varepsilon\cdot\|f_d\|_{L^2}^2\}\cap\mathcal{P}_{\varepsilon_0}) + \mathbb{P}(\mathcal{P}_{\varepsilon_0}^c)\\
&\leqslant C'\frac{d^{\gamma}}{N} + \mathbb{P}(\mathcal{P}_{\varepsilon_0}^c).
\end{aligned}$$

The assumption $N = \omega_d(d^{\gamma})$ and lemma 8 conclude the proof.

## Appendix G. Proof of theorem 7(a): lower bound for NT model

### G.1. Preliminaries

We consider the activation function $\sigma:\mathbb{R}\to\mathbb{R}$ with weak derivative $\sigma'$. Consider $\sigma'_{d,\boldsymbol{\tau}}:\mathrm{ps}^{\boldsymbol{d}}\to\mathbb{R}$ defined as follows

$$\begin{aligned}
\sigma'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R) &= \sigma'\left(\sum_{q\in[Q]}\tau^{(q)}\cdot(r_q/R)\cdot\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\right)\\
&\equiv \sigma'_{d,\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right).
\end{aligned}\tag{100}$$

Consider the expansion of $\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of product of Gegenbauer polynomials. We have

$$\sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle/R) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right), \tag{101}$$

where

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}^{(1)}_1, \dots, \overline{x}^{(Q)}_1\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\,\overline{x}^{(1)}_1, \dots, \sqrt{d_Q}\,\overline{x}^{(Q)}_1\right)\right],$$

where the expectation is taken over $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(Q)}) \sim \mu_{\boldsymbol{d}}$.

**Lemma 11.** *Let $\sigma$ be an activation function that satisfies assumptions 3(a) and (b). Define for $\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}$ and $\boldsymbol{\tau}\in\mathbb{R}^Q_{\geqslant 0}$,*

$$A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}} = r_q^2 \cdot [t_{d_q,k_q-1}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k}_{q-}) + s_{d_q,k_q+1}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k}_{q+})], \tag{102}$$

*with $\mathbf{k}_{q+} = (k_1, \dots, k_q+1, \dots, k_Q)$ and $\mathbf{k}_{q-} = (k_1, \dots, k_q-1, \dots, k_Q)$, and*

$$s_{d,k} = \frac{k}{2k+d-2}, \qquad t_{d,k} = \frac{k+d-2}{2k+d-2},$$

*with the convention $t_{d,-1} = 0$. Then there exists constants $\varepsilon_0 > 0$ and $C > 0$ such that for $d$ large enough, we have for any $\boldsymbol{\tau}\in[1-\varepsilon_0, 1+\varepsilon_0]^Q$ and $\boldsymbol{k}\in\mathcal{Q}_{\mathrm{NT}}(\gamma)^c$,*

$$\frac{A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}}}{B(\boldsymbol{d},\boldsymbol{k})} \leqslant \begin{cases} Cd^{\xi} d^{-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)} & \text{if } k_q > 0, \\ Cd^{\eta_q+2\kappa_q-\xi} d^{-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)} & \text{if } k_q = 0, \end{cases}$$

*where we recall $S(\mathbf{k}) \subset [Q]$ is the subset of indices corresponding to the non zero integers $k_q > 0$.*

**Proof of lemma 11.** Let us fix an integer $M$ such that $\mathcal{Q} \subset [M]^Q$. We will denote $\mathcal{Q} \equiv \mathcal{Q}_{\mathrm{NT}}(\gamma)$ for simplicity. Following the same proof as in lemma 9, there exists $\varepsilon_0 > 0$, $d_0$ and $C > 0$ such that for any $d \geqslant d_0$ and $\boldsymbol{\tau}\in[1-\varepsilon_0, 1+\varepsilon_0]^Q$, we have for any $\boldsymbol{k}\in\mathcal{Q}^c\cap[M]^Q$,

$$\begin{aligned} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 &\leqslant Cd^{-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)}, \\ \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 &\leqslant Cd^{\xi-\kappa_q-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)}, \\ \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 &\leqslant Cd^{\kappa_q-\xi-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)}, \end{aligned}$$

while for $\boldsymbol{k}\notin[M]^Q$, we get

$$\max\{\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2, \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2, \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2\} \leqslant Cd^{-(M-1)\min_{q\in[Q]}\eta_q}.$$

Injecting this bound in the formula (102) of $\boldsymbol{A}_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}$, we get for $d \geqslant d_0$, $\boldsymbol{\tau} \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$ and any $\boldsymbol{k} \in \mathcal{Q}^c \cap [M]^Q$: if $k_q > 0$,

$$\frac{A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})} \leqslant C' d^{\eta_q+\kappa_q} d^{\xi-\kappa_q-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)} = C' d^{\xi} d^{-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)},$$

while for $k_q = 0$,

$$\frac{A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})} \leqslant C' d^{\eta_q+\kappa_q} d^{-\eta_q} d^{\kappa_q+\eta_q-\xi-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)} = C' d^{\eta_q+2\kappa_q-\xi} d^{-\gamma-(\xi-\min_{q\in S(\boldsymbol{k})}\kappa_q)},$$

where we used that for $k_q \in [M]$, there exists a constant $c > 0$ such that $s_{d_q,k_q} \leqslant c d^{-\eta_q}$ and $t_{d_q,k_q} \leqslant c$. Similarly, we get for $\boldsymbol{k} \notin [M]^Q$

$$\frac{A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})} \leqslant C'' d^{\kappa_q+\eta_q-(M-1)\min_{q\in[Q]}\eta_q},$$

where we used that $s_{d_q,k}, t_{d_q,k} \leqslant 1$ for any $k \in \mathbb{Z}_{\geqslant 0}$. Taking $M$ sufficiently large yields the result. □

### G.2. Proof of theorem 7(a): outline

The structure of the proof for the NT model is the same as for the RF case, however some parts of the proof requires more work.

We define the random vector $\boldsymbol{V} = (\boldsymbol{V}_1, \dots, \boldsymbol{V}_N)^{\mathsf{T}} \in \mathbb{R}^{Nd}$, where, for each $j \leqslant N$, $\boldsymbol{V}_j \in \mathbb{R}^D$, and analogously $\boldsymbol{V}_{\mathcal{Q}} = (\boldsymbol{V}_{1,\mathcal{Q}}, \dots, \boldsymbol{V}_{N,\mathcal{Q}})^{\mathsf{T}} \in \mathbb{R}^{ND}$, $\boldsymbol{V}_{\mathcal{Q}^c} = (\boldsymbol{V}_{1,\mathcal{Q}^c}, \dots, \boldsymbol{V}_{N,\mathcal{Q}^c})^{\mathsf{T}} \in \mathbb{R}^{ND}$, as follows

$$\begin{aligned}
\boldsymbol{V}_{i,\mathcal{Q}} &= \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}} f_d](\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\boldsymbol{x}],\\
\boldsymbol{V}_{i,\mathcal{Q}^c} &= \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\boldsymbol{x}],\\
\boldsymbol{V}_i &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\boldsymbol{x}] = \boldsymbol{V}_{i,\mathcal{Q}} + \boldsymbol{V}_{i,\mathcal{Q}^c}.
\end{aligned}$$

We define the random matrix $\boldsymbol{U} = (\boldsymbol{U}_{ij})_{i,j\in[N]} \in \mathbb{R}^{ND\times ND}$, where for each $i, j \leqslant N$, $\boldsymbol{U}_{ij} \in \mathbb{R}^{D\times D}$, is given by

$$\boldsymbol{U}_{ij} = \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle\boldsymbol{x},\boldsymbol{\theta}_i\rangle/R)\sigma'(\langle\boldsymbol{x},\boldsymbol{\theta}_j\rangle/R)\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}]. \tag{103}$$

Proceeding as for the RF model, we obtain

$$\begin{aligned}
&\left|R_{\mathrm{NT}}(f_d) - R_{\mathrm{NT}}(\mathsf{P}_{\mathcal{Q}} f_d) - \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2\right|\\
&\quad = \left|\boldsymbol{V}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}} - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}\right|\\
&\quad = \left|\boldsymbol{V}_{\mathcal{Q}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}} - (\boldsymbol{V}_{\mathcal{Q}} + \boldsymbol{V}_{\mathcal{Q}^c})^{\mathsf{T}}\boldsymbol{U}^{-1}(\boldsymbol{V}_{\mathcal{Q}} + \boldsymbol{V}_{\mathcal{Q}^c})\right|\\
&\quad = \left|2\boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}^c} - \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}^c}\right|\\
&\quad \leqslant 2\|\boldsymbol{U}^{-1/2}\boldsymbol{V}_{\mathcal{Q}^c}\|_2\|f_d\|_{L^2} + \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\mathcal{Q}^c}.
\end{aligned}$$

We claim that we have

$$\| \boldsymbol{U}^{-1/2} \boldsymbol{V}_{\mathcal{Q}^c} \|_2^2 = \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{\mathcal{Q}^c} = o_{d,\mathbb{P}}(\|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2). \tag{104}$$

To show this result, we will need the following two propositions.

**Proposition 3 (Expected norm of V).** *Let $\sigma$ be a weakly differentiable activation function with weak derivative $\sigma'$ and $\mathcal{Q} \subset \mathbb{Z}_{\geqslant 0}^Q$. Let $\varepsilon > 0$ and define $\mathcal{E}_{\mathcal{Q}^c,\varepsilon}^{(q)}$ by*

$$\mathcal{E}_{\mathcal{Q}^c,\varepsilon}^{(q)} \equiv \mathbb{E}_{\boldsymbol{\theta}_\varepsilon} \left[ \langle \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x}) \sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle / R) \boldsymbol{x}^{(q)}], \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x}) \sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle / R) \boldsymbol{x}^{(q)}] \rangle \right],$$

*where the expectation is taken with respect to $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(Q)}) \sim \mu_{\boldsymbol{d}}^{\boldsymbol{\kappa}}$. Then,*

$$\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0}^{(q)} \leqslant \left[ \max_{\boldsymbol{k} \in \mathcal{Q}^c} B(\boldsymbol{d}, \boldsymbol{k})^{-1} \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}} [A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}] \right] \cdot \|\mathsf{P}_{\mathcal{Q}^c} f_d\|_{L^2}^2.$$

**Proposition 4 (Lower bound on the kernel matrix).** *Let $N = o_d(d^\gamma)$ for some $\gamma > 0$, and $(\boldsymbol{\theta}_i)_{i \in [N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ independently. Let $\sigma$ be an activation that satisfies assumptions 3(a) and (b). Let $\boldsymbol{U} \in \mathbb{R}^{ND \times ND}$ be the kernel matrix with $i, j$ block $\boldsymbol{U}_{ij} \in \mathbb{R}^{D \times D}$ defined by equation (103). Then there exists two matrices $\mathbf{D}$ and $\boldsymbol{\Delta}$ such that*

$$\boldsymbol{U} \succeq \boldsymbol{D} + \boldsymbol{\Delta},$$

*with $\mathbf{D} = \mathrm{diag}(\mathbf{D}_{ii})$ block diagonal. Furthermore, $\mathbf{D}$ and $\boldsymbol{\Delta}$ verifies the following properties:*

*(a)* $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(d^{-\max_{q \in [Q]} \kappa_q})$.

*(b) For each $i \in [N]$, we can decompose the matrix $\boldsymbol{D}_{ii}$ into block matrix form $(\boldsymbol{D}_{ii}^{qq'})_{q,q' \in [Q]} \in \mathbb{R}^{DN \times DN}$ with $\boldsymbol{D}_{ii}^{qq'} \in \mathbb{R}^{d_q N \times d_{q'} N}$ such that*

- *For any $q \in [Q]$, there exists constants $c_q, C_q > 0$ such that we have with high probability*

$$\begin{aligned} 0 < c_q \frac{r_q^2}{d_q} = c_q d^{\kappa_q} \leqslant \min_{i \in [N]} \lambda_{\min}(\boldsymbol{D}_{ii}^{qq}) \leqslant \max_{i \in [N]} \lambda_{\max}(\boldsymbol{D}_{ii}^{qq}) \\ \leqslant C_q \frac{r_q^2}{d_q} = C_q d^{\kappa_q} < \infty, \end{aligned} \tag{105}$$

  *as $d \to \infty$.*
- *For any $q \neq q' \in [Q]$, we have*

$$\max_{i \in [N]} \sigma_{\max}(\boldsymbol{D}_{ii}^{qq'}) = o_{d,\mathbb{P}}(r_q r_{q'} / \sqrt{d_q d_{q'}}). \tag{106}$$

The proofs of these two propositions are provided in the next sections.

From proposition 4, we can upper bound equation (104) as follows

$$\boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{\mathcal{Q}^c} \preceq \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} (\boldsymbol{D} + \boldsymbol{\Delta})^{-1} \boldsymbol{V}_{\mathcal{Q}^c} = \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} \boldsymbol{D}^{-1} \boldsymbol{V}_{\mathcal{Q}^c} - \boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} \boldsymbol{D}^{-1} \boldsymbol{\Delta} (\boldsymbol{D} + \boldsymbol{\Delta})^{-1} \boldsymbol{V}_{\mathcal{Q}^c}. \tag{107}$$

Let us fix $\varepsilon_0 > 0$ as prescribed in lemma 11. We decompose the vector $\boldsymbol{V}_{i,\mathcal{Q}^c} = (\boldsymbol{V}_{i,\mathcal{Q}^c}^{(q)})_{q\in[Q]}$ where

$$\boldsymbol{V}_{i,\mathcal{Q}^c}^{(q)} = \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c} f_d](\boldsymbol{x}) \sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle / R) \boldsymbol{x}^{(q)}].$$

We denote $\boldsymbol{V}_{\mathcal{Q}^c}^{(q)} = (\boldsymbol{V}_{1,\mathcal{Q}^c}^{(q)}, \ldots, \boldsymbol{V}_{N,\mathcal{Q}^c}^{(q)}) \in \mathbb{R}^{d_q N}$. From proposition 3, we have

$$\frac{d_q}{r_q^2} \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[\| \boldsymbol{V}_{\mathcal{Q}^c}^{(q)} \|_2^2] \leqslant \left[ \max_{\boldsymbol{k}\in\mathcal{Q}^c} N d^{-\kappa_q} B(\boldsymbol{d}, \boldsymbol{k})^{-1} \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}] \right] \cdot \| \mathsf{P}_{\mathcal{Q}^c} f_d \|_{L^2}^2.$$

Hence, using the upper bounds on $A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}$ in lemma 11, we get for $\boldsymbol{k} \in \mathcal{Q}^c$ with $k_q > 0$:

$$N d^{-\kappa_q} B(\boldsymbol{d}, \boldsymbol{k})^{-1} \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}] \leqslant C N d^{-\kappa_q} d^{-\gamma + \min_{q\in S(\boldsymbol{k})} \kappa_q} = o_d(1),$$

where we used that $N = o_d(d^{\gamma})$ and $\kappa_q \geqslant \min_{q\in S(\boldsymbol{k})} \kappa_q$ (we have $k_q > 0$ and therefore $q \in S(\boldsymbol{k})$ by definition). Similarly for $\boldsymbol{k} \in \mathcal{Q}^c$ with $k_q = 0$:

$$N d^{-\kappa_q} B(\boldsymbol{d}, \boldsymbol{k})^{-1} \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}] \leqslant C N d^{\eta_q + \kappa_q - \xi} d^{-\gamma - (\xi - \min_{q\in S(\boldsymbol{k})} \kappa_q)} = o_d(1),$$

where we used that by definition of $\xi$ we have $\eta_q + \kappa_q \leqslant \xi$ and $\min_{q\in S(\boldsymbol{k})} \kappa_q \leqslant \xi$. We deduce that

$$\frac{d_q}{r_q^2} \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[\| \boldsymbol{V}_{\mathcal{Q}^c}^{(q)} \|_2^2] = o_d(1) \cdot \| \mathsf{P}_{\mathcal{Q}^c} f_d \|_{L^2}^2,$$

and therefore by Markov's inequality that

$$\frac{d_q}{r_q^2} \| \boldsymbol{V}_{\mathcal{Q}^c}^{(q)} \|_2^2 = o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\mathcal{Q}^c} f_d \|_{L^2}^2. \tag{108}$$

Notice that the properties (105) and (106) imply that there exists $c > 0$ such that with high probability $\lambda_{\min}(\boldsymbol{D}) \geqslant \min_{i\in[N]} \lambda_{\min}(\boldsymbol{D}_{ii}) \geqslant c$. In particular, we deduce that $\|(\boldsymbol{D} + \boldsymbol{\Delta})^{-1}\|_{\mathrm{op}} \leqslant c^{-1}/2$ with high probability. Combining these bounds and equation (108) and recalling that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(d^{-\max_{q\in[Q]} \kappa_q})$ show that

$$\begin{aligned} |\boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}} \boldsymbol{D}^{-1} \boldsymbol{\Delta} (\boldsymbol{D} + \boldsymbol{\Delta})^{-1} \boldsymbol{V}_{\mathcal{Q}^c}| &\leqslant \|\boldsymbol{D}^{-1}\|_{\mathrm{op}} \|(\boldsymbol{D} + \boldsymbol{\Delta})^{-1}\|_{\mathrm{op}} \sum_{q\in[Q]} \|\boldsymbol{\Delta}\|_{\mathrm{op}} \| \boldsymbol{V}_{\mathcal{Q}^c}^{(q)} \|_2^2 \\ &= o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\mathcal{Q}^c} f_d \|_{L^2}^2. \end{aligned} \tag{109}$$

We are now left to show $\boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{V}_{\mathcal{Q}^c} = o_{d,\mathbb{P}}(1)$. For each $i \in [N]$, denote $\boldsymbol{B}_{ii} = \boldsymbol{D}_{ii}^{-1}$ and notice that we can apply lemma 24 to $\boldsymbol{B}_{ii}$ and get

$$\max_{i\in[N]} \|\boldsymbol{B}_{ii}^{qq}\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(\frac{d_q}{r_q^2}\right), \qquad \max_{i\in[N]} \|\boldsymbol{B}_{ii}^{qq'}\|_{\mathrm{op}} = o_{d,\mathbb{P}}\left(\frac{\sqrt{d_q d_{q'}}}{r_q r_{q'}}\right).$$

Therefore,

$$\begin{aligned}\boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{V}_{\mathcal{Q}^c} &= \sum_{i\in[N]}\sum_{q,q'\in[Q]} (\boldsymbol{V}_{i,\mathcal{Q}^c}^{(q)})^{\mathsf{T}}\boldsymbol{B}_{ii}^{qq'}\boldsymbol{V}_{i,\mathcal{Q}^c}^{(q')} \\ &\leqslant \sum_{q,q'\in[Q]} O_{d,\mathbb{P}}(1)\cdot\left(\frac{d_q}{r_q^2}\|\boldsymbol{V}_{\mathcal{Q}^c}^{(q)}\|_2^2\right)^{1/2}\left(\frac{d_{q'}}{r_{q'}^2}\|\boldsymbol{V}_{\mathcal{Q}^c}^{(q')}\|_2^2\right)^{1/2}. \qquad (110)\end{aligned}$$

Using equation (108) in equation (110), we get

$$\boldsymbol{V}_{\mathcal{Q}^c}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{V}_{\mathcal{Q}^c} = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\mathcal{Q}^c}f_d\|_{L^2}^2. \qquad (111)$$

Combining equations (109) and (111) yields equation (104). This proves the theorem.

### G.3. Proof of proposition 3

**Proof of proposition 3.** Let us consider $\varepsilon_0 > 0$ as prescribed in lemma 11. We have for $q \in [Q]$

$$\begin{aligned}\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0}^{(q)} &= \mathbb{E}_{\boldsymbol{\theta}_\varepsilon}\left[\langle\mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\mathcal{Q}^c}f_d](\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R)\boldsymbol{x}^{(q)}],\mathbb{E}_{\boldsymbol{y}}[[\mathsf{P}_{\mathcal{Q}^c}f_d](\boldsymbol{y})\sigma'(\langle\boldsymbol{\theta},\boldsymbol{y}\rangle/R)\boldsymbol{y}^{(q)}]\rangle\right] \\ &= \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}\left[[\mathsf{P}_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{x}})[\mathsf{P}_{\mathcal{Q}^c}\overline{f}](\overline{\boldsymbol{y}})H_{\boldsymbol{\tau}}^{(q)}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}})\right]\right],\end{aligned}$$

where we denoted $H_{\boldsymbol{\tau}}^{(q)}$ the kernel given by

$$H_{\boldsymbol{\tau}}^{(q)}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}) = \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]\langle\boldsymbol{x}^{(q)},\boldsymbol{y}^{(q)}\rangle.$$

Then we have

$$H_{\boldsymbol{\tau}}^{(q)}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}) = \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)} Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right), \qquad (112)$$

where $A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}$ is given in lemma 12. Hence we get

$$\begin{aligned}\mathcal{E}_{\mathcal{Q}^c,\varepsilon_0}^{(q)} &= \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}[P_{\mathcal{Q}^c}\overline{f}_d(\overline{\boldsymbol{x}})P_{\mathcal{Q}^c}\overline{f}_d(\overline{\boldsymbol{y}})H_{\boldsymbol{\tau}}^{(q)}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}})]] \\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)}]\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}\left[Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)[P_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{x}})[P_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{y}})\right].\end{aligned}$$

We have

$$
\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)[P_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{x}})[P_{\mathcal{Q}^c}\overline{f}_d](\overline{\boldsymbol{y}})\right]\\
&\quad=\sum_{\boldsymbol{l},\boldsymbol{l}'\in\mathcal{Q}^c}\ \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{l})]}\ \sum_{\boldsymbol{s}'\in[B(\boldsymbol{d},\boldsymbol{l}')]}\lambda^{\boldsymbol{d}}_{\boldsymbol{l},\boldsymbol{s}}(\overline{f}_d)\lambda^{\boldsymbol{d}}_{\boldsymbol{l}',\boldsymbol{s}'}(\overline{f}_d)\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)Y^{\boldsymbol{d}}_{\boldsymbol{l},\boldsymbol{s}}(\overline{\boldsymbol{x}})Y^{\boldsymbol{d}}_{\boldsymbol{l}',\boldsymbol{s}'}(\overline{\boldsymbol{y}})\right]\\
&\quad=\delta_{\boldsymbol{k}\in\mathcal{Q}^c}\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\frac{\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)^2}{B(\boldsymbol{d},\boldsymbol{k})},
\end{aligned}
$$

where we used in the third line

$$
\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)Y^{\boldsymbol{d}}_{\boldsymbol{l},\boldsymbol{s}}(\overline{\boldsymbol{x}})Y^{\boldsymbol{d}}_{\boldsymbol{l}',\boldsymbol{s}'}(\overline{\boldsymbol{y}})\right]\\
&\quad=\mathbb{E}_{\overline{\boldsymbol{y}}}\left[\mathbb{E}_{\overline{\boldsymbol{x}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)Y^{\boldsymbol{d}}_{\boldsymbol{l},\boldsymbol{s}}(\overline{\boldsymbol{x}})\right]Y^{\boldsymbol{d}}_{\boldsymbol{l}',\boldsymbol{s}'}(\overline{\boldsymbol{y}})\right]\\
&\quad=\frac{\delta_{\boldsymbol{k},\boldsymbol{l}}}{B(\boldsymbol{d},\boldsymbol{k})}\mathbb{E}_{\overline{\boldsymbol{y}}}\left[Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{\boldsymbol{y}})Y^{\boldsymbol{d}}_{\boldsymbol{l}',\boldsymbol{s}'}(\overline{\boldsymbol{y}})\right]\\
&\quad=\frac{\delta_{\boldsymbol{k},\boldsymbol{l}}\delta_{\boldsymbol{k},\boldsymbol{l}'}\delta_{\boldsymbol{s},\boldsymbol{s}'}}{B(\boldsymbol{d},\boldsymbol{k})}.
\end{aligned}
$$

We conclude that

$$
\begin{aligned}
\mathcal{E}^{(q)}_{\mathcal{Q}^c,\varepsilon_0}&=\sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}}]\sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}\frac{\lambda^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}(\overline{f}_d)^2}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k}\in\mathcal{Q}^c}\\
&=\sum_{\boldsymbol{k}\in\mathcal{Q}^c}\frac{\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}}]}{B(\boldsymbol{d},\boldsymbol{k})}\|P_{\boldsymbol{k}}f_d\|^2_{L^2}\\
&\leqslant\left[\max_{\boldsymbol{k}\in\mathcal{Q}^c}B(\boldsymbol{d},\boldsymbol{k})^{-1}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}}]\right]\cdot\|\mathsf{P}_{\mathcal{Q}^c}f_d\|^2_{L^2}.
\end{aligned}
$$

□

**Lemma 12.** *Let $\sigma$ be a weakly differentiable activation function with weak derivative $\sigma'$. For a fixed $\boldsymbol{\tau}\in\mathbb{R}^Q_{\geqslant 0}$, define the kernels for $q\in[Q]$,*

$$
H^{(q)}_{\boldsymbol{\tau}}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}})=\frac{r_q^2}{d_q}\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right]\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle.
$$

*Then, we have the following decomposition in terms of product of Gegenbauer polynomials*

$$
H^{(q)}_{\boldsymbol{\tau}}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}})=\sum_{\boldsymbol{k}\in\mathbb{Z}^2_{\geqslant 0}}A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}}Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{x}}^{(q)},\overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right),
$$

*where*

$$A^{(q)}_{\boldsymbol{\tau},\boldsymbol{k}} = r_q^2 \cdot [t_{d_q,k_q-1}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k}_{q-}) + s_{d_q,k_q+1}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k}_{q+})],$$

*with* $\mathbf{k}_{q+} = (k_1, \dots, k_q + 1, \dots, k_Q)$ *and* $\mathbf{k}_{q-} = (k_1, \dots, k_q - 1, \dots, k_Q)$, *and*

$$s_{d,k} = \frac{k}{2k + d - 2}, \qquad t_{d,k} = \frac{k + d - 2}{2k + d - 2},$$

*with the convention* $t_{d,-1} = 0$.

**Proof of lemma 12.** Recall the decomposition of $\sigma'$ in terms of tensor product of Gegenbauer polynomials,

$$\sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle / R) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right),$$

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}^{(1)}_1, \dots, \overline{x}^{(Q)}_1\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\,\overline{x}^{(1)}_1, \dots, \sqrt{d_q}\,\overline{x}^{(Q)}_1\right)\right].$$

Injecting this decomposition into the definition of $H^{(q)}_{\boldsymbol{\tau}}$ yields

$$\begin{aligned} H^{(q)}_{\boldsymbol{\tau}}(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}) &= \frac{r_q^2}{d_q}\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle / \sqrt{d_q}\}_{q\in[Q]}\right)\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle / \sqrt{d_q}\}_{q\in[Q]}\right)\right]\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle \\ &= \frac{r_q^2}{d_q}\sum_{\boldsymbol{k},\boldsymbol{k}'\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})\lambda^{\boldsymbol{d}}_{\boldsymbol{k}'}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) B(\boldsymbol{d},\boldsymbol{k}) B(\boldsymbol{d},\boldsymbol{k}') \\ &\quad\times \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)\right]\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle. \end{aligned}$$

Recalling equation (36), we have

$$\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}'}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)\right] = \delta_{\boldsymbol{k},\boldsymbol{k}'}\frac{Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)}{B(\boldsymbol{d},\boldsymbol{k})}.$$

Hence,

$$\begin{aligned} H^{(q)}_{\boldsymbol{\tau}}(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}) &= \frac{r_q^2}{d_q}\sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle\}_{q\in[Q]}\right)\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle \\ &= r_q^2 \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d},\boldsymbol{k})\left[Q^{(d_q)}_{k_q}(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle)\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle / d_q\right] \\ &\quad\times \prod_{q'\neq q} Q^{\boldsymbol{d}}_{k_{q'}}(\langle \overline{\boldsymbol{x}}^{(q')}, \overline{\boldsymbol{y}}^{(q')}\rangle). \end{aligned}$$

By the recurrence relationship for Gegenbauer polynomials (23), we have

$$\frac{t}{d_q} Q_{k_q}^{(d_q)}(t) = s_{d_q,k_q} Q_{k_q-1}^{(d_q)}(t) + t_{d_q,k_q} Q_{k_q+1}^{(d_q)}(t),$$

where (we use the convention $t_{d_q,-1} = 0$)

$$s_{d_q,k_q} = \frac{k_q}{2k_q + d_q - 2}, \qquad t_{d_q,k_q} = \frac{k_q + d_q - 2}{2k_q + d_q - 2}.$$

Hence we get,

$$\begin{aligned} H_{\boldsymbol{\tau}}^{(q)}(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}) &= r_q^2 \sum_{\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}} (\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d}, \boldsymbol{k}) \left[ Q_{k_q}^{(d_q)}(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)} \rangle) \langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)} \rangle / d_q \right] \\ &\quad \times \prod_{q' \neq q} Q_{k_{q'}}^{\boldsymbol{d}}(\langle \overline{\boldsymbol{x}}^{(q')}, \overline{\boldsymbol{y}}^{(q')} \rangle) \\ &= r_q^2 \sum_{\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}} (\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d}, \boldsymbol{k}) \left[ s_{d_q,k_q} Q_{k_q-1}^{(d_q)}(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)} \rangle) + t_{d_q,k_q} Q_{k_q+1}^{(d_q)}(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)} \rangle) \right] \\ &\quad \times \prod_{q' \neq q} Q_{k_{q'}}^{\boldsymbol{d}}(\langle \overline{\boldsymbol{x}}^{(q')}, \overline{\boldsymbol{y}}^{(q')} \rangle) \\ &= \sum_{\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^2} A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)} Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ \langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)} \rangle \}_{q \in [Q]} \right), \end{aligned}$$

where we get by matching the coefficients,

$$A_{\boldsymbol{\tau},\boldsymbol{k}}^{(q)} = r_q^2 \cdot [t_{d_q,k_q-1} \lambda_{\boldsymbol{k}_{q-}}^{\boldsymbol{d}} (\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d}, \boldsymbol{k}_{q-}) + s_{d_q,k_q+1} \lambda_{\boldsymbol{k}_{q+}}^{\boldsymbol{d}} (\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})^2 B(\boldsymbol{d}, \boldsymbol{k}_{q+})],$$

with $\boldsymbol{k}_{q+} = (k_1, \ldots, k_q + 1, \ldots, k_Q)$ and $\boldsymbol{k}_{q-} = (k_1, \ldots, k_q - 1, \ldots, k_Q)$. □

## G.4. Proof of proposition 4

### *G.4.1. Preliminaries.*

**Lemma 13.** *Let $\psi : \mathbb{R}^Q \to \mathbb{R}$ be a function such that $\psi(\{\langle \boldsymbol{e}_q, \cdot \rangle\}_{q \in [Q]}) \in L^2(\mathrm{PS}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}})$. We will consider for integers $\boldsymbol{i} = (i_1, \ldots, i_Q) \in \mathbb{Z}_{\geqslant 0}^Q$, the associated function $\psi^{(i)}$ given by:*

$$\psi^{(\boldsymbol{i})}\left(\overline{x}_1^{(1)}, \ldots, \overline{x}_1^{(Q)}\right) = \left(\overline{x}_1^{(1)}\right)^{i_1} \ldots \left(\overline{x}_1^{(Q)}\right)^{i_Q} \psi\left(\overline{x}_1^{(1)}, \ldots, \overline{x}_1^{(Q)}\right).$$

*Assume that* $\psi^{(\boldsymbol{i})}(\{\langle \boldsymbol{e}_q, \cdot\rangle\}_{q\in[Q]}) \in L^2(\mathrm{PS}^{\boldsymbol{d}}, \mu_{\boldsymbol{d}})$*. Let* $\{\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\psi)\}_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}}$ *be the coefficients of the expansion of* $\psi$ *in terms of the product of Gegenbauer polynomials*

$$\psi\left(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)}\right) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\psi) B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\ldots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right),$$

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\psi) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\psi\left(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)}\right) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\ldots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right].$$

*Then we can write*

$$\psi^{(\boldsymbol{i})}\left(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)}\right) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}} \lambda^{\boldsymbol{d},\boldsymbol{i}}_{\boldsymbol{k}}(\psi) B(\boldsymbol{d},\boldsymbol{k}) Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\ldots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right),$$

*where the coefficients* $\lambda^{\boldsymbol{d},\boldsymbol{i}}_{\boldsymbol{k}}(\psi)$ *are given recursively: denoting* $i_{q+} = (i_1,\ldots,i_q+1,\ldots,i_Q)$*, if* $k_q = 0$,

$$\lambda^{\boldsymbol{d},\boldsymbol{i}_{q+}}_{\boldsymbol{k}}(\psi) = \sqrt{d_q}\lambda^{\boldsymbol{d},\boldsymbol{i}}_{\boldsymbol{k}_{q+}}(\psi),$$

*and for* $k_q > 0$,

$$\lambda^{\boldsymbol{d},\boldsymbol{i}_{q+}}_{\boldsymbol{k}}(\psi) = \sqrt{d_q}\frac{k_q+d_q-2}{2k_q+d_q-2}\lambda^{\boldsymbol{d},\boldsymbol{i}}_{\boldsymbol{k}_{q+}}(\psi) + \sqrt{d_q}\frac{k_q}{2k_q+d_q-2}\lambda^{\boldsymbol{d},\boldsymbol{i}}_{\boldsymbol{k}_{q-}}(\psi),$$

*where we recall the notations* $\mathbf{k}_{q+} = (k_1,\ldots,k_q+1,\ldots,k_Q)$ *and* $\mathbf{k}_{q-} = (k_1,\ldots,k_q-1,\ldots,k_Q)$.

**Proof of lemma 13.** We recall the following two formulas for $k \geqslant 1$ (see appendix B.2):

$$\frac{x}{d}Q^{(d)}_k(x) = \frac{k}{2k+d-2}Q^{(d)}_{k-1}(x) + \frac{k+d-2}{2k+d-2}Q^{(d)}_{k+1}(x),$$

$$B(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{k-1}.$$

Furthermore, we have $Q^{(d)}_0(x) = 1$, $Q^{(d)}_1(x) = x/d$ and therefore therefore $xQ^{(d)}_0(x) = dQ^{(d)}_1(x)$. Similarly to the proof of lemma 6 in [21], we insert these expressions in the expansion of the function $\psi$. Matching the coefficients of the expansion yields the result. □

Let $\boldsymbol{u}\colon \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}^{D\times D}$ be a matrix-valued function defined by

$$\boldsymbol{u}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle\boldsymbol{\theta}_1,\boldsymbol{x}\rangle/R)\sigma'(\langle\boldsymbol{\theta}_2,\boldsymbol{x}\rangle/R)\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}].$$

We can write this function as a $Q$ by $Q$ block matrix function $\boldsymbol{u} = (\boldsymbol{u}^{(qq')})_{q,q'\in[Q]}$, where $\boldsymbol{u}^{(qq')}: \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}^{d_q\times d_{q'}}$ are given by

$$\boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle\boldsymbol{\theta}_1,\boldsymbol{x}\rangle/R)\sigma'(\langle\boldsymbol{\theta}_2,\boldsymbol{x}\rangle/R)\boldsymbol{x}^{(q)}(\boldsymbol{x}^{(q')})^{\mathsf{T}}].$$

We have the following lemma which is a generalization of lemma 7 in [21] that shows essentially the same decomposition of the matrix $\boldsymbol{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as by integration by part if we had $\boldsymbol{x} \sim N(0, \mathbf{I})$.

**Lemma 14.** *For $q \in [Q]$, there exists functions $u_1^{(qq)}, u_2^{(qq)}, u_{3,1}^{(qq)}, u_{3,2}^{(qq)} : \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}$ such that*

$$\begin{aligned}\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = u_1^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\mathbf{I}_{d_q} + u_2^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)[\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_2^{(q)})^\mathsf{T} + \boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_1^{(q)})^\mathsf{T}] \\ + u_{3,1}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q)})^\mathsf{T} + u_{3,2}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_2^{(q)})^\mathsf{T}.\end{aligned}$$

*For $q, q' \in [Q]$, there exists functions $u_{2,1}^{(qq')}, u_{2,2}^{(qq')}, u_{3,1}^{(qq')}, u_{3,2}^{(qq')} : \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}$ such that*

$$\begin{aligned}\boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = u_{2,1}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_2^{(q')})^\mathsf{T} + u_{2,2}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_1^{(q')})^\mathsf{T} \\ + u_{3,1}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q')})^\mathsf{T} + u_{3,2}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_2^{(q')})^\mathsf{T}.\end{aligned}$$

**Proof of lemma 14.** Denote $\gamma^{(q)} = \langle \overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{\theta}}_2^{(q)} \rangle / d_q$. Let us rotate each sphere $q \in [Q]$ such that

$$\begin{aligned}\boldsymbol{\theta}_1^{(q)} &= \left(\tau_1^{(q)}\sqrt{d_q}, 0, \ldots, 0\right), \\ \boldsymbol{\theta}_2^{(q)} &= \left(\tau_2^{(q)}\sqrt{d_q}\gamma^{(q)}, \tau_2^{(q)}\sqrt{d_q}\sqrt{1-(\gamma^{(q)})^2}, 0, \ldots, 0\right).\end{aligned} \tag{113}$$

**Step 1: $\boldsymbol{u}^{(qq)}$.**
Let us start with $\boldsymbol{u}^{(qq)}$. For clarity, we will denote (in the rotated basis (113))

$$\begin{aligned}\alpha_1 &= \langle \boldsymbol{\theta}_1, \boldsymbol{x} \rangle / R = \sum_{q \in [Q]} \tau_1^{(q)}\sqrt{d_q}/R \cdot \overline{x}_1^{(q)}, \\ \alpha_2 &= \langle \boldsymbol{\theta}_2, \boldsymbol{x} \rangle / R = \sum_{q \in [Q]} \left[\tau_2^{(q)}\sqrt{d_q}\gamma^{(q)}/R \cdot \overline{x}_1^{(q)} + \tau_2^{(q)}\sqrt{d_q}\sqrt{1-(\gamma^{(q)})^2}/R \cdot \overline{x}_2^{(q)}\right].\end{aligned}$$

Then it is easy to show that we can rewrite

$$\begin{aligned}\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)\boldsymbol{x}^{(q)}(\boldsymbol{x}^{(q)})^\mathsf{T} \\ &= \begin{bmatrix} \boldsymbol{u}_{1:2,1:2}^{(qq)} & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)(x_3^{(q)})^2]\mathbf{I}_{d_q-2} \end{bmatrix},\end{aligned}$$

with

$$\boldsymbol{u}_{1:2,1:2}^{(qq)} = \begin{bmatrix} \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)(x_1^{(q)})^2] & \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)x_1^{(q)}x_2^{(q)}] \\ \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)x_2^{(q)}x_1^{(q)}] & \mathbb{E}_{\boldsymbol{x}}[\sigma'(\alpha_1)\sigma'(\alpha_2)(x_2^{(q)})^2] \end{bmatrix}.$$

https://doi.org/10.1088/1742-5468/ac3a81

**Case (a): $\boldsymbol{\theta}_1^{(q)} \neq \boldsymbol{\theta}_2^{(q)}$.**
Given any functions $u_1^{(qq)}, u_2^{(qq)}, u_{3,1}^{(qq)}, u_{3,2}^{(qq)} : \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}$, we define

$$\begin{aligned}\tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = {} & u_1^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\mathbf{I}_{d_1} + u_2^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)[\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_2^{(q)})^{\mathsf{T}} + \boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_1^{(q)})^{\mathsf{T}}] \\ & + u_{3,1}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q)})^{\mathsf{T}} + u_{3,2}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_2^{(q)})^{\mathsf{T}}.\end{aligned}$$

In the rotated basis (113), we have

$$\tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \begin{bmatrix} \tilde{\boldsymbol{u}}_{1:2,1:2}^{(qq)} & \mathbf{0} \\ \mathbf{0} & u_1^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\mathbf{I}_{d_q-2} \end{bmatrix},$$

where (we dropped the dependency on $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for clarity)

$$\begin{aligned}\boldsymbol{u}_{11}^{(qq)} &= u_1^{(qq)} + 2\tau_1^{(q)}\tau_2^{(q)}d_q\gamma^{(q)}u_2^{(qq)} + (\tau_1^{(q)})^2 d_q u_{3,1}^{(qq)} + (\tau_2^{(q)})^2 d_q(\gamma^{(q)})^2 u_{3,2}^{(qq)}, \\ \boldsymbol{u}_{12}^{(qq)} &= \tau_1^{(q)}\tau_2^{(q)}d_q\sqrt{1-(\gamma^{(q)})^2}u_2^{(qq)} + (\tau_2^{(q)})^2 d_q\gamma^{(q)}\sqrt{1-(\gamma^{(q)})^2}u_{3,2}^{(qq)}, \\ \boldsymbol{u}_{22}^{(qq)} &= u_1^{(qq)} + (\tau_2^{(q)})^2 d_q(1-(\gamma^{(q)})^2)u_{3,2}^{(qq)}.\end{aligned}$$

We see that $\boldsymbol{u}^{(qq)}$ and $\tilde{\boldsymbol{u}}^{(qq)}$ will be equal if and only if we have the following equalities:

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) &= \mathrm{Tr}(\tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \\ &= d_q u_1^{(qq)} + 2\tau_1^{(q)}\tau_2^{(q)}d_q\gamma^{(q)}u_2^{(qq)} + (\tau_1^{(q)})^2 d_q u_{3,1}^{(qq)} + (\tau_2^{(q)})^2 d_q u_{3,2}^{(qq)}, \\ \langle\boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}\rangle &= \langle\boldsymbol{\theta}_1^{(q)}, \tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}\rangle \\ &= \tau_1^{(q)}\tau_2^{(q)}d_q\gamma^{(q)}u_1^{(qq)} + (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q^2(1+(\gamma^{(q)})^2)u_2^{(qq)} \\ &\quad + (\tau_1^{(q)})^3\tau_2^{(q)}d_q^2\gamma^{(q)}u_{3,1}^{(qq)} + \tau_1^{(q)}(\tau_2^{(q)})^3 d_q^2\gamma^{(q)}u_{3,1}^{(qq)}, \\ \langle\boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}\rangle &= \langle\boldsymbol{\theta}_1^{(q)}, \tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}\rangle \\ &= (\tau_1^{(q)})^2 d_q u_1^{(qq)} + 2(\tau_1^{(q)})^3\tau_2^{(q)}d_q^2\gamma^{(q)}u_2^{(qq)} \\ &\quad + (\tau_1^{(q)})^4 d_q^2 u_{3,1}^{(qq)} + (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q^2(\gamma^{(q)})^2 u_{3,2}^{(qq)}, \\ \langle\boldsymbol{\theta}_2^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}\rangle &= \langle\boldsymbol{\theta}_2^{(q)}, \tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}\rangle \\ &= (\tau_2^{(q)})^2 d_q u_1^{(qq)} + 2\tau_1^{(q)}(\tau_2^{(q)})^3 d_q^2\gamma^{(q)}u_2^{(qq)} \\ &\quad + (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q^2(\gamma^{(q)})^2 u_{3,1}^{(qq)} + (\tau_2^{(q)})^4 d_q^2 u_{3,2}^{(qq)}.\end{aligned}$$

Hence $\tilde{\boldsymbol{u}}^{(qq)} = \boldsymbol{u}^{(qq)}$ if and only if

$$\begin{bmatrix} u_1^{(qq)} \\ u_2^{(qq)} \\ u_{3,1}^{(qq)} \\ u_{3,2}^{(qq)} \end{bmatrix} = d_q^{-1}(\boldsymbol{M}^{(qq)})^{-1} \times \begin{bmatrix} \mathrm{Tr}(\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \\ \langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)} \rangle \\ \langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)} \rangle \\ \langle \boldsymbol{\theta}_2^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)} \rangle \end{bmatrix}, \tag{114}$$

where

$$\boldsymbol{M}^{(qq)} = \begin{bmatrix} 1 & 2\tau_1^{(q)}\tau_2^{(q)}\gamma^{(q)} & (\tau_1^{(q)})^2 & (\tau_2^{(q)})^2 \\ \tau_1^{(q)}\tau_2^{(q)}\gamma^{(q)} & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(1+(\gamma^{(q)})^2) & (\tau_1^{(q)})^3\tau_2^{(q)} d_q\gamma^{(q)} & \tau_1^{(q)}(\tau_2^{(q)})^3 d_q\gamma^{(q)} \\ (\tau_1^{(q)})^2 & 2(\tau_1^{(q)})^3\tau_2^{(q)} d_q\gamma^{(q)} & (\tau_1^{(q)})^4 d_q & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(\gamma^{(q)})^2 \\ (\tau_2^{(q)})^2 & 2\tau_1^{(q)}(\tau_2^{(q)})^3 d_q\gamma^{(q)} & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(\gamma^{(q)})^2 & (\tau_2^{(q)})^4 d_q \end{bmatrix}$$

is invertible almost surely (for $\tau_1^{(q)}, \tau_2^{(q)} \neq 0$ and $\gamma^{(q)} \neq 1$).

**Case (b):** $\boldsymbol{\theta}_1^{(q)} = \boldsymbol{\theta}_2^{(q)}$.

Similarly, for some fixed $\alpha$ and $\beta$, we define

$$\tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) = \alpha \mathbf{I}_{d_q} + \beta \boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q)})^{\mathsf{T}}.$$

Then $\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)$ and $\tilde{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)$ are equal if and only if

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = d_q^{-1}(\boldsymbol{M}_{\parallel}^{(qq)})^{-1} \times \begin{bmatrix} \mathrm{Tr}(\boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)) \\ \langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)\boldsymbol{\theta}_1^{(q)} \rangle \end{bmatrix},$$

where

$$\boldsymbol{M}_{\parallel}^{(qq)} = \begin{bmatrix} 1 & (\tau_1^{(q)})^2 \\ (\tau_1^{(q)})^2 & (\tau_1^{(q)})^4 d_q \end{bmatrix}.$$

**Step 2:** $\boldsymbol{u}^{(qq')}$ for $q \neq q'$.

Similarly to the two previous steps, we define for any functions $u_{2,1}^{(qq')}, u_{2,2}^{(qq')}, u_{3,1}^{(qq')}, u_{3,2}^{(qq')} : \mathbb{S}^{D-1}(\sqrt{D}) \times \mathbb{S}^{D-1}(\sqrt{D}) \to \mathbb{R}$,

$$\begin{aligned} \tilde{\boldsymbol{u}}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = {} & u_{2,1}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_2^{(q')})^{\mathsf{T}} + u_{2,2}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_1^{(q')})^{\mathsf{T}} \\ & + u_{3,1}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q')})^{\mathsf{T}} + u_{3,2}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)}(\boldsymbol{\theta}_2^{(q')})^{\mathsf{T}}. \end{aligned}$$

We can rewrite $\tilde{\boldsymbol{u}}^{(qq')}$ as

$$\tilde{\boldsymbol{u}}^{(qq')}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \begin{bmatrix} \tilde{\boldsymbol{u}}_{1:2,1:2}^{(qq')} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where

$$
\begin{aligned}
\tilde{\boldsymbol{u}}_{11}^{(qq')} &= u_{2,1}^{(qq')}\tau_1^{(q)}\tau_2^{(q')}\gamma^{(q')} + u_{2,2}^{(qq')}\tau_2^{(q)}\tau_1^{(q')}\gamma^{(q)} + u_{3,1}^{(qq')}\tau_1^{(q)}\tau_1^{(q')} + u_{3,2}^{(qq')}\tau_2^{(q)}\tau_2^{(q')}\gamma^{(q)}\gamma^{(q')},\\
\tilde{\boldsymbol{u}}_{12}^{(qq')} &= u_{2,1}^{(qq')}\tau_1^{(q)}\tau_2^{(q')}\sqrt{1-(\gamma^{(q')})^2} + u_{3,2}^{(qq')}\tau_2^{(q)}\tau_2^{(q')}\gamma^{(q)}\sqrt{1-(\gamma^{(q')})^2},\\
\tilde{\boldsymbol{u}}_{21}^{(qq')} &= u_{2,2}^{(qq')}\tau_2^{(q)}\tau_1^{(q')}\sqrt{1-(\gamma^{(q)})^2} + u_{3,2}^{(qq')}\tau_2^{(q)}\tau_2^{(q')}\sqrt{1-(\gamma^{(q)})^2}\gamma^{(q')},\\
\tilde{\boldsymbol{u}}_{22}^{(qq')} &= u_{3,2}^{(qq')}\tau_2^{(q)}\tau_2^{(q')}\sqrt{1-(\gamma^{(q)})^2}\sqrt{1-(\gamma^{(q')})^2}.
\end{aligned}
$$

**Case (a): $\boldsymbol{\theta}_1^{(q)} \neq \boldsymbol{\theta}_2^{(q)}$.**
We have equality $\tilde{\boldsymbol{u}}^{(qq')} = \boldsymbol{u}^{(qq')}$ if and only if

$$
\begin{bmatrix} u_{2,1}^{(qq')} \\ u_{2,2}^{(qq')} \\ u_{3,1}^{(qq')} \\ u_{3,2}^{(qq')} \end{bmatrix} = (d_q d_{q'})^{-1}(\boldsymbol{M}^{(qq')})^{-1} \times \begin{bmatrix} \langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q')}\rangle \\ \langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q')}\rangle \\ \langle \boldsymbol{\theta}_2^{(q)}, \boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q')}\rangle \\ \langle \boldsymbol{\theta}_2^{(q)}, \boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q')}\rangle \end{bmatrix},
$$

where $\boldsymbol{M}^{(qq')}$ is given by

$$
\begin{bmatrix}
(\tau_1^{(q)})^2\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q')} & \tau_1^{(q)}\tau_2^{(q)}(\tau_1^{(q')})^2\gamma^{(q)} & (\tau_1^{(q)})^2(\tau_1^{(q')})^2 & \tau_1^{(q)}\tau_2^{(q)}\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q)}\gamma^{(q')} \\
(\tau_1^{(q)})^2(\tau_2^{(q')})^2 & \tau_1^{(q)}\tau_2^{(q)}\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q)}\gamma^{(q')} & (\tau_1^{(q)})^2\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q')} & \tau_1^{(q)}\tau_2^{(q)}(\tau_2^{(q')})^2\gamma^{(q)} \\
\tau_1^{(q)}\tau_2^{(q)}\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q)}\gamma^{(q')} & (\tau_2^{(q)})^2(\tau_1^{(q')})^2 & \tau_1^{(q)}\tau_2^{(q)}(\tau_1^{(q')})^2\gamma^{(q)} & (\tau_2^{(q)})^2\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q')} \\
\tau_1^{(q)}\tau_2^{(q)}(\tau_2^{(q')})^2\gamma^{(q)} & (\tau_2^{(q)})^2\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q')} & \tau_1^{(q)}\tau_2^{(q)}\tau_1^{(q')}\tau_2^{(q')}\gamma^{(q)}\gamma^{(q')} & (\tau_2^{(q)})^2(\tau_2^{(q')})^2
\end{bmatrix},
$$

which is invertible almost surely (for $\tau_1^{(q)}, \tau_2^{(q)} \neq 0$ and $\gamma^{(q)} \neq 1$).

**Case (b): $\boldsymbol{\theta}_1^{(q)} = \boldsymbol{\theta}_2^{(q)}$.**
It is straightforward to check that

$$
\boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) = \beta\boldsymbol{\theta}_1^{(q)}(\boldsymbol{\theta}_1^{(q')})^{\mathsf{T}},
$$

where

$$
\beta = (d_q d_{q'})^{-1}(\tau_1^{(q)}\tau_1^{(q')})^{-2}\left\langle \boldsymbol{\theta}_1^{(q)}, \boldsymbol{u}^{(qq')}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)\boldsymbol{\theta}_1^{(q')}\right\rangle.
$$

□

*G.4.2. Proof of proposition 4.* **Step 1. Construction of the activation function $\hat{\sigma}$.**

Recall the definition of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in equation (100) and its expansion in terms of tensor product of Gegenbauer polynomials:

$$
\sigma'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R) = \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right),
$$

$$
\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)}\right)Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\ldots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right].
$$

We recall the definition of $q_\xi = \mathrm{argmax}_{q\in[Q]}\{\eta_q + \kappa_q\}$. Let $l_2 > l_1 \geqslant 2L+5$ be two indices that satisfy the conditions of assumption 3(b) and we define $\boldsymbol{l}_1 = (0,\dots,0,l_1,0,\dots,0)$ ($l_1$ at position $q_\xi$) and $\boldsymbol{l}_2 = (0,\dots,0,l_2,0,\dots,0)$ ($l_2$ at position $q_\xi$). Using the Gegenbauer coefficients of $\sigma'$, we define a new activation function $\hat{\sigma}'$ by

$$\hat{\sigma}'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R) = \sum_{\boldsymbol{k}\in\mathbb{Z}^Q_{\geqslant 0}\setminus\{\boldsymbol{l}_1,\boldsymbol{l}_2\}} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q^{\boldsymbol{d}}_{\boldsymbol{k}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) \tag{115}$$

$$+\sum_{t=1,2}(1-\delta_t)\lambda^{\boldsymbol{d}}_{\boldsymbol{l}_t}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})B(d_{q_\xi},l_t)Q^{(d_{q_\xi})}_{l_t}(\langle\overline{\boldsymbol{\theta}}^{(q_\xi)},\overline{\boldsymbol{x}}^{(q_\xi)}\rangle)), \tag{116}$$

for some $\delta_1, \delta_2$ that we will fix later (with $|\delta_t| \leqslant 1$).

**Step 2. The functions $\boldsymbol{u}$, $\hat{\boldsymbol{u}}$ and $\bar{\boldsymbol{u}}$.**

Let $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$ be the matrix-valued functions associated respectively to $\sigma'$ and $\hat{\sigma}'$

$$\boldsymbol{u}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle\boldsymbol{\theta}_1,\boldsymbol{x}\rangle/R)\sigma'(\langle\boldsymbol{\theta}_2,\boldsymbol{x}\rangle/R)\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}], \tag{117}$$

$$\hat{\boldsymbol{u}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}}[\hat{\sigma}'(\langle\boldsymbol{\theta}_1,\boldsymbol{x}\rangle/R)\hat{\sigma}'(\langle\boldsymbol{\theta}_2,\boldsymbol{x}\rangle/R)\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}]. \tag{118}$$

From lemma 14, there exists functions $u_1^{ab}, u_{2,1}^{ab}, u_{2,2}^{ab}, u_{3,1}^{ab}, u_{3,2}^{ab}$ and $\hat{u}_1^{ab}, \hat{u}_{2,1}^{ab}, \hat{u}_{2,2}^{ab}, \hat{u}_{3,1}^{ab}, \hat{u}_{3,2}^{ab}$ (for $a, b \in [Q]$), which decompose $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$ along $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ vectors. We define $\bar{\boldsymbol{u}} = \boldsymbol{u} - \hat{\boldsymbol{u}}$. Then we have the same decomposition for $\bar{u}_{k,j}^{ab} = u_{k,j}^{ab} - \hat{u}_{k,j}^{ab}$ for $a, b \in [Q]$, $k = 1, 2, 3$, $j = 1, 2$.

**Step 3. Construction of the kernel matrices.**

Let $\boldsymbol{U}$, $\hat{\boldsymbol{U}}$, $\bar{\boldsymbol{U}} \in \mathbb{R}^{ND\times ND}$ with $i, j$th block (for $i, j \in [N]$) given by

$$\boldsymbol{U}_{ij} = \boldsymbol{u}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j), \tag{119}$$

$$\hat{\boldsymbol{U}}_{ij} = \hat{\boldsymbol{u}}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j), \tag{120}$$

$$\bar{\boldsymbol{U}}_{ij} = \bar{\boldsymbol{u}}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) = \boldsymbol{u}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - \hat{\boldsymbol{u}}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j). \tag{121}$$

Note that we have $\boldsymbol{U} = \hat{\boldsymbol{U}} + \bar{\boldsymbol{U}}$. By equations (118) and (120), it is easy to see that $\hat{\boldsymbol{U}} \succeq 0$. Then we have $\boldsymbol{U} \succeq \bar{\boldsymbol{U}}$. In the following, we would like to lower bound matrix $\bar{\boldsymbol{U}}$.

We decompose $\bar{\boldsymbol{U}}$ as

$$\bar{\boldsymbol{U}} = \boldsymbol{D} + \boldsymbol{\Delta},$$

where $\boldsymbol{D} \in \mathbb{R}^{DN\times DN}$ is a block-diagonal matrix, with

$$\boldsymbol{D} = \mathrm{diag}(\bar{\boldsymbol{U}}_{11},\dots,\bar{\boldsymbol{U}}_{NN}), \tag{122}$$

and $\boldsymbol{\Delta} \in \mathbb{R}^{DN\times DN}$ is formed by blocks $\boldsymbol{\Delta}_{ij} \in \mathbb{R}^{D\times D}$ for $i, j \in [n]$, defined by

$$\boldsymbol{\Delta}_{ij} = \begin{cases} 0, & i = j, \\ \bar{\boldsymbol{U}}_{ij}, & i \neq j. \end{cases} \tag{123}$$

In the rest of the proof, we will prove that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(d^{-\max_{q\in[Q]}\kappa_q})$ and the block matrix $\boldsymbol{D}$ verifies the properties (105) and (106).

**Step 4. Prove that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(d^{-\max_{q\in[Q]}\kappa_q})$.**

We will prove in fact that $\|\boldsymbol{\Delta}\|_F^2 = o_{d,\mathbb{P}}(d^{-2\max_{q\in[Q]}\kappa_q})$. For the rest of the proof, we fix $\varepsilon_0 \in (0,1)$ and we restrict ourselves without loss of generality to the set $\mathcal{P}_{\varepsilon_0}$.

Let us start with $\overline{\boldsymbol{u}}^{(qq)}$ for $q \in [Q]$. Denoting $\gamma_{ij}^{(q)} = \langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)} \rangle / d_q < 1$, we get, from equation (114),

$$\begin{bmatrix} \bar{u}_1^{(qq)}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ \bar{u}_2^{(qq)}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ \bar{u}_{3,1}^{(qq)}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ \bar{u}_{3,2}^{(qq)}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \end{bmatrix} = \begin{bmatrix} u_1(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - \hat{u}_1(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ u_2(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - \hat{u}_2(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ u_{3,1}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - \hat{u}_{3,1}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \\ u_{3,2}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - \hat{u}_{3,2}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) \end{bmatrix}$$

$$= d_q^{-1}(\boldsymbol{M}_{ij}^{(qq)})^{-1} \times \begin{bmatrix} \mathrm{Tr}(\bar{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)) \\ \langle \boldsymbol{\theta}_1^{(q)}, \bar{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)} \rangle \\ \langle \boldsymbol{\theta}_1^{(q)}, \bar{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_1^{(q)} \rangle \\ \langle \boldsymbol{\theta}_2^{(q)}, \bar{\boldsymbol{u}}^{(qq)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{\theta}_2^{(q)} \rangle \end{bmatrix}, \tag{124}$$

where $\boldsymbol{M}_{ij}^{(qq)}$ is given by

$$\begin{bmatrix} 1 & 2\tau_1^{(q)}\tau_2^{(q)}\gamma_{ij}^{(q)} & (\tau_1^{(q)})^2 & (\tau_2^{(q)})^2 \\ \tau_1^{(q)}\tau_2^{(q)}\gamma_{ij}^{(q)} & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(1+(\gamma_{ij}^{(q)})^2) & (\tau_1^{(q)})^3\tau_2^{(q)} d_q\gamma_{ij}^{(q)} & \tau_1^{(q)}(\tau_2^{(q)})^3 d_q\gamma_{ij}^{(q)} \\ (\tau_1^{(q)})^2 & 2(\tau_1^{(q)})^3\tau_2^{(q)} d_q\gamma^{(q)} & (\tau_1^{(q)})^4 d_q & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(\gamma_{ij}^{(q)})^2 \\ (\tau_2^{(q)})^2 & \tau_1^{(q)}(\tau_2^{(q)})^3 d_q\gamma_{ij}^{(q)} & (\tau_1^{(q)})^2(\tau_2^{(q)})^2 d_q(\gamma^{(q)})^2 & (\tau_2^{(q)})^4 d_q \end{bmatrix}. \tag{125}$$

Using the notations of lemma 13, we get

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{U}_{ij}^{(qq)}) &= \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)\|\boldsymbol{x}^{(q)}\|_2^2] \\
&= r_q^2 \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle\}_{q\in[Q]}\right), \\
\langle \boldsymbol{\theta}_i^{(q)}, \boldsymbol{U}_{ij}^{(qq)}\boldsymbol{\theta}_j^{(q)}\rangle &= \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\langle \boldsymbol{\theta}_i^{(q)}, \boldsymbol{z}^{(q)}\rangle\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)\langle \boldsymbol{\theta}_j^{(q)}, \boldsymbol{x}^{(q)}\rangle] \\
&= r_q^2\tau_i^{(q)}\tau_j^{(q)} \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\boldsymbol{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i})\lambda_{\boldsymbol{k}}^{\boldsymbol{d},\boldsymbol{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle\}_{q\in[Q]}\right), \\
\langle \boldsymbol{\theta}_i^{(q)}, \boldsymbol{U}_{ij}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle &= \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\langle \boldsymbol{\theta}_i^{(q)}, \boldsymbol{x}^{(q)}\rangle^2\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)] \\
&= r_q^2(\tau_i^{(q)})^2 \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\boldsymbol{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i})\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle\}_{q\in[Q]}\right), \\
\langle \boldsymbol{\theta}_j^{(q)}, \boldsymbol{U}_{ij}^{(qq)}\boldsymbol{\theta}_j^{(q)}\rangle &= \mathbb{E}_{\boldsymbol{x}}[\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)\langle \boldsymbol{\theta}_j^{(q)}, \boldsymbol{x}^{(q)}\rangle^2] \\
&= r_q^2(\tau_j^{(q)})^2 \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i})\lambda_{\boldsymbol{k}}^{\boldsymbol{d},\boldsymbol{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle\}_{q\in[Q]}\right),
\end{aligned}$$

where we denoted $\mathbf{1}_q = (0, \dots, 0, 1, 0, \dots, 0)$ (namely the $q$'th coordinate vector in $\mathbb{R}^Q$) and $\mathbf{2}_q = (0, \dots, 0, 2, 0, \dots, 0) = 2\mathbf{1}_q$.

We get similar expressions for $\hat{\boldsymbol{U}}_{ij}$ with $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})$ replaced by $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\hat{\sigma}'_{\boldsymbol{d},\boldsymbol{\tau}})$. Because we defined $\sigma'$ and $\hat{\sigma}'$ by only modifying the $\boldsymbol{l}_1$th and $\boldsymbol{l}_2$th coefficients, we get

$$\begin{aligned}\mathrm{Tr}(\bar{\boldsymbol{U}}_{ij}^{(qq)}) &= \mathrm{Tr}(\boldsymbol{U}_{ij}^{(qq)} - \hat{\boldsymbol{U}}_{ij}^{(qq)}) \\ &= r_q^2 \sum_{t=1,2} \delta_t (2-\delta_t) \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) B(\boldsymbol{d}, \boldsymbol{l}_t) Q_{\boldsymbol{l}_t}^{\boldsymbol{d}} \left( \{ d_q \gamma_{ij}^{(q)} \}_{q \in [Q]} \right).\end{aligned} \tag{126}$$

Recalling that $\lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{1}_q}$ only depend on $\lambda_{\boldsymbol{k}-\mathbf{1}_q}^{\boldsymbol{d}}$ and $\lambda_{\boldsymbol{k}+\mathbf{1}_q}^{\boldsymbol{d}}$, and $\lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{2}_q}$ on $\lambda_{\boldsymbol{k}-\mathbf{2}_q}^{\boldsymbol{d}}$, $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}$ and $\lambda_{\boldsymbol{k}+\mathbf{2}_q}^{\boldsymbol{d}}$, (lemma 13), we get

$$\begin{aligned}\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)} \boldsymbol{\theta}_j^{(q)} \rangle &= r_q^2 \tau_i^{(q)} \tau_j^{(q)} \sum_{t=\{1,2\}, \boldsymbol{k} \in \{\boldsymbol{l}_t \pm \mathbf{1}_q\}} \delta_t (2-\delta_t) \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) \\ &\quad \times B(\boldsymbol{d}, \boldsymbol{k}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ d_q \gamma_{ij}^{(q)} \}_{q \in [Q]} \right), \\ \langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)} \boldsymbol{\theta}_i^{(q)} \rangle &= r_q^2 (\tau_i^{(q)})^2 \sum_{t \in \{1,2\}, \boldsymbol{k} \in \{\boldsymbol{l}_t, \boldsymbol{l}_t \pm \mathbf{2}_q\}} \delta_t (2-\delta_t) \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) \\ &\quad \times B(\boldsymbol{d}, \boldsymbol{k}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ d_q \gamma_{ij}^{(q)} \}_{q \in [Q]} \right), \\ \langle \boldsymbol{\theta}_j^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)} \boldsymbol{\theta}_j^{(q)} \rangle &= r_q^2 (\tau_j^{(q)})^2 \sum_{t \in \{1,2\}, \boldsymbol{k} \in \{\boldsymbol{l}_t, \boldsymbol{l}_t \pm \mathbf{2}_q\}} \delta_t (2-\delta_t) \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{k}}^{\boldsymbol{d},\mathbf{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) \\ &\quad \times B(\boldsymbol{d}, \boldsymbol{k}) Q_{\boldsymbol{k}}^{\boldsymbol{d}} \left( \{ d_q \gamma_{ij}^{(q)} \}_{q \in [Q]} \right),\end{aligned} \tag{127}$$

where we used the convention $\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) = 0$ if one of the coordinates verifies $k_q < 0$.

From lemmas 13, 19 and 20, we get for $t = 1, 2$ and $q \neq q_\xi$:

$$\begin{aligned}\lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j) \to (+\infty,\mathbf{1},\mathbf{1})} \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) B(\boldsymbol{d}, \boldsymbol{l}_t) &= \frac{\mu_{l_t}(\sigma')^2}{l_t!}, \\ \lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j) \to (+\infty,\mathbf{1},\mathbf{1})} \lambda_{\boldsymbol{l}_t+\mathbf{1}_q}^{\boldsymbol{d},\mathbf{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{l}_t+\mathbf{1}_q}^{\boldsymbol{d},\mathbf{1}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) B(\boldsymbol{d}, \boldsymbol{l}_t + \mathbf{1}_q) &= \frac{\mu_{l_t}(\sigma')^2}{l_t!}, \\ \lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j) \to (+\infty,\mathbf{1},\mathbf{1})} \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d},\mathbf{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{l}_t}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) B(\boldsymbol{d}, \boldsymbol{l}_t) &= \frac{\mu_{l_t}(\sigma')^2}{l_t!}, \\ \lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j) \to (+\infty,\mathbf{1},\mathbf{1})} \lambda_{\boldsymbol{l}_t+\mathbf{2}_q}^{\boldsymbol{d},\mathbf{2}_q}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) \lambda_{\boldsymbol{l}_t+\mathbf{2}_q}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_j}) B(\boldsymbol{d}, \boldsymbol{l}_t + \mathbf{2}_q) &= 0,\end{aligned} \tag{128}$$

while for $q = q_\xi$ and $u \in \{-1, 1\}$,

$$\begin{aligned}&\lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j) \to (+\infty,\mathbf{1},\mathbf{1})} \lambda_{\boldsymbol{l}_t+u\mathbf{1}_{q_\xi}}^{\boldsymbol{d},\mathbf{1}_{q_\xi}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i}) [B(\boldsymbol{d}, \boldsymbol{l}_t + u\mathbf{1}_{q_\xi})(l_t+u)!]^{1/2} \\ &= \mu_{l_t+u+1}(\sigma') + (l_t+u)\mu_{l_t+u-1}(\sigma'),\end{aligned} \tag{129}$$

and for $v \in \{-2, 0, 2\}$,

$$\begin{aligned}&\lim_{(d,\boldsymbol{\tau}_i,\boldsymbol{\tau}_j)\to(+\infty,\mathbf{1},\mathbf{1})} \lambda^{\boldsymbol{d},\mathbf{2}_{q_\xi}}_{\boldsymbol{l}_t+v\mathbf{1}_{q_\xi}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}_i})[B(\boldsymbol{d},\boldsymbol{l}_t+v\mathbf{1}_{q_\xi})(l_t+v)!]^{1/2}\\ &\quad= \mu_{l_t+v+2}(\sigma') + (2l_t+2v+1)\mu_{l_t+v}(\sigma') + (l_t+v)(l_t+v-1)\mu_{l_t+v-2}(\sigma'). \end{aligned} \tag{130}$$

From lemma 26, we recall that the coefficients of the $k$th Gegenbauer polynomial $Q_k^{(d)}(x) = \sum_{s=0}^{k} p_{k,s}^{(d)} x^s$ satisfy

$$p_{k,s}^{(d)} = O_d(d^{-k/2-s/2}). \tag{131}$$

Furthermore, lemma 27 shows that $\max_{i\neq j}|\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle| = O_{d,\mathbb{P}}(\sqrt{d_q \log d_q})$. We deduce that

$$\max_{i\neq j}|Q_{k_q}^{(d_q)}(\langle \overline{\boldsymbol{\theta}}_i^{(q)}, \overline{\boldsymbol{\theta}}_j^{(q)}\rangle)| = \tilde{O}_{d,\mathbb{P}}(d_q^{-k_q/2}). \tag{132}$$

Plugging the estimates (128) and (132) into equations (126) and (127), we obtain that

$$\begin{aligned}&\max_{i\neq j}\left\{\left|\mathrm{Tr}(\bar{\boldsymbol{U}}_{ij}^{(qq)})\right|,\ \left|\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)}\boldsymbol{\theta}_j^{(q)}\rangle\right|,\ \left|\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle,\ \left|\langle \boldsymbol{\theta}_j^{(q)}, \bar{\boldsymbol{U}}_{ij}^{(qq)}\boldsymbol{\theta}_j^{(q)}\rangle\right|\right.\right\}\\ &\quad= \tilde{O}_{d,\mathbb{P}}(d^{2\xi}d^{-\eta_q l_1/2}). \end{aligned} \tag{133}$$

From equation (125), using the fact that $\max_{i\neq j}|\gamma_{ij}^{(q)}| = O_{d,\mathbb{P}}(\sqrt{(\log d_q)/d_q})$ and Cramer's rule for matrix inversion, it is easy to see that

$$\max_{i\neq j}\max_{l,k\in[4]}\left|((\boldsymbol{M}_{ij}^{(qq)})^{-1})_{lk}\right| = O_{d,\mathbb{P}}(1). \tag{134}$$

We deduce from (124), (133) and (134) that for $a \in [3], b \in [2]$,

$$\max_{i\neq j}\{|\bar{u}_{a,b}^{(qq)}(\boldsymbol{\theta}_i^{(q)}, \boldsymbol{\theta}_j^{(q)})|\} = \tilde{O}_{d,\mathbb{P}}(d^{2\xi}d^{-\eta_q l_1/2}). \tag{135}$$

As a result, combining equation (135) with equation (121) in the expression of $\overline{u}^{(qq)}$ given in lemma 14, we get

$$\begin{aligned}&\max_{i\neq j}\|\,\bar{\boldsymbol{U}}_{ij}^{(qq)}\,\|_F^2\\ &\quad= \max_{i\neq j}\|\bar{u}_1^{(qq)}\mathbf{I}_{d_q} + \bar{u}_2^{(qq)}[\boldsymbol{\theta}_i^{(q)}(\boldsymbol{\theta}_j^{(q)})^{\mathsf{T}} + \boldsymbol{\theta}_j^{(q)}(\boldsymbol{\theta}_i^{(q)})^{\mathsf{T}}] + \bar{u}_{3,1}^{(qq)}\boldsymbol{\theta}_i^{(q)}(\boldsymbol{\theta}_i^{(q)})^{\mathsf{T}} + \bar{u}_{3,2}^{(qq)}\boldsymbol{\theta}_j^{(q)}(\boldsymbol{\theta}_j^{(q)})^{\mathsf{T}}\|_F^2\\ &\quad\leqslant \tilde{O}_{d,\mathbb{P}}(d^{6\xi}d^{-\eta_q l_1}). \end{aligned}$$

A similar computation shows that

$$\max_{i\neq j}\|\,\bar{\boldsymbol{U}}_{ij}^{(qq')}\,\|_F^2 \leqslant \tilde{O}_{d,\mathbb{P}}(d^{6\xi}d^{-\eta_q l_1}).$$

By the expression of $\boldsymbol{\Delta}$ given by (123), we conclude that

$$\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2 \leqslant \|\boldsymbol{\Delta}\|_F^2 = \sum_{q,q'\in[Q]}\sum_{i,j=1,i\neq j}^{N} \|\bar{\boldsymbol{U}}_{ij}^{(qq')}\|_F^2 = \tilde{O}_{d,\mathbb{P}}(N^2 d^{6\xi-\eta_q l_1}).$$

By assumption, $N = o_d(d^{\gamma})$. Hence, since by assumption $\eta_q l_1 \geqslant 2\gamma + 7\xi$, we deduce that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(d^{-\xi}) = o_{d,\mathbb{P}}(d^{-\max_{q\in[Q]}\kappa_q})$.

**Step 5. Checking the properties of matrix $\boldsymbol{D}$.**

By lemma 14, we can express $\bar{\boldsymbol{U}}_{ii}$ as a block matrix with

$$\bar{\boldsymbol{U}}_{ii}^{(qq)} = \alpha^{(q)}\mathbf{I}_{d_q} + \beta^{(q)}\boldsymbol{\theta}_i^{(q)}(\boldsymbol{\theta}_i^{(q)})^{\mathsf{T}}, \qquad \bar{\boldsymbol{U}}_{ii}^{(qq')} = \beta^{(qq')}\boldsymbol{\theta}_i^{(q)}(\boldsymbol{\theta}_i^{(q')})^{\mathsf{T}},$$

with coefficients given by

$$\begin{aligned}\begin{bmatrix}\alpha^{(q)}\\ \beta^{(q)}\end{bmatrix} &= [d_q(d_q-1)(\tau_i^{(q)})^4]^{-1}\begin{bmatrix} d_q(\tau_i^{(q)})^4 & -(\tau_i^{(q)})^2\\ -(\tau_i^{(q)})^2 & 1\end{bmatrix}\times\begin{bmatrix}\mathrm{Tr}(\bar{\boldsymbol{U}}_{ii}^{(qq)})\\ \langle\boldsymbol{\theta}_i^{(q)},\bar{\boldsymbol{U}}_{ii}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle\end{bmatrix},\\ \beta^{(qq')} &= (d_q d_{q'})^{-1}(\tau_i^{(q)}\tau_i^{(q')})^{-2}\langle\boldsymbol{\theta}_i^{(q)},\bar{\boldsymbol{U}}_{ii}^{(qq')}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)\boldsymbol{\theta}_i^{(q')}\rangle.\end{aligned} \tag{136}$$

Let us first focus on the $q = q_\xi$ sphere. Using equations (126) and (127) with the expressions (129) and (130), we get the following convergence in probability (using that $\{\tau_i^{(q)}\}_{i\in[N]}$ concentrates on 1),

$$\begin{aligned}\sup_{i\in[N]}\left|r_{q_\xi}^{-2}\mathrm{Tr}(\bar{\boldsymbol{U}}_{ii}^{(q_\xi q_\xi)}) - F_1(\boldsymbol{\delta})\right| &\xrightarrow{\mathbb{P}} 0,\\ \sup_{i\in[N]}\left|r_{q_\xi}^{-2}\langle\boldsymbol{\theta}_i^{(q_\xi)},\bar{\boldsymbol{U}}_{ii}^{(q_\xi q_\xi)}\boldsymbol{\theta}_i^{(q_\xi)}\rangle - F_2(\boldsymbol{\delta})\right| &\xrightarrow{\mathbb{P}} 0,\end{aligned} \tag{137}$$

where we denoted $\boldsymbol{\delta} = (\delta_1,\delta_2)$ (where $\delta_1,\delta_2$ first appears in the definition of $\hat{\sigma}$ in equation (115), and till now $\delta_1,\delta_2$ are still not determined) and, similarly to the proof of proposition 5 in [21] and letting $\mu_k \equiv \ \mu_k(\sigma')$, we have

$$F_1(\boldsymbol{\delta}) = \sum_{t\in\{1,2\}}\delta_t(2-\delta_t)\frac{\mu_{l_t}^2}{l_t!}, \tag{138}$$

while, for $l_2 \neq l_1 + 2$

$$\begin{aligned}F_2(\boldsymbol{\delta}) = \sum_{t\in\{1,2\}}\Bigg\{&\frac{1}{(l_t-1)!}\left[(\mu_{l_t}+(l_t-1)\mu_{l_t-2})^2 - ((1-\delta_t)\mu_{l_t}+(l_t-1)\mu_{l_t-2})^2\right]\\ &+\frac{1}{(l_t+1)!}\left[(\mu_{l_t+2}+(l_t+1)\mu_{l_t})^2 - (\mu_{l_t+2}+(1-\delta_t)(l_t+1)\mu_{l_t})^2\right]\Bigg\},\end{aligned}$$

while, for $l_2 = l_1 + 2$

$$\begin{aligned} F_2(\boldsymbol{\delta}) = &\frac{1}{(l_1-1)!}\left[(\mu_{l_1} + (l_1-1)\mu_{l_1-2})^2 - ((1-\delta_1)\mu_{l_1} + (l_1-1)\mu_{l_1-2})^2\right] \\ &+ \frac{1}{(l_1+1)!}\left[(\mu_{l_1+2} + (l_1+1)\mu_{l_1})^2 - ((1-\delta_2)\mu_{l_1+2} + (1-\delta_1)(l_1+1)\mu_{l_1})^2\right] \\ &+ \frac{1}{(l_2+1)!}\left[(\mu_{l_2+2} + (l_2+1)\mu_{l_2})^2 - (\mu_{l_2+2} + (1-\delta_2)(l_2+1)\mu_{l_2})^2\right]. \end{aligned}$$

We have from equation (136),

$$\begin{aligned} &\lambda_{\min}(\bar{U}_{ii}^{(qq)}) \\ &\quad = \min\left\{\alpha^{(q)}, \alpha^{(q)} + \beta^{(q)} d_q (\tau_i^{(q)})^2\right\} \\ &\quad = \min\left\{\frac{1}{d_q-1}\mathrm{Tr}(\bar{\boldsymbol{U}}_{ii}^{(qq)}) - \frac{1}{d_q(d_q-1)(\tau_i^{(q)})^2}\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ii}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle, \frac{1}{d_q(\tau_i^{(q)})^2}\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ii}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle\right\}. \end{aligned}$$

Hence, using equation (137), we get

$$\sup_{i\in[N]}\left|\frac{d_{q_\xi}}{r_{q_\xi}^2}\lambda_{\min}(\bar{U}_{ii}^{(q_\xi q_\xi)}) - \min\{F_1(\boldsymbol{\delta}), F_2(\boldsymbol{\delta})\}\right| \xrightarrow{\mathbb{P}} 0. \tag{139}$$

Following the same reasoning as in proposition 5 in [21], we can verify that under assumption 3(b), we have $\nabla F_1(\mathbf{0}), \nabla F_2(\mathbf{0}) \neq \mathbf{0}$ and $\det(\nabla F_1(\mathbf{0}), \nabla F_2(\mathbf{0})) \neq 0$. We can therefore find $\boldsymbol{\delta} = (\delta_1, \delta_2)$ such that $F_1(\boldsymbol{\delta}) > 0$, $F_2(\boldsymbol{\delta}) > 0$. Furthermore,

$$\sup_{i\in[N]}\left|\frac{d_{q_\xi}}{r_{q_\xi}^2}\lambda_{\max}(\bar{U}_{ii}^{(q_\xi q_\xi)}) - \max\{F_1(\boldsymbol{\delta}), F_2(\boldsymbol{\delta})\}\right| \xrightarrow{\mathbb{P}} 0. \tag{140}$$

Similarly, we get for $q \neq q_\xi$ from equations (126) and (127) with the expressions (128) (recalling that $\{\tau_i^{(q)}\}_{i\in[N]}$ concentrates on 1),

$$\begin{aligned} \sup_{i\in[N]}\left|r_q^{-2}\mathrm{Tr}(\bar{\boldsymbol{U}}_{ii}^{(qq)}) - F_1(\boldsymbol{\delta})\right| &\xrightarrow{\mathbb{P}} 0, \\ \sup_{i\in[N]}\left|r_q^{-2}\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ii}^{(qq)}\boldsymbol{\theta}_i^{(q)}\rangle - F_1(\boldsymbol{\delta})\right| &\xrightarrow{\mathbb{P}} 0, \\ \sup_{i\in[N]}\left|(r_q r_{q'})^{-1}\langle \boldsymbol{\theta}_i^{(q)}, \bar{\boldsymbol{U}}_{ii}^{(qq')}\boldsymbol{\theta}_i^{(q')}\rangle\right| &\xrightarrow{\mathbb{P}} 0. \end{aligned} \tag{141}$$

We deduce that for $q \neq q_\xi$ and $q \neq q'$,

$$\sup_{i\in[N]} \left| \frac{d_q}{r_q^2} \lambda_{\min}(\bar{U}_{ii}^{(qq)}) - F_1(\boldsymbol{\delta}) \right| \xrightarrow{\mathbb{P}} 0,$$

$$\sup_{i\in[N]} \left| \frac{d_q}{r_q^2} \lambda_{\max}(\bar{U}_{ii}^{(qq)}) - F_1(\boldsymbol{\delta}) \right| \xrightarrow{\mathbb{P}} 0,$$

$$\sup_{i\in[N]} \left| \frac{(d_q d_{q'})^{1/2}}{r_q r_{q'}} \sigma_{\max}(\bar{U}_{ii}^{(qq')}) \right| \xrightarrow{\mathbb{P}} 0,$$

which finishes to prove properties (105) and (106).

## Appendix H. Proof of theorem 7(b): upper bound for NT model

### H.1. Preliminaries

**Lemma 15.** *Let $\sigma$ be an activation function that satisfies assumptions 3(a) and (c) for some level $\gamma > 0$. Let $\mathcal{Q} = \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma)$ as defined in equation (49). Define for integer $\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q}$ and $\boldsymbol{\tau}, \boldsymbol{\tau}' \in \mathbb{R}_{\geqslant 0}^{Q}$,*

$$A^{(q)}_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}} = r_q^2 \cdot \Big[ t_{d_q,k_q-1} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q-}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}) B(\boldsymbol{d}, \boldsymbol{k}_{q-}) + s_{d_q,k_q+1} \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}_{q+}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}) B(\boldsymbol{d}, \boldsymbol{k}_{q+}) \Big], \tag{142}$$

*with $\mathbf{k}_{q+} = (k_1, \ldots, k_q + 1, \ldots, k_Q)$ and $\mathbf{k}_{q-} = (k_1, \ldots, k_q - 1, \ldots, k_Q)$, and*

$$s_{d,k} = \frac{k}{2k+d-2}, \qquad t_{d,k} = \frac{k+d-2}{2k+d-2},$$

*with the convention $t_{d,-1} = 0$.*

*Then there exists constants $\varepsilon_0 > 0$ and $C > 0$ such that for $d$ large enough, we have for any $\boldsymbol{\tau}, \boldsymbol{\tau}' \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$,*

$$\max_{\boldsymbol{k}\in\mathcal{Q}} \frac{B(\boldsymbol{d}, \boldsymbol{k})}{A^{(q)}_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}}} \leqslant C d^{\gamma-\kappa_q}.$$

**Proof of lemma 15.** From assumptions 3(a) and (c) and lemma 19, there exists $c > 0$ and $\varepsilon_0 > 0$ such that for any $\boldsymbol{\tau}, \boldsymbol{\tau}' \in [1-\varepsilon_0, 1+\varepsilon_0]^Q$ and $\boldsymbol{k} \in \mathcal{Q}$,

$$\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}) \geqslant c \prod_{q\in[Q]} d^{k_q(\kappa_q-\xi)}.$$

Hence for $k_q > 0$, we get $\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}) \geqslant c d^{-\gamma-\xi+\kappa_q}$, and for $k_q = 0$, we get $\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) \lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}) \geqslant c d^{-\gamma+\xi-\eta_q-\kappa_q}$. Carefully injecting these bounds in equation (142) yields the lemma. □

### H.2. Proof of theorem 7(b): outline

In this proof, we will consider $Q$ sub-classes of functions corresponding to the NT model restricted to the $q$th sphere:

$$\mathcal{F}_{\mathrm{NT}^{(q)}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}^{(q)} \rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle / R) : \boldsymbol{a}_i \in \mathbb{R}^{d_q}, i \in [N] \right\}.$$

We define similarly the risk associated to this sub-model

$$R_{\mathrm{NT}^{(q)}}(f_d, \boldsymbol{W}) = \inf_{f \in \mathcal{F}_{\mathrm{NT}^{(q)}}(\boldsymbol{W})} \mathbb{E}[(f_d(\boldsymbol{x}) - f(\boldsymbol{x}))^2].$$

and approximation subspace

$$\begin{aligned} \overline{\mathcal{Q}}_{\mathrm{NT}^{(q)}}(\gamma) = & \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q} \,\middle|\, k_q > 0 \quad \text{and} \quad \sum_{q \in [Q]} k_q(\xi - \kappa_q) \leqslant \gamma + (\xi - \kappa_q) \right\} \\ & \cup \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q} \,\middle|\, k_q = 0 \quad \text{and} \quad \sum_{q \in [Q]} k_q(\xi - \kappa_q) \leqslant \gamma - (\xi - \kappa_q - \eta_q) \right\}. \end{aligned} \tag{143}$$

**Theorem 8.** *Let $\{f_d \in L^2(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}, \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}})\}_{d \geqslant 1}$ be a sequence of functions. Let $\boldsymbol{W} = (\boldsymbol{w}_i)_{i \in [N]}$ with $(\boldsymbol{w}_i)_{i \in [N]} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ independently. Assume $N \geqslant \omega_d(d^\gamma)$ for some positive constant $\gamma > 0$, and $\sigma$ satisfy assumptions 3(a) and (c) at level $\gamma$. Then for any $\varepsilon > 0$, the following holds with high probability:*

$$0 \leqslant R_{\mathrm{NT}^{(q)}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) \leqslant \varepsilon \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2, \tag{144}$$

*where $\mathcal{Q} \equiv \overline{\mathcal{Q}}_{\mathrm{NT}^{(q)}}(\gamma)$ is defined in equation (143).*

**Remark 5.** From the proof of theorem 7(a), we have a matching lower bound for $\mathcal{F}_{\mathrm{NT}^{(q)}}$. We recall

$$\overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^{Q} \,\middle|\, \sum_{q=1}^{Q} (\xi - \kappa_q) k_q \leqslant \gamma + \left( \xi - \min_{q \in S(\boldsymbol{k})} \kappa_q \right) \right\}.$$

Notice that

$$\overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma) = \bigcup_{q \in Q} \overline{\mathcal{Q}}_{\mathrm{NT}^{(q)}}(\gamma).$$

Denote $q_{\boldsymbol{k}} = \mathrm{argmin}_{q \in S(\boldsymbol{k})} \kappa_q$, such that $\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{NT}^{(q_{\boldsymbol{k}})}}$ for any $\boldsymbol{k} \in \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma)$. Furthermore, notice that by definition for any $f \in L^2(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}, \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}})$ and $q \in [Q]$,

$$R_{\mathrm{NT}}(f, \boldsymbol{W}) \leqslant R_{\mathrm{NT}^{(q)}}(f, \boldsymbol{W}).$$

Let us deduce theorem 7(b) from theorem 8. Denote $\mathcal{Q} = \overline{\mathcal{Q}}_{\mathrm{NT}}(\gamma)$. We divide the $N$ neurons in $|\mathcal{Q}|$ sections of size $N' = N/|\mathcal{Q}|$, i.e. $\boldsymbol{W} = (\boldsymbol{W}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{Q}}$ where $\boldsymbol{W}_{\boldsymbol{k}} \in \mathbb{R}^{N'\times d}$. For any $\varepsilon > 0$, we get from theorem 8 that with high probability

$$R_{\mathrm{NT}}(\mathsf{P}_{\mathcal{Q}} f_d, \boldsymbol{W}) \leqslant \sum_{\boldsymbol{k}\in\mathcal{Q}} R_{\mathrm{NT}^{(\boldsymbol{q_k})}}(\mathsf{P}_{\boldsymbol{k}} f, \boldsymbol{W}_{\boldsymbol{k}}) \leqslant \sum_{\boldsymbol{k}\in\mathcal{Q}} \varepsilon \|\mathsf{P}_{\boldsymbol{k}} f_d\|_{L^2}^2 = \varepsilon \|\mathsf{P}_{\mathcal{Q}} f_d\|_{L^2}^2.$$

## H.3. Proof of theorem 8

*H.3.1. Properties of the limiting kernel.* Similarly to the proof of theorem 6(b), we construct a limiting kernel which is used as a proxy to upper bound the $\mathrm{NT}^{(q)}$ risk.

We recall the definition of $\mathrm{PS}^{\boldsymbol{d}} = \prod_{q\in[Q]} \mathbb{S}^{d_q-1}(\sqrt{d_q})$. We introduce $\mathcal{L} = L^2(\mathrm{PS}^{\boldsymbol{d}} \to \mathbb{R}, \mu_{\boldsymbol{d}})$ and $\mathcal{L}_{d_q} = L^2(\mathrm{PS}^{\boldsymbol{d}} \to \mathbb{R}^{d_q}, \mu_{\boldsymbol{d}})$. For a given $\boldsymbol{\theta} \in \mathbb{S}^{D-1}(\sqrt{D})$ and associated vector $\boldsymbol{\tau} \in \mathbb{R}^Q_{\geqslant 0}$, recall the definition of $\sigma'_{\boldsymbol{d},\boldsymbol{\tau}} \in \mathcal{L}$:

$$\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) = \sigma'\left(\sum_{q\in[Q]} \tau^{(q)} \cdot (r_q/R) \cdot \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\right).$$

For any $\boldsymbol{\tau} \in \mathbb{R}^Q_{\geqslant 0}$, define the operator $\mathbb{T}_{\boldsymbol{\tau}} : \mathcal{L} \to \mathcal{L}_{d_q}$, such that for any $g \in \mathcal{L}$,

$$\mathbb{T}_{\boldsymbol{\tau}} g(\overline{\boldsymbol{\theta}}) = \frac{r_q}{\sqrt{d_q}} \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\overline{\boldsymbol{x}}^{(q)} \sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) g(\overline{\boldsymbol{x}})\right].$$

The adjoint operator $\mathbb{T}^*_{\boldsymbol{\tau}} : \mathcal{L}_{d_q} \to \mathcal{L}$ verifies for any $h \in \mathcal{L}_{d_q}$,

$$\mathbb{T}^*_{\boldsymbol{\tau}} h(\overline{\boldsymbol{x}}) = \frac{r_q}{\sqrt{d_q}} (\overline{\boldsymbol{x}}^{(q)})^{\mathsf{T}} \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) h(\overline{\boldsymbol{\theta}})\right].$$

We define the operator $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} : \mathcal{L}_{d_q} \to \mathcal{L}_{d_q}$ as $\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} \equiv \mathbb{T}_{\boldsymbol{\tau}} \mathbb{T}^*_{\boldsymbol{\tau}'}$. For $h \in \mathcal{L}_{d_q}$, we can write

$$\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'} h(\overline{\boldsymbol{\theta}}_1) = \mathbb{E}_{\overline{\boldsymbol{\theta}}_2}[\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'}(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2) h(\overline{\boldsymbol{\theta}}_2)],$$

where

$$\begin{aligned}
&\mathbb{K}_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2}(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2) \\
&\quad = \frac{r_q^2}{d_q} \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\overline{\boldsymbol{x}}^{(q)} (\overline{\boldsymbol{x}}^{(q)})^{\mathsf{T}} \sigma'_{\boldsymbol{d},\boldsymbol{\tau}_1}\left(\{\langle \overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) \sigma'_{\boldsymbol{d},\boldsymbol{\tau}_2}\left(\{\langle \overline{\boldsymbol{\theta}}_2^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right].
\end{aligned}$$

Define $\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'} : \mathcal{L} \to \mathcal{L}$ as $\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'} \equiv \mathbb{T}^*_{\boldsymbol{\tau}} \mathbb{T}_{\boldsymbol{\tau}'}$. For $g \in \mathcal{L}$, we can write

$$\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'} g(\overline{\boldsymbol{x}}_1) = \mathbb{E}_{\overline{\boldsymbol{x}}_2}[\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'}(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2) g(\overline{\boldsymbol{x}}_2)]$$

where

$$\begin{aligned}
&\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'}(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2) \\
&\quad = \frac{r_q^2}{d_q} \mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}_1^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right) \sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}_2^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right] \langle \overline{\boldsymbol{x}}_1^{(q)}, \overline{\boldsymbol{x}}_2^{(q)}\rangle.
\end{aligned}$$

We recall the decomposition of $\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of tensor product of Gegenbauer polynomials:

$$\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)}) = \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\dots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right),$$

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{x}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\dots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right].$$

Following the same computations as in lemma 12, we get

$$\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'}(\overline{\boldsymbol{x}}_1,\overline{\boldsymbol{x}}_2) = \sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q} A_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}}^{(q)} Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{x}}_1^{(q)},\overline{\boldsymbol{x}}_2^{(q)}\rangle\}_{q\in[Q]}\right),$$

where

$$A_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}}^{(q)} = r_q^2\cdot\left[t_{d_q,k_q-1}\lambda_{\boldsymbol{k}_{q-}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})\lambda_{\boldsymbol{k}_{q-}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'})B(\boldsymbol{d},\boldsymbol{k}_{q-})\right.$$
$$\left.+\, s_{d_q,k_q+1}\lambda_{\boldsymbol{k}_{q+}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}})\lambda_{\boldsymbol{k}_{q+}}^{\boldsymbol{d}}(\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'})B(\boldsymbol{d},\boldsymbol{k}_{q+})\right], \tag{145}$$

with $\boldsymbol{k}_{q+} = (k_1,\dots,k_q+1,\dots,k_Q)$ and $\boldsymbol{k}_{q-} = (k_1,\dots,k_q-1,\dots,k_Q)$, and convention $t_{d_q,-1} = 0$,

$$s_{d_q,k_q} = \frac{k_q}{2k_q+d_q-2}, \qquad t_{d_q,k_q} = \frac{k_q+d_q-2}{2k_q+d_q-2}.$$

Recall that for $\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q$ and $\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]$, $Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}} = \bigotimes_{q\in[Q]} Y_{k_q s_q}^{(d_q)}$ forms an orthogonal basis of $\mathcal{L}$ and that

$$\mathbb{E}_{\overline{\boldsymbol{x}}_2}\left[Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{x}}_1,\overline{\boldsymbol{x}}_2\rangle\}_{q\in[Q]}\right)Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_2)\right] = \frac{1}{B(\boldsymbol{d},\boldsymbol{k})}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_1)\delta_{\boldsymbol{k},\boldsymbol{s}}.$$

We deduce that

$$\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_1) = \sum_{\boldsymbol{k}'\in\mathbb{Z}_{\geqslant 0}^Q} A_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}'}^{(q)}\mathbb{E}_{\overline{\boldsymbol{x}}_2}\left[Q_{\boldsymbol{k}'}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{x}}_1,\overline{\boldsymbol{x}}_2\rangle\}_{q\in[Q]}\right)Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_2)\right]$$
$$= \frac{A_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{\boldsymbol{x}}_1).$$

Consider $\{\mathbb{T}_{\boldsymbol{\tau}}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}\}_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q,\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]}$. We have:

$$\langle\mathbb{T}_{\boldsymbol{\tau}}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}},\mathbb{T}_{\boldsymbol{\tau}'}Y_{\boldsymbol{k}',\boldsymbol{s}'}^{\boldsymbol{d}}\rangle_{L^2} = \langle Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}},\mathbb{H}_{\boldsymbol{\tau},\boldsymbol{\tau}'}Y_{\boldsymbol{k}',\boldsymbol{s}'}^{\boldsymbol{d}}\rangle_{L^2} = \frac{A_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})}\delta_{\boldsymbol{k},\boldsymbol{k}'}\delta_{\boldsymbol{s},\boldsymbol{s}'},$$

$$\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}'}\mathbb{T}_{\boldsymbol{\tau}''}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}} = \mathbb{T}_{\boldsymbol{\tau}}\mathbb{H}_{\boldsymbol{\tau}',\boldsymbol{\tau}''}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}} = \frac{A_{(\boldsymbol{\tau}',\boldsymbol{\tau}''),\boldsymbol{k}}^{(q)}}{B(\boldsymbol{d},\boldsymbol{k})}\mathbb{T}_{\boldsymbol{\tau}}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}.$$

Hence $\{\mathbb{T}_{\boldsymbol{\tau}''}Y_{\boldsymbol{k},\boldsymbol{s}}^{(\boldsymbol{d})}\}$ forms an orthogonal basis that diagonalizes $\mathbb{K}_{\boldsymbol{\tau}',\boldsymbol{\tau}''}$ (notice that $\mathbb{T}_{\boldsymbol{\tau}}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}$ is parallel to $\mathbb{T}_{\boldsymbol{\tau}'}Y_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}$ for any $\boldsymbol{\tau},\boldsymbol{\tau}'\in\mathbb{R}_{\geqslant 0}^Q$). Let us consider the subspace $\mathbb{T}_{\boldsymbol{\tau}}(V_{\mathcal{Q}}^{\boldsymbol{d}})$, the image of $V_{\mathcal{Q}}^{\boldsymbol{d}}$ by the operator $\mathbb{T}_{\boldsymbol{\tau}}$. From assumptions 3(a) and (b) and lemma 19, there exists $\varepsilon_0 \in (0, 1)$ and $d_0$ such that for any $\boldsymbol{\tau}, \boldsymbol{\tau}' \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$ and $d \geqslant d_0$, we have $A^{(q)}_{(\boldsymbol{\tau},\boldsymbol{\tau}'),\boldsymbol{k}} > 0$ for any $\boldsymbol{k} \in \mathcal{Q}$, and therefore the inverse $\mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}'}|_{\mathbb{T}_{\boldsymbol{\tau}}(V_{\mathcal{Q}}^{\boldsymbol{d}})}$ (restricted to $\mathbb{T}_{\boldsymbol{\tau}}(V_{\mathcal{Q}}^{\boldsymbol{d}})$) is well defined.

*H.3.2. Proof of theorem 8.* Let us assume that $\{f_d\}$ is contained in $\bigoplus_{\boldsymbol{k}\in\mathcal{Q}} \boldsymbol{V}_{\boldsymbol{k}}^{\boldsymbol{d}}$, i.e. $\overline{f}_d = \mathsf{P}_{\mathcal{Q}}\overline{f}_d$.

Consider

$$\hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}^{(q)} \rangle \sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / R).$$

Define $\boldsymbol{\alpha}_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}) \equiv \mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}} \mathbb{T}_{\boldsymbol{\tau}} \overline{f}_d(\overline{\boldsymbol{\theta}})$ and choose $\boldsymbol{a}_i^* = N^{-1} \boldsymbol{\alpha}_{\boldsymbol{\tau}_i}(\overline{\boldsymbol{\theta}}_i)$, where we denoted $\overline{\boldsymbol{\theta}}_i = (\overline{\boldsymbol{\theta}}_i^{(q)})_{q\in[Q]}$ with $\overline{\boldsymbol{\theta}}_i^{(q)} = \boldsymbol{\theta}_i^{(q)}/\tau_i^{(q)} \in \mathbb{S}^{d_q - 1}(\sqrt{d_q})$ independent of $\boldsymbol{\tau}_i$.

Fix $\varepsilon_0 > 0$ as prescribed in lemma 15 and consider the expectation over $\mathcal{P}_{\varepsilon_0}$ of the $\mathrm{NT}^{(q)}$ risk (in particular, $\boldsymbol{a}^* = (\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_N^*) \in \mathbb{R}^{N d_q}$ are well defined):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}[R_{\mathrm{NT}^{(q)}}(f_d, \boldsymbol{\Theta})] &= \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\inf_{\boldsymbol{a}\in\mathbb{R}^{N d_q}} \mathbb{E}_{\boldsymbol{x}}[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}))^2]\right] \\
&\leqslant \mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{a}^*(\boldsymbol{\Theta})))^2\right]\right].
\end{aligned}$$

We can expand the squared loss at $\boldsymbol{a}$ as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2] = \|f_d\|_{L^2}^2 - 2\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i, \boldsymbol{x}^{(q)} \rangle \sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / R) f_d(\boldsymbol{x})] \\
+ \sum_{i,j=1}^{N} \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i, \boldsymbol{x}^{(q)} \rangle \langle \boldsymbol{a}_j, \boldsymbol{x}^{(q)} \rangle \sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / R) \sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x} \rangle / R)].
\end{aligned} \tag{146}$$

The second term of the expansion (146) around $\boldsymbol{a}^*$ verifies

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i^*, \boldsymbol{x}^{(q)} \rangle \sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / R) f_d(\boldsymbol{x})]\right] \\
&\quad = \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\boldsymbol{\alpha}_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}})^{\mathsf{T}} \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\boldsymbol{x}^{(q)} \sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)} \rangle / \sqrt{d_q}\}_{q\in[Q]}\right) \overline{f}_d(\overline{\boldsymbol{x}})\right]\right]\right] \\
&\quad = \mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}\left[\langle \mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}} \mathbb{T}_{\boldsymbol{\tau}} \overline{f}_d, \mathbb{T}_{\boldsymbol{\tau}} \overline{f}_d \rangle_{L^2}\right] \\
&\quad = \|f_d\|_{L^2}^2,
\end{aligned} \tag{147}$$

where we used that for each $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$, we have $\mathbb{T}_{\boldsymbol{\tau}}^* \mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}} \mathbb{T}_{\boldsymbol{\tau}} = \mathbf{I}|_{V_{\mathcal{Q}}^{\boldsymbol{d}}}$.

Let us consider the third term in the expansion (146) around $\boldsymbol{a}^*$: the non diagonal term verifies

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\neq j}\mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i^*, \boldsymbol{x}^{(q)}\rangle\langle \boldsymbol{a}_j^*, \boldsymbol{x}^{(q)}\rangle\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)]\right]\\
&\quad= (1-N^{-1})\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0},\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2}\left[\boldsymbol{\alpha}_{\boldsymbol{\tau}^1}(\overline{\boldsymbol{\theta}}_1)^{\mathsf{T}}E_{\overline{\boldsymbol{x}}}\left[\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle\overline{\boldsymbol{\theta}}_1^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\right.\right.\\
&\qquad\left.\left.\times\,\sigma'_{\boldsymbol{d},\boldsymbol{\tau}'}\left(\{\langle\overline{\boldsymbol{\theta}}_2^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right)\boldsymbol{x}^{(q)}(\boldsymbol{x}^{(q)})^{\mathsf{T}}\right]\boldsymbol{\alpha}_{\boldsymbol{\tau}^2}(\overline{\boldsymbol{\theta}}_2)\right]\\
&\quad= (1-N^{-1})\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0},\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2}\left[\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\overline{f}_d(\overline{\boldsymbol{\theta}}_1)^{\mathsf{T}}\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}(\overline{\boldsymbol{\theta}}_1,\overline{\boldsymbol{\theta}}_2)\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}\overline{f}_d(\overline{\boldsymbol{\theta}}_2)\right]\\
&\quad= (1-N^{-1})\mathbb{E}_{\boldsymbol{\tau}^1_{\varepsilon_0},\boldsymbol{\tau}^2_{\varepsilon_0}}\left[\langle\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\overline{f}_d, \mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}\overline{f}_d\rangle_{L^2}\right].
\end{aligned}
$$

For $\boldsymbol{k}\in\mathcal{Q}$ and $\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]$ and $\boldsymbol{\tau}^1,\boldsymbol{\tau}^2\in[1-\varepsilon_0, 1+\varepsilon_0]^Q$, we have

$$
\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}} = \left(\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{T}_{\boldsymbol{\tau}^1}\right)\cdot\left(\mathbb{T}^*_{\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2}\right)\cdot Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}} = Y^{\boldsymbol{d}}_{\boldsymbol{k},\boldsymbol{s}}.
$$

Hence for any $\boldsymbol{\tau}^1,\boldsymbol{\tau}^2\in[1-\varepsilon_0, 1+\varepsilon_0]^Q$, $\mathbb{T}^*_{\boldsymbol{\tau}^1}\mathbb{K}^{-1}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^1}\mathbb{K}_{\boldsymbol{\tau}^1,\boldsymbol{\tau}^2}\mathbb{K}^{-1}_{\boldsymbol{\tau}^2,\boldsymbol{\tau}^2}\mathbb{T}_{\boldsymbol{\tau}^2} = \mathbf{I}|_{V^{\boldsymbol{d}}_{\mathcal{Q}}}$. Hence

$$
\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\neq j}\mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i^*, \boldsymbol{x}^{(q)}\rangle\langle \boldsymbol{a}_j^*, \boldsymbol{x}^{(q)}\rangle\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)]\right] = (1-N^{-1})\|f_d\|^2_{L^2}. \tag{148}
$$

The diagonal term verifies

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\in[N]}\mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i^*, \boldsymbol{x}^{(q)}\rangle^2\sigma'(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/R)\sigma'(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/R)]\right]\\
&\quad= N^{-1}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0},\overline{\boldsymbol{\theta}}}\left[\boldsymbol{\alpha}_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}\,)^{\mathsf{T}}\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\boldsymbol{\alpha}_{\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}}\,)\right]\\
&\quad\leqslant N^{-1}\left[\max_{\overline{\boldsymbol{\theta}},\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q}\|\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\|_{\mathrm{op}}\right]\cdot\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\|\mathbb{K}^{-1}_{\boldsymbol{\tau},\boldsymbol{\tau}}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|^2_{L^2}].
\end{aligned}
$$

We have, from lemma 14,

$$
\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}) = \alpha^{(q)}\mathbf{I}_{d_q} + \beta^{(q)}\overline{\boldsymbol{\theta}}^{(q)}(\overline{\boldsymbol{\theta}}^{(q)})^2
$$

where

$$
\begin{bmatrix}\alpha^{(q)}\\ \beta^{(q)}\end{bmatrix} = [d_q(d_q-1)(\tau^{(q)})^4]^{-1}\begin{bmatrix}d_q(\tau^{(q)})^4 & -(\tau^{(q)})^2\\ -(\tau^{(q)})^2 & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}_{\boldsymbol{x}}[\langle\boldsymbol{x}^{(q)},\boldsymbol{x}^{(q)}\rangle\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)})^2]\\ \mathbb{E}_{\boldsymbol{x}}[(\overline{x}_1^{(q)})^2\sigma'_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}_1^{(1)},\ldots,\overline{x}_1^{(Q)})^2]\end{bmatrix}.
$$

Hence from lemma 17 and for $\varepsilon_0$ small enough, there exists $C > 0$ such that for $d$ large enough

$$\sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q} \|\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\|_{\mathrm{op}} \leqslant C\frac{r_q^2}{d_q} = Cd^{\kappa_q}.$$

Furthermore

$$\begin{aligned}\|\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|_{L^2}^2 &= \sum_{\boldsymbol{k}\in\mathcal{Q}} \frac{B(\boldsymbol{d},\boldsymbol{k})}{A_{(\boldsymbol{\tau},\boldsymbol{\tau}),\boldsymbol{k}}^{(q)}} \sum_{\boldsymbol{s}\in[B(\boldsymbol{d},\boldsymbol{k})]} \lambda_{\boldsymbol{k},\boldsymbol{s}}^{\boldsymbol{d}}(\overline{f}_d)^2 \\ &\leqslant \left[\max_{\boldsymbol{k}\in\mathcal{Q}} \frac{B(\boldsymbol{d},\boldsymbol{k})}{A_{(\boldsymbol{\tau},\boldsymbol{\tau}),\boldsymbol{k}}^{(q)}}\right] \cdot \|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}^2.\end{aligned}$$

From lemma 15, we get

$$\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0}}[\|\mathbb{K}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1}\mathbb{T}_{\boldsymbol{\tau}}\overline{f}_d\|_{L^2}^2] \leqslant Cd^{\gamma-\kappa_q} \cdot \|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}^2.$$

Hence,

$$\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\sum_{i\in[N]} \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{a}_i^*, \boldsymbol{x}^{(q)}\rangle^2 \sigma'(\langle\boldsymbol{\theta}_i,\boldsymbol{x}\rangle/R)\sigma'(\langle\boldsymbol{\theta}_j,\boldsymbol{x}\rangle/R)]\right] \leqslant C\frac{d^\gamma}{N}\|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}^2. \qquad (149)$$

Combining equations (147)–(149), we get

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\Theta}_{\varepsilon_0}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(f_d(\boldsymbol{x}) - \hat{f}(\boldsymbol{x};\boldsymbol{\Theta},\boldsymbol{a}^*(\boldsymbol{\Theta})))^2\right]\right] \\ &\quad= \|f_d\|_{L^2}^2 - 2\|f_d\|_{L^2}^2 + (1-N^{-1})\|f_d\|_{L^2}^2 + N^{-1}\mathbb{E}_{\boldsymbol{\tau}_{\varepsilon_0},\overline{\boldsymbol{\theta}}}\left[\boldsymbol{\alpha}_\tau(\overline{\boldsymbol{\theta}})^{\mathsf{T}}\mathbb{K}_{\tau,\tau}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}})\boldsymbol{\alpha}_\tau(\overline{\boldsymbol{\theta}})\right] \\ &\quad\leqslant C\frac{d^\gamma}{N}\|\mathsf{P}_{\mathcal{Q}}f_d\|_{L^2}^2.\end{aligned}$$

By Markov's inequality, we get for any $\varepsilon > 0$ and $d$ large enough,

$$\begin{aligned}\mathbb{P}(R_{\mathrm{NT}^{(q)}}(f_d,\boldsymbol{\Theta}) > \varepsilon\cdot\|f_d\|_{L^2}^2) &\leqslant \mathbb{P}(\{R_{\mathrm{NT}^{(q)}}(f_d,\boldsymbol{\Theta}) > \varepsilon\cdot\|f_d\|_{L^2}^2\}\cap\mathcal{P}_{\varepsilon_0}) + \mathbb{P}(\mathcal{P}_{\varepsilon_0}^c) \\ &\leqslant C'\frac{d^\gamma}{N} + \mathbb{P}(\mathcal{P}_{\varepsilon_0}^c).\end{aligned}$$

The assumption that $N = \omega_d(d^\gamma)$ and lemma 8 conclude the proof.

## Appendix I. Proof of theorem 4 in the main text

### Step 1. Show that $R_{\mathrm{NN},2N}(f_*) \leqslant \inf_{\boldsymbol{W}\in\mathbb{R}^{N\times d}} R_{\mathrm{NT},N}(f_*,\boldsymbol{W})$.

Define the neural tangent model with $N$ neurons by $\hat{f}_{\mathrm{NT},N}(\boldsymbol{x};\boldsymbol{s};\boldsymbol{W}) = \sum_{i=1}^N \langle\boldsymbol{s}_i,\boldsymbol{x}\rangle\sigma'(\langle\boldsymbol{w}_i,\boldsymbol{x}\rangle)$ and the NN with $N$ neurons by $\hat{f}_{\mathrm{NN},N}(\boldsymbol{x};\boldsymbol{W},\boldsymbol{b}) = \sum_{i=1}^N b_i\sigma(\langle\boldsymbol{w}_i,\boldsymbol{x}\rangle)$.

For any $\boldsymbol{W} \in \mathbb{R}^{N\times d}$, $\boldsymbol{s} \in \mathbb{R}^N$, and $\varepsilon > 0$, we define

$$\hat{g}_N(\boldsymbol{x};\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon) \equiv \varepsilon^{-1}\left(\hat{f}_{\mathrm{NN},N}(\boldsymbol{x};\, \boldsymbol{W} + \varepsilon \boldsymbol{s}, \boldsymbol{1}) - \hat{f}_{\mathrm{NN},N}(\boldsymbol{x};\, \boldsymbol{W}, \boldsymbol{1})\right),$$

$$\mathcal{E}(\boldsymbol{x};\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon) = \hat{g}_N(\boldsymbol{x};\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon) - \hat{f}_{\mathrm{NT},N}(\boldsymbol{x};\, \boldsymbol{s};\, \boldsymbol{W}).$$

Then by Taylor expansion, there exists $(\tilde{\boldsymbol{w}}_i)_{i\in[N]}$ such that

$$|\mathcal{E}(\boldsymbol{x};\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon)| = \frac{\varepsilon}{2}\left|\sum_{i=1}^{N} \langle \boldsymbol{s}_i, \boldsymbol{x}\rangle^2 \sigma''(\langle \tilde{\boldsymbol{w}}_i, \boldsymbol{x}\rangle)\right|.$$

By the boundedness assumption of $\sup_{x\in\mathbb{R}}|\sigma''(x)|$, we have

$$\lim_{\varepsilon\to 0+} \|\mathcal{E}(\cdot;\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon)\|_{L^2}^2 = 0,$$

and hence

$$\lim_{\varepsilon\to 0+} \|f_* - \hat{g}_N(\cdot;\, \boldsymbol{W}, \boldsymbol{s}, \varepsilon)\|_{L^2}^2 = \|f_* - \hat{f}_{\mathrm{NT},N}(\cdot;\, \boldsymbol{s};\, \boldsymbol{W})\|_{L^2}^2.$$

Note that $\hat{g}_N$ can be regarded as a function in $\mathcal{F}_{\mathrm{NN}}^{2N}$ and $\hat{f}_{\mathrm{NT},N} \in \mathcal{F}_{\mathrm{NN}}^{N}(\boldsymbol{W})$, this implies that

$$R_{\mathrm{NN},2N}(f_*) \leqslant \inf_{\boldsymbol{W}\in\mathbb{R}^{N\times d}} R_{\mathrm{NT},N}(f_*, \boldsymbol{W}). \tag{150}$$

**Step 2. Give upper bound of $\inf_{\boldsymbol{W}\in\mathbb{R}^{N\times d}} R_{\mathrm{NT},N}(f_*, \boldsymbol{W})$.** We take $\overline{\boldsymbol{W}} = (\bar{\boldsymbol{w}}_i)_{i\leqslant N}$ with $\bar{\boldsymbol{w}}_i = \boldsymbol{U}\bar{\boldsymbol{v}}_i$, where $\bar{\boldsymbol{v}}_i \sim \mathrm{Unif}(\mathbb{S}^{d_0-1}(r^{-1}))$, and denote $\overline{\boldsymbol{V}} = (\bar{\boldsymbol{v}}_i)_{i\leqslant N}$. Then we have

$$\mathcal{G}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}}) \equiv \left\{ f(\boldsymbol{x}) = \bar{f}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}) : \bar{f}(\boldsymbol{z}) = \sum_{i=1}^{N} \langle \bar{\boldsymbol{s}}_i, \boldsymbol{z}\rangle \sigma'(\langle \bar{\boldsymbol{v}}_i, \boldsymbol{z}\rangle), \bar{\boldsymbol{s}}_i \in \mathbb{R}^{d_0}, i \leqslant N \right\} \subseteq \mathcal{F}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{W}}).$$

It is easy to see that, when $f_*(\boldsymbol{x}) = \varphi(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{x})$, we have

$$\inf_{\hat{f}\in\mathcal{G}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}})} \mathbb{E}[(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2] = \inf_{\hat{f}\in\mathcal{F}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}})} \mathbb{E}[(\varphi(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}))^2],$$

where $\mathcal{F}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}})$ is the class of neural tangent model on $\mathbb{R}^{d_0}$

$$\mathcal{F}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}}) = \left\{ \bar{f}(\boldsymbol{z}) = \sum_{i=1}^{N} \langle \bar{\boldsymbol{s}}_i, \boldsymbol{z}\rangle \sigma'(\langle \bar{\boldsymbol{v}}_i, \boldsymbol{z}\rangle) : \bar{\boldsymbol{s}}_i \in \mathbb{R}^{d_0}, i \leqslant N \right\}.$$

Moreover, by theorem 3 in the main text, when $d_0^{\ell+\delta} \leqslant N \leqslant d_0^{\ell+1-\delta}$ for some $\delta > 0$ independent of $N, d$, we have

$$\inf_{\hat{f}\in\mathcal{F}_{\mathrm{NT}}^{N}(\overline{\boldsymbol{V}})} \mathbb{E}[(\varphi(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}))^2] = (1 + o_{d,\mathbb{P}}(1)) \cdot \|\mathsf{P}_{>\ell+1}\varphi\|_{L^2}^2 = (1 + o_{d,\mathbb{P}}(1)) \cdot \|\mathsf{P}_{>\ell+1} f_*\|_{L^2}^2.$$

As a consequence, we have

$$\begin{aligned}\inf_{\boldsymbol{W}\in\mathbb{R}^{N\times d}} R_{\mathrm{NT},N}(f_*,\boldsymbol{W}) &\leqslant \inf_{\hat{f}\in\mathcal{F}^N_{\mathrm{NT}}(\overline{\boldsymbol{W}})} \mathbb{E}[(f_*(\boldsymbol{x})-\hat{f}(\boldsymbol{x}))^2] \leqslant \inf_{\hat{f}\in\mathcal{G}^N_{\mathrm{NT}}(\overline{\boldsymbol{V}})} \mathbb{E}[(f_*(\boldsymbol{x})-\hat{f}(\boldsymbol{x}))^2] \\ &= \inf_{\hat{f}\in\mathcal{F}^N_{\mathrm{NT}}(\overline{\boldsymbol{V}})} \mathbb{E}[(\varphi(\boldsymbol{z})-\hat{f}(\boldsymbol{z}))^2] = (1+o_{d,\mathbb{P}}(1))\cdot\|\mathsf{P}_{>\ell+1}f_*\|^2_{L^2}.\end{aligned}$$

Combining with equation (150) gives that, when $d_0^{\ell+\delta}\leqslant N\leqslant d_0^{\ell+1-\delta}$, we have

$$R_{\mathrm{NN},N}(f_*)\leqslant(1+o_d(1))\cdot\|\mathsf{P}_{>\ell+1}f_*\|^2_{L^2}.$$

**Step 3. Show that $R_{\mathrm{NN},N}(f_*)$ is independent of $\kappa$.**

We let $\tilde{r}=d^{\tilde{\kappa}/2}$ and $\mathring{r}=d^{\mathring{r}/2}$ for some $\tilde{\kappa}\neq\mathring{\kappa}$. Suppose we have $\tilde{\boldsymbol{x}}=\boldsymbol{U}\tilde{\boldsymbol{z}}_1+\boldsymbol{U}^{\perp}\boldsymbol{z}_2$ and $\mathring{\boldsymbol{x}}=\boldsymbol{U}\mathring{\boldsymbol{z}}_1+\boldsymbol{U}^{\perp}\boldsymbol{z}_2$, where $\tilde{\boldsymbol{z}}_1\sim\mathrm{Unif}(\mathbb{S}^{d_0-1}(\tilde{r}\sqrt{d_0}))$, $\mathring{\boldsymbol{z}}_1\sim\mathrm{Unif}(\mathbb{S}^{d_0-1}(\mathring{r}\sqrt{d_0}))$, and $\boldsymbol{z}_2\sim\mathrm{Unif}(\mathbb{S}^{d-d_0-1}(\sqrt{d-d_0}))$. Moreover, we let $\tilde{f}_*(\tilde{\boldsymbol{x}})=\varphi(\boldsymbol{U}^{\mathsf{T}}\tilde{\boldsymbol{x}}/\tilde{r})$ and $\mathring{f}_*(\mathring{\boldsymbol{x}})=\varphi(\boldsymbol{U}^{\mathsf{T}}\mathring{\boldsymbol{x}}/\mathring{r})$ for some function $\varphi:\mathbb{R}^{d_0}\to\mathbb{R}$.

Then, for any $\tilde{\boldsymbol{W}}=(\tilde{\boldsymbol{w}}_i)_{i\leqslant N}\subseteq\mathbb{R}^d$ and $\tilde{\boldsymbol{b}}=(\tilde{b}_i)_{i\leqslant N}\subseteq\mathbb{R}$, there exists $(\tilde{\boldsymbol{v}}_{1,i})_{i\leqslant N}\subseteq\mathbb{R}^{d_0}$ and $(\tilde{\boldsymbol{v}}_{2,i})_{i\leqslant N}\subseteq\mathbb{R}^{d-d_0}$ such that $\tilde{\boldsymbol{w}}_i=\boldsymbol{U}\tilde{\boldsymbol{v}}_{1,i}+\boldsymbol{U}^{\perp}\tilde{\boldsymbol{v}}_{2,i}$. We define $\mathring{\boldsymbol{v}}_{1,i}=\tilde{r}\cdot\tilde{\boldsymbol{v}}_{1,i}/\mathring{r}$, $\mathring{\boldsymbol{w}}_i=\boldsymbol{U}\mathring{\boldsymbol{v}}_{1,i}+\boldsymbol{U}^{\perp}\tilde{\boldsymbol{v}}_{2,i}$, $\mathring{\boldsymbol{W}}=(\mathring{\boldsymbol{w}}_i)_{i\leqslant N}$, and $\mathring{\boldsymbol{b}}=\tilde{\boldsymbol{b}}$. Then we have

$$\mathbb{E}_{\mathring{\boldsymbol{x}}}[(\mathring{f}_*(\mathring{\boldsymbol{x}})-f_{\mathrm{NN},N}(\mathring{\boldsymbol{x}};\mathring{\boldsymbol{W}},\mathring{\boldsymbol{b}}))^2]=\mathbb{E}_{\tilde{\boldsymbol{x}}}[(\tilde{f}_*(\tilde{\boldsymbol{x}})-f_{\mathrm{NN},N}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{W}},\tilde{\boldsymbol{b}}))^2].$$

On the other hand, for any $\mathring{\boldsymbol{W}}=(\mathring{\boldsymbol{w}}_i)_{i\leqslant N}\subseteq\mathbb{R}^d$ and $\mathring{\boldsymbol{b}}=(\mathring{b}_i)_{i\leqslant N}\subseteq\mathbb{R}$, we can find $\tilde{\boldsymbol{W}}=(\tilde{\boldsymbol{w}}_i)_{i\leqslant N}\subseteq\mathbb{R}^d$ and $\tilde{\boldsymbol{b}}=(\tilde{b}_i)_{i\leqslant N}\subseteq\mathbb{R}$ such that the above equation holds. This proves that $R_{\mathrm{NN},N}(f_*)$ is independent of $\kappa$.

## Appendix J. Convergence of the Gegenbauer coefficients

In this section, we prove a string of lemmas that are used to show convergence of the Gegenbauer coefficients.

### J.1. Technical lemmas

First recall that for $q\in[Q]$ we denote $\tau^{(q)}\equiv\|\boldsymbol{\theta}^{(q)}\|_2/\sqrt{d_q}$ where $\boldsymbol{\theta}^{(q)}$ are the $d_q$ coordinates of $\boldsymbol{\theta}\sim\mathrm{Unif}(\mathbb{S}^{D-1}(\sqrt{D}))$ associated to the $q$th sphere of $\mathrm{PS}^{\boldsymbol{d}}$. We show that $\tau^{(q)}$ is $(1/d_q)$-sub-Gaussian.

**Lemma 16.** *There exists constants $c,C>0$ such that for any $\varepsilon>0$,*

$$\mathbb{P}(|\tau^{(q)}-1|>\varepsilon)\leqslant C\,\exp(-cd_q\varepsilon^2).$$

**Proof of lemma 16.** Let $\boldsymbol{G}\sim N(0,\mathbf{I}_D)$. We consider the random vector $\boldsymbol{U}\equiv\boldsymbol{G}/\|\boldsymbol{G}\|_2\in\mathbb{R}^D$. We have $\boldsymbol{U}\sim\mathrm{Unif}(\mathbb{S}^{D-1}(1))$. We denote $N_{d_q}=G_1^2+\cdots+G_{d_q}^2$ and $N_D=$

$G_1^2 + \cdots + G_D^2$. The random variable $\tau^{(q)}$ has the same distribution as

$$\tau^{(q)} \equiv \|\boldsymbol{\theta}^{(q)}\|_2/\sqrt{d_q} \overset{\mathrm{d}}{=} \frac{\sqrt{N_{d_q}/d_q}}{\sqrt{N_D/D}}.$$

Hence,

$$\begin{aligned}\mathbb{P}(|\tau^{(q)} - 1| > \varepsilon) &= \mathbb{P}\left(\left|\frac{\sqrt{N_{d_q}/d_q}}{\sqrt{N_D/D}} - 1\right| > \varepsilon\right) \\ &\leqslant \mathbb{P}\left(\left|\sqrt{N_{d_q}/d_q} - 1\right| > \varepsilon/2\right) + \mathbb{P}\left(\left|\sqrt{N_D/D} - 1\right| > \varepsilon/(2+2\varepsilon)\right),\end{aligned} \tag{151}$$

where we used the fact that

$$|a-1| \leqslant \frac{\varepsilon}{2} \quad \text{and} \quad |b-1| \leqslant \frac{\varepsilon}{2+2\varepsilon} \Rightarrow \left|\frac{a}{b} - 1\right| \leqslant \varepsilon.$$

Let us first consider $N_{d_q}$ with $\varepsilon \in (0, 2]$. The $G_i^2$ are sub-exponential random variables with

$$\mathbb{E}\left[\mathrm{e}^{\lambda(G_i^2 - 1)}\right] \leqslant \mathrm{e}^{2\lambda^2}, \quad \forall\ |\lambda| < 1/4.$$

From standard sub-exponential concentration inequality, we get

$$\mathbb{P}\left(\left|N_{d_q}/d_q - 1\right| > \varepsilon\right) \leqslant 2\ \exp\left(-d_q\varepsilon\ \min(1, \varepsilon)/8\right). \tag{152}$$

Hence, for $\varepsilon \in (0, 2]$, we have

$$\mathbb{P}\left(\left|\sqrt{N_{d_q}/d_q} - 1\right| > \varepsilon/2\right) \leqslant \mathbb{P}\left(\left|N_{d_q}/d_q - 1\right| > \varepsilon/2\right) \leqslant 2\ \exp\left(-d_q\varepsilon^2/32\right),$$

while for $\varepsilon > 2$,

$$\begin{aligned}\mathbb{P}\left(\left|\sqrt{N_{d_q}/d_q} - 1\right| > \varepsilon/2\right) \leqslant \mathbb{P}\left(N_{d_q}/d_q > (\varepsilon/2 + 1)^2\right) \leqslant \mathbb{P}\left(N_{d_q}/d_q - 1 > \varepsilon^2/4\right) \\ \leqslant \exp\left(-d_q\varepsilon^2/32\right).\end{aligned}$$

In the case of $N_D$, applying (152) with $\varepsilon/(2+2\varepsilon) \leqslant 1$ shows that

$$\begin{aligned}\mathbb{P}\left(\left|\sqrt{N_D/D} - 1\right| > \varepsilon/(2+2\varepsilon)\right) &\leqslant \mathbb{P}\left(|N_D/D - 1| > \varepsilon/(2+2\varepsilon)\right) \\ &\leqslant 2\ \exp\left(-D\varepsilon^2/(32(1+\varepsilon)^2)\right).\end{aligned}$$

Combining the above bounds into (151) yields for $\varepsilon \geqslant 0$,

$$\begin{aligned}\mathbb{P}(|\tau^{(q)} - 1| > \varepsilon) &\leqslant 2\ \exp\left(-d_q\varepsilon^2/32\right) + 2\ \exp\left(-D\varepsilon^2/(32(1+\varepsilon)^2)\right) \\ &\leqslant 4\ \exp\left(-\varepsilon^2\ \min\left(d_q, D/(1+\varepsilon)^2\right)/32\right).\end{aligned}$$

Notice that $|\tau^{(q)} - 1| \leqslant \sqrt{D/d_q} - 1$ and we only need to consider $\varepsilon \in [0, \sqrt{D/d_q} - 1]$. We conclude that for any $\varepsilon \geqslant 0$, we have

$$\mathbb{P}(|\tau^{(q)} - 1| > \varepsilon) \leqslant 4\,\exp\left(-d_q\varepsilon^2/32\right).$$

□

We consider an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Fix $\boldsymbol{\theta} \in \mathbb{S}^{D-1}(\sqrt{D})$ and recall that $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(Q)}) \in \mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}$. We recall that $\boldsymbol{x} \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}}_{\boldsymbol{\kappa}}) = \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}}$ while $\overline{\boldsymbol{x}} \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}}) = \mu_{\boldsymbol{d}}$. Therefore, for a given $\overline{\boldsymbol{\theta}}$, $\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]} \sim \tilde{\mu}^1_{\boldsymbol{d}}$ as defined in equation (33). Therefore we reformulate $\sigma(\langle \boldsymbol{\theta}, \cdot\rangle/R)$ as a function $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ from $\mathrm{ps}^{\boldsymbol{d}}$ to $\mathbb{R}$:

$$\begin{aligned}\sigma(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle/R) &= \sigma\left(\sum_{q\in[Q]} \tau^{(q)} \cdot (r_q/R) \cdot \langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\right)\\ &\equiv \sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \overline{\boldsymbol{\theta}}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle/\sqrt{d_q}\}_{q\in[Q]}\right).\end{aligned} \tag{153}$$

We will denote in the rest of this section $\alpha_q = \tau^{(q)} r_q/R$ for $q = 1, \ldots, Q$. Notice in particular that $\alpha_q \propto d^{\eta_q + \kappa_q - \xi}$ where we recall that $\xi = \max_{q\in[Q]}\{\eta_q + \kappa_q\}$. Without loss of generality, we will assume that the (unique) maximum is attained on the first sphere, i.e. $\xi = \eta_1 + \kappa_1$ and $\xi > \eta_q + \kappa_q$ for $q \geqslant 2$.

**Lemma 17.** *Assume $\sigma$ is an activation function with $\sigma(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ almost surely, for some constants $c_0 > 1$ and $c_1 < 1$. We consider the function $\sigma_{\boldsymbol{d},\boldsymbol{\tau}} : \mathrm{ps}^{\boldsymbol{d}} \to \mathbb{R}$ associated to $\sigma$, as defined in equation (75).*

*Then*

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

*(b) Let $\boldsymbol{w}^{(q)}$ be unit vectors in $\mathbb{R}^{d_q}$ for $q = 1, \ldots, Q$. There exists $\varepsilon_0 = \varepsilon_0(c_1)$ and $d_0 = d_0(c_1)$ such that, for $\overline{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(Q)}) \sim \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}}$,*

$$\sup_{d\geqslant d_0} \sup_{\boldsymbol{\tau}\in[1-\varepsilon_0, 1+\varepsilon_0]^Q} \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] < \infty. \tag{154}$$

*(c) Let $\boldsymbol{w}^{(q)}$ be unit vectors in $\mathbb{R}^{d_q}$ for $q = 1, \ldots, Q$. Fix integers $\boldsymbol{k} = (k_1, \ldots, k_Q) \in \mathbb{Z}^Q_{\geqslant 0}$. Then for any $\delta > 0$, there exists constants $\varepsilon_0 = \varepsilon_0(c_1, \delta)$ and $d_0 = d_0(c_1, \delta)$, and a coupling of $G \sim N(0, 1)$ and $\overline{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(Q)}) \sim \mu^{\boldsymbol{\kappa}}_{\boldsymbol{d}}$ such that for any $d \geqslant d_0$ and $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$*

$$\mathbb{E}_{\overline{\boldsymbol{x}},G}\left[\left(\left[\prod_{q\in[Q]}\left(1 - \langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle^2/d_q\right)^{k_q}\right]\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right) - \sigma(G)\right)^2\right] < \delta. \tag{155}$$

**Proof of lemma 17.** Part (a) is straightforward.

For part (b), recall that the probability distribution of $\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle$ when $\overline{\boldsymbol{x}}^{(q)} \sim \mathrm{Unif}(\mathbb{S}^{d_q-1}(\sqrt{d_q}))$ is given by

$$\tilde{\tau}^1_{d_q-1}(\mathrm{d}x) = C_{d_q}\left(1-\frac{x^2}{d_q}\right)^{\frac{d_q-3}{2}} \mathbf{1}_{x\in[-\sqrt{d_q},\sqrt{d_q}]}\mathrm{d}x, \tag{156}$$

$$C_{d_q} = \frac{\Gamma(d_q-1)}{2^{d_q-2}\sqrt{d_q}\,\Gamma((d_q-1)/2)^2}. \tag{157}$$

A simple calculation shows that $C_n \to (2\pi)^{-1/2}$ as $n\to\infty$, and hence $\sup_n C_n \leqslant \overline{C} < \infty$. Therefore for $\boldsymbol{\tau} \in [1-\varepsilon, 1+\varepsilon]^Q$, we have

$$\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] \\
&\quad= \int_{\prod_{q\in[Q]}[-\sqrt{d_q},\sqrt{d_q}]} \sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)}\right)^2 \prod_{q\in[Q]}\left(C_{d_q}\left(1-\frac{(\overline{x}_1^{(q)})^2}{d_q}\right)^{\frac{d_q-3}{2}}\mathrm{d}\overline{x}_1^{(q)}\right) \\
&\quad\leqslant \overline{C}^Q \int_{\mathbb{R}^Q} c_0 \exp\left(c_1\left(\sum_{q\in[Q]}\alpha_q \overline{x}_1^{(q)}\right)^2/2\right)\prod_{q\in[Q]}\left(\exp\left(-\frac{d_q-3}{2d_q}(\overline{x}_1^{(q)})^2\right)\mathrm{d}\overline{x}_1^{(q)}\right) \\
&\quad= c_0\overline{C}^Q\int_{\mathbb{R}^Q}\exp\left(-\overline{\boldsymbol{x}}_1^{\mathsf{T}}\boldsymbol{M}\overline{\boldsymbol{x}}_1/2\right)\left(\prod_{q\in[Q]}\mathrm{d}\overline{x}_1^{(q)}\right)
\end{aligned}$$

where we denoted $\overline{\boldsymbol{x}}_1 = (\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)})$ and $\boldsymbol{M}\in\mathbb{R}^{Q\times Q}$ with

$$M_{qq} = \frac{d_q-3}{d_q} - c_1^2\alpha_q^2, \qquad M_{qq'} = -c_1\alpha_q\alpha_{q'}, \quad \text{for } q\neq q'\in[Q].$$

Recalling the definition of $\alpha_q = \tau^{(q)}r_q/R$, with $r_q = d^{(\eta_q+\kappa_q)/2}$ and $R = d^{\xi/2}(1+o_d(1))$. Hence for any $\varepsilon>0$, uniformly on $\boldsymbol{\tau}\in[1-\varepsilon,1+\varepsilon]^Q$, we have $\alpha_q\to 0$ for $q\geqslant 2$ and $\limsup_{d\to\infty}|\alpha_1-1|\leqslant\varepsilon$. Hence if we choose $\varepsilon_0 < c_1^{-1}-1$, there exists $c>0$ such that for $d$ sufficiently large $\boldsymbol{M}\succeq c\mathbf{I}_Q$ and for any $\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q$

$$\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] \leqslant c_0\overline{C}^Q\int_{\mathbb{R}^Q}\exp\left(-c\|\overline{\boldsymbol{x}}_1\|_2^2/2\right)\left(\prod_{q\in[Q]}\mathrm{d}\overline{x}_1^{(q)}\right) < \infty.$$

Finally, for part (c), without loss of generality we will take $\boldsymbol{w}^{(q)} = \boldsymbol{e}_1^{(q)}$ so that $\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle = \overline{x}_1^{(q)}$. From part (b), there exists $\varepsilon>0$ and $d_0$ such that

$$\begin{aligned}
&\sup_{d\geqslant d_0}\sup_{\tau\in[1-\varepsilon,1+\varepsilon]}\mathbb{E}_{\overline{\boldsymbol{x}}}\mathbb{E}_{\overline{\boldsymbol{x}},G}\left[\left[\prod_{q\in[Q]}\left(1-\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle^2/d_q\right)^{2k_q}\right]\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] \\
&\quad\leqslant \sup_{d\geqslant d_0}\sup_{\tau\in[1-\varepsilon,1+\varepsilon]}\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\{\langle \boldsymbol{w}^{(q)}, \overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right)^2\right] < \infty.
\end{aligned}$$

Consider $\boldsymbol{G} \sim \mathsf{N}(0, \mathbf{I}_Q)$ and an arbitrary coupling between $\overline{\boldsymbol{x}}$ and $\boldsymbol{G}$. For any $M > 0$ we can choose $\sigma_M$ bounded continuous so that for any $d$ and $\boldsymbol{\tau} \in [1-\varepsilon, 1+\varepsilon]^Q$,

$$\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{x}},\boldsymbol{G}}\left[\left(\prod_{q\in[Q]}\left(1-(\overline{x}_1^{(q)})^2/d_q\right)^{k_q}\cdot\sigma\left(\sum_{q\in[Q]}\alpha_q\overline{x}_1^{(q)}\right)\right.\right.\\
&\qquad\left.\left.-\prod_{q\in[Q]}\left(1-G_q^2/d_q\right)^{k_q}\cdot\sigma\left(\sum_{q\in[Q]}\alpha_qG_q\right)\right)^2\right]\\
&\quad\leqslant \mathbb{E}_{\overline{\boldsymbol{x}},\boldsymbol{G}}\left[\left(\prod_{q\in[Q]}\left(1-(\overline{x}_1^{(q)})^2/d_q\right)^{k_q}\cdot\sigma_M\left(\sum_{q\in[Q]}\alpha_q\overline{x}_1^{(q)}\right)\right.\right.\\
&\qquad\left.\left.-\prod_{q\in[Q]}\left(1-G_q^2/d_q\right)^{k_q}\cdot\sigma_M\left(\sum_{q\in[Q]}\alpha_qG_q\right)\right)^2\right]+\frac{1}{M}.
\end{aligned} \tag{158}$$

It is therefore sufficient to prove the claim for $\sigma_M$. Letting $\boldsymbol{\xi}_q \sim \mathsf{N}(0, \mathbf{I}_{d_q-1})$ independently for each $q \in [Q]$ and independent of $\boldsymbol{G}$, we construct the coupling via

$$\overline{x}_1^{(q)} = \frac{G_q\sqrt{d_q}}{\sqrt{G_q^2+\|\boldsymbol{\xi}_q\|_2^2}}, \qquad \overline{\boldsymbol{x}}_{-1}^{(q)} = \frac{\boldsymbol{\xi}_q\sqrt{d_q}}{\sqrt{G_q^2+\|\boldsymbol{\xi}_q\|_2^2}}, \qquad q \in [Q], \tag{159}$$

where we set $\overline{\boldsymbol{x}}^{(q)} = (\overline{x}_1^{(q)}, \overline{\boldsymbol{x}}_{-1}^{(q)})$ for each $q \in [Q]$. We thus have $(\overline{x}_1^{(q)}, \overline{\boldsymbol{x}}_{-1}^{(q)}) \to \boldsymbol{G}$ almost surely, hence the limit superior of equation (158) is by weak convergence bounded by $1/M$ for any arbitrary $M$. Furthermore, noticing that $\alpha_q \to 0$ uniformly on $\boldsymbol{\tau} \in [1-\varepsilon, 1+\varepsilon]^Q$ for $q \geqslant 2$, we have by bounded convergence

$$\lim_{d\to\infty}\sup_{\boldsymbol{\tau}\in[1-\varepsilon,1+\varepsilon]^Q}\mathbb{E}_{\overline{\boldsymbol{x}},\boldsymbol{G}}\left[\left(\prod_{q\in[Q]}\left(1-G_q^2/d_q\right)^{k_q}\cdot\sigma\left(\sum_{q\in[Q]}\alpha_qG_q\right)-\sigma\left(\alpha_1G_1\right)\right)^2\right]=0. \tag{160}$$

We further have $\lim_{(d,\tau^{(1)})\to(\infty,1)}\alpha_1 = 1$. Hence, by bounded convergence,

$$\lim_{(d,\tau^{(1)})\to(\infty,1)}\mathbb{E}_{G_1}\left[\left(\sigma\left(\alpha_1G_1\right)-\sigma(G_1)\right)^2\right]=0. \tag{161}$$

Combining equation (158) with the coupling (159) and equations (160) and (161) yields the result. □

Consider the expansion of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$ in terms of tensor product of Gegenbauer polynomials. We have

$$\sigma(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/R)=\sum_{\boldsymbol{k}\in\mathbb{Z}_{\geqslant 0}^Q}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})B(\boldsymbol{d},\boldsymbol{k})Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\{\langle\overline{\boldsymbol{\theta}}^{(q)},\overline{\boldsymbol{x}}^{(q)}\rangle\}_{q\in[Q]}\right),$$

where

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)}\right) Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\dots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right],$$

with the expectation taken over $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}^{(1)},\dots,\overline{\boldsymbol{x}}^{(Q)}) \sim \mu_{\boldsymbol{d}} \equiv \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$. We will need the following lemma, which is direct consequence of Rodrigues formula, to get the scaling of the Gegenbauer coefficients of $\sigma_{\boldsymbol{d},\boldsymbol{\tau}}$.

**Lemma 18.** *Let $\boldsymbol{k} = (k_1,\dots,k_Q) \in \mathbb{Z}_{\geqslant 0}^Q$ and denote $|\mathbf{k}| = k_1 + \cdots + k_Q$. Assume that the activation function $\sigma$ is $|\mathbf{k}|$-times weakly differentiable and denote $\sigma^{(|\mathbf{k}|)}$ its $|\mathbf{k}|$-weak derivative. Let $\alpha_q = \tau^{(q)} r_q/R$ for $q = 1,\dots,Q$. Then*

$$\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \left(\prod_{q\in[Q]} \alpha_q^{k_q}\right) \cdot R(\boldsymbol{d},\boldsymbol{k}) \cdot \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\left(\prod_{q\in[Q]}\left(1 - \frac{(\overline{x}_1^{(q)})^2}{d_q}\right)^{k_q}\right) \cdot \sigma^{(|\boldsymbol{k}|)}\left(\sum_{q\in[Q]} \alpha_q \overline{x}_1^{(q)}\right)\right], \tag{162}$$

*where $\overline{\boldsymbol{x}} \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$ and*

$$R(\boldsymbol{d},\boldsymbol{k}) = \prod_{q\in[Q]} \frac{d_q^{k_q/2}\Gamma((d_q-1)/2)}{2^{k_q}\Gamma(k_q + (d_q-1)/2)}.$$

*Furthermore,*

$$\lim_{d\to\infty} B(\boldsymbol{d},\boldsymbol{k}) R(\boldsymbol{d},\boldsymbol{k})^2 = \frac{1}{\boldsymbol{k}!}, \tag{163}$$

*where $\mathbf{k}! = k_1!\dots k_Q!$.*

**Proof of lemma 18.** We have

$$\begin{aligned}\lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) &= \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}\left(\overline{x}_1^{(1)},\dots,\overline{x}_1^{(Q)}\right) Q_{\boldsymbol{k}}^{\boldsymbol{d}}\left(\sqrt{d_1}\overline{x}_1^{(1)},\dots,\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right] \\ &= \mathbb{E}_{\overline{\boldsymbol{x}}^{(1)},\dots,\overline{\boldsymbol{x}}^{(Q-1)}}\left[\mathbb{E}_{\overline{\boldsymbol{x}}^{(Q)}}\left[\sigma\left(\sum_{q\in[Q-1]} \alpha_q \overline{x}_1^{(q)} + \alpha_Q \overline{x}_1^{(Q)}\right) Q_{k_Q}^{(d_Q)}(\sqrt{d_Q}\overline{x}_1^{(Q)})\right]\right. \\ &\qquad \left.\times \prod_{q\in[Q-1]} Q_{k_q}^{(d_q)}(\sqrt{d_q}\overline{x}_1^{(q)})\right],\end{aligned} \tag{164}$$

where we used the definition (34) of tensor product of Gegenbauer polynomials.

Consider the integration with respect to $\overline{\boldsymbol{x}}^{(Q)}$. Denote for ease of notations $u = \alpha_1\overline{x}_1^{(1)} + \cdots + \alpha_{Q-1}\overline{x}_1^{(Q-1)}$. We use the Rodrigues formula for the Gegenbauer polynomials (see equation (24)):

$$
\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{x}}^{(Q)}\sim \mathrm{Unif}(\mathbb{S}^{d_Q-1}(\sqrt{d_Q}))}\left[\sigma\left(u+\alpha_Q\overline{x}_1^{(Q)}\right)Q_{k_Q}^{(d_Q)}\left(\sqrt{d_Q}\overline{x}_1^{(Q)}\right)\right]\\
&\quad=\frac{\omega_{d_Q-2}}{\omega_{d_Q-1}}\int_{[-1,1]}\sigma\left(\alpha_Q\sqrt{d_Q}t+u\right)Q_{k_Q}^{(d_Q)}(d_Qt)(1-t^2)^{(d_Q-3)/2}\mathrm{d}t\\
&\quad=(-1/2)^{k_Q}\frac{\Gamma((d_Q-1)/2)}{\Gamma(k_Q+(d_Q-1)/2)}\cdot\frac{\omega_{d_Q-2}}{\omega_{d_Q-1}}\\
&\qquad\times\int_{[-1,1]}\sigma\left(\alpha_Q\sqrt{d_Q}t+u\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k_Q}(1-t^2)^{k_Q+(d_Q-3)/2}\mathrm{d}t\\
&\quad=\alpha_Q^{k_Q}2^{-k_Q}d_Q^{k_Q/2}\frac{\Gamma((d_Q-1)/2)}{\Gamma(k_Q+(d_Q-1)/2)}\cdot\frac{\omega_{d_Q-2}}{\omega_{d_Q-1}}\\
&\qquad\times\int_{[-1,1]}(1-t^2)^{k_Q}\sigma^{(k_Q)}\left(\alpha_Q\sqrt{d_Q}t+u\right)(1-t^2)^{(d_Q-3)/2}\mathrm{d}t\\
&\quad=\alpha_Q^{k_Q}\frac{d_Q^{k_Q/2}\Gamma((d_Q-1)/2)}{2^{k_Q}\Gamma(k_Q+(d_Q-1)/2)}\mathbb{E}_{\overline{\boldsymbol{x}}^{(Q)}\sim \mathrm{Unif}(\mathbb{S}^{d_Q-1}(\sqrt{d_Q}))}\\
&\qquad\times\left[\left(1-(\overline{x}_1^{(Q)})^2/d_Q\right)^{k_Q}\sigma^{(k_Q)}\left(\alpha_Q\overline{x}_1^{(Q)}+u\right)\right].
\end{aligned}
\tag{165}
$$

Iterating equation (165) over $q\in[Q]$ and equation (164) yield the desired formula (162).

Furthermore, for each $q\in[Q]$,

$$
\begin{aligned}
k_q!B(d_q,k_q)&=(2k_q+d_q-2)\prod_{j=0}^{k_q-2}(j+d_q-1),\\
\frac{\Gamma((d_q-1)/2)}{2^{k_q}\Gamma(k_q+(d_q-1)/2)}&=\prod_{j=0}^{k_q-1}\frac{1}{2j+d_q-1}.
\end{aligned}
$$

Combining these two equations yields

$$
\begin{aligned}
&k_q!B(d_q,k_q)\frac{d_q^{k_q}\Gamma((d_q-1)/2)^2}{2^{2k_q}\Gamma(k_q+(d_q-1)/2)^2}\\
&\quad=\frac{2k_q+d_q-2}{2k_q+d_q-3}\cdot\left(\prod_{j=0}^{k_q-2}\frac{j+d_q-1}{2j+d_q-1}\right)\cdot\left(\prod_{j=0}^{k_q-1}\frac{d_q}{2j+d_q-1}\right),
\end{aligned}
\tag{166}
$$

which converges to 1 when $d_q\to\infty$. We deduce that

$$
\lim_{d\to\infty}B(\boldsymbol{d},\boldsymbol{k})R(\boldsymbol{d},\boldsymbol{k})^2=\frac{1}{\boldsymbol{k}!}.
$$

□

### J.2. Proof of convergence in probability of the Gegenbauer coefficients

**Lemma 19.** *Let $\boldsymbol{k} = (k_1, \dots, k_Q) \in \mathbb{Z}_{\geqslant 0}^Q$ and denote $|\mathbf{k}| = k_1 + \dots + k_Q$. Assume that the activation function $\sigma$ is $|\mathbf{k}|$-times weakly differentiable and denote $\sigma^{(|\mathbf{k}|)}$ its $|\mathbf{k}|$-weak derivative. Assume furthermore that there exist constants $c_0 > 0$ and $c_1 < 1$ such that $\sigma^{(|\mathbf{k}|)}(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ almost surely.*

*Then for any $\delta > 0$, there exists $\varepsilon_0 \in (0, 1)$ and $d_0$ such that for any $d \geqslant d_0$ and $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$,*

$$\left| \left( \prod_{q \in [Q]} d^{(\xi - \eta_q - \kappa_q)k_q} \right) B(\boldsymbol{d}, \boldsymbol{k}) \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 - \frac{\mu_{|\boldsymbol{k}|}(\sigma)^2}{\boldsymbol{k}!} \right| \leqslant \delta.$$

**Proof of lemma 19.** From lemma 18, we have

$$\begin{aligned} &\left( \prod_{q \in [Q]} d^{(\xi - \eta_q - \kappa_q)k_q} \right) B(\boldsymbol{d}, \boldsymbol{k}) \lambda_{\boldsymbol{k}}^{\boldsymbol{d}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 \\ &\quad = \left( \prod_{q \in [Q]} \alpha_q^{2k_q} d^{(\xi - \eta_q - \kappa_q)k_q} \right) \cdot [B(\boldsymbol{d}, \boldsymbol{k}) R(\boldsymbol{d}, \boldsymbol{k})^2] \\ &\qquad \times \mathbb{E}_{\overline{\boldsymbol{x}}} \left[ \prod_{q \in [Q]} \left( 1 - \frac{(\overline{x}_1^{(q)})^2}{d_q} \right)^{k_q} \cdot \sigma^{(|\boldsymbol{k}|)} \left( \sum_{q \in [Q]} \alpha_q \overline{x}_1^{(q)} \right) \right]^2. \end{aligned} \tag{167}$$

Recall $\alpha_q = \tau^{(q)} r_q / R$ with $r_q = d^{(\kappa_q + \eta_q)/2}$ and $R = d^{\xi/2}(1 + o_d(1))$. Hence, we have

$$\lim_{(d,\boldsymbol{\tau}) \to (\infty, \mathbf{1})} \prod_{q \in [Q]} \alpha_q^{2k_q} d^{(\xi - \eta_q - \kappa_q)k_q} = 1. \tag{168}$$

Furthermore, from lemma 18, we have

$$\lim_{d \to \infty} B(\boldsymbol{d}, \boldsymbol{k}) R(\boldsymbol{d}, \boldsymbol{k})^2 = \frac{1}{\boldsymbol{k}!}. \tag{169}$$

We can apply lemma 17 to the activation function $\sigma^{(|\boldsymbol{k}|)}$. In particular part $(c)$ of the lemma implies that there exists $\varepsilon_0 \in (0, 1)$ such that for $d$ sufficiently large, we have for any $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$,

$$\left| E_{\overline{\boldsymbol{x}}} \left[ \left( \prod_{q \in [Q]} \left( 1 - \frac{(\overline{x}_1^{(q)})^2}{d_q} \right)^{k_q} \right) \cdot \sigma^{(|\boldsymbol{k}|)} \left( \sum_{q \in [Q]} \alpha_q \overline{x}_1^{(q)} \right) \right] - \mathbb{E}_G[\sigma^{(|\boldsymbol{k}|)}(G)] \right| \leqslant \delta/2. \tag{170}$$

From equation (28), we have $\mathbb{E}_G[\sigma^{(|\boldsymbol{k}|)}(G)] = \mu_{|\boldsymbol{k}|}(\sigma)$. Combining equations (168) and (170) into equation (167) yields the result. □

**Lemma 20.** *Let $k$ be a non negative integer and denote $\boldsymbol{k} = (k, 0, \ldots, 0) \in \mathbb{Z}^Q_{\geqslant 0}$, where we recall that without loss of generality we choose $q = 1$ as the unique* $\mathrm{argmax}_{q\in[Q]}\{\eta_q + \kappa_q\}$. *Assume that the activation function $\sigma$ verifies $\sigma(u)^2 \leqslant c_0 \exp(c_1 u^2/2)$ almost surely for some constants $c_0 > 0$ and $c_1 < 1$.*

*Then for any $\delta > 0$, there exists $\varepsilon_0 = \varepsilon_0(c_1, \delta)$ and $d_0 = d_0(c_1, \delta)$ such that for any $d \geqslant d_0$ and $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$,*

$$\left| B(d_1, k)\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}})^2 - \frac{\mu_k(\sigma)^2}{k!} \right| \leqslant \delta.$$

**Proof of lemma 20.** 20. Recall the correspondence (29) between Gegenbauer and Hermite polynomials. Note for any monomial $m_l(x) = x^k$, we can apply lemma 17(c) to $m_l(\overline{x}^{(q_\xi)}_1)\sigma$ and find a coupling such that for any $\eta > 0$, there exists $\varepsilon_0 > 0$ and

$$\lim_{d\to\infty} \sup_{\boldsymbol{\tau}\in[1-\varepsilon_0,1+\varepsilon_0]^Q} \mathbb{E}_{\overline{\boldsymbol{x}},G}\left[\left(m_k(\overline{x}^{(q_\xi)}_1)\sigma_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}^{(1)}_1, \ldots, \overline{x}^{(Q)}_1) - m_k(G)\sigma(G)\right)^2\right] \leqslant \eta. \tag{171}$$

We have

$$[B(d_1, k)k!]^{1/2}\lambda^{\boldsymbol{d}}_{\boldsymbol{k}}(\sigma_{\boldsymbol{d},\boldsymbol{\tau}}) = \mathbb{E}_{\overline{\boldsymbol{x}}}[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}^{(1)}_1, \ldots, \overline{x}^{(Q)}_1)Q^{(d_1)}_k(\sqrt{d_1}\overline{x}^{(1)}_1)[B(d_1, k)k!]^{1/2}].$$

Using the asymptotic correspondence between Gegenbauer polynomials and Hermite polynomials (29)

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q^{(d)}_k(\sqrt{\mathrm{d}}x)B(d, k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\mathrm{He}_k(x)\right\},$$

and equation (171), we get for any $\delta > 0$, there exists $\varepsilon_0 > 0$ such that for $d$ sufficiently large, we have for any $\boldsymbol{\tau} \in [1 - \varepsilon_0, 1 + \varepsilon_0]^Q$,

$$\left| E_{\overline{\boldsymbol{x}}}\left[\sigma_{\boldsymbol{d},\boldsymbol{\tau}}(\overline{x}^{(1)}_1, \ldots, \overline{x}^{(Q)}_1)Q^{(d_1)}_k(\sqrt{d_1}\overline{x}^{(1)}_1)[B(d_1, k)k!]^{1/2}\right] - \mathbb{E}_G[\sigma(G)\mathrm{He}_k(G)] \right| \leqslant \delta,$$

which concludes the proof. □

## Appendix K. Bound on the operator norm of Gegenbauer polynomials

**Proposition 5 (Bound on the Gram matrix).** *Let $\boldsymbol{k} \in \mathbb{Z}^Q_{\geqslant 0}$ and denote $\gamma = \sum_{q\in[Q]} \eta_q k_q$. Let $n \leqslant d^\gamma/\mathrm{e}^{A_d\sqrt{\log d}}$ for any $A_d \to \infty$. Let $(\overline{\boldsymbol{x}}_i)_{i\in[n]}$ with $\overline{\boldsymbol{x}}_i = (\{\overline{\boldsymbol{x}}^{(q)}_i\}_{q\in[Q]}) \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$*

*independently, and* $Q_{k_q}^{(d_q)}$ *be the* $k_q$*'th Gegenbauer polynomial with domain* $[-d_q, d_q]$. *Consider the random matrix* $\boldsymbol{W} = (\boldsymbol{W}_{ij})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$, *with*

$$\boldsymbol{W}_{ij} = Q_{\boldsymbol{k}}^{\boldsymbol{d}}(\{\langle \overline{\boldsymbol{x}}_i^{(q)}, \overline{\boldsymbol{x}}_j^{(q)}\rangle\}_{q\in[Q]}) = \prod_{q\in[Q]} Q_{k_q}^{(d_q)}(\langle \overline{\boldsymbol{x}}_i^{(q)}, \overline{\boldsymbol{x}}_j^{(q)}\rangle).$$

*Then we have*

$$\lim_{d,n\to\infty} \mathbb{E}[\|\boldsymbol{W} - \mathbf{I}_n\|_{\mathrm{op}}] = 0.$$

**Corollary 1 (Uniform bound on the Gram matrix).** *Let* $n \leqslant d^\gamma/\mathrm{e}^{A_d\sqrt{\log d}}$ *for some* $\gamma > 0$ *and any* $A_d \to \infty$. *Let* $(\overline{\boldsymbol{x}}_i)_{i\in[N]}$ *with* $\overline{\boldsymbol{x}}_i = (\{\overline{\boldsymbol{x}}_i^{(q)}\}_{q\in[Q]}) \sim \mathrm{Unif}(\mathrm{PS}^{\boldsymbol{d}})$ *independently. Consider for any* $\boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q$, *the random matrix* $\boldsymbol{W}_{\boldsymbol{k}} = ((\boldsymbol{W}_{\boldsymbol{k}})_{ij})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ *as defined in proposition 5. Denote:*

$$\mathcal{Q} = \left\{ \boldsymbol{k} \in \mathbb{Z}_{\geqslant 0}^Q \,\middle|\, \sum_{q\in[Q]} \eta_q k_q < \gamma \right\}.$$

*Then we have*

$$\sup_{\boldsymbol{k}\in\mathcal{Q}^c} \mathbb{E}[\|\boldsymbol{W}_{\boldsymbol{k}} - \mathbf{I}_n\|_{\mathrm{op}}] = o_{d,\mathbb{P}}(1).$$

**Proof of corollary 1.** For each $q \in [Q]$, we consider $\boldsymbol{\Delta}^{(q)} = \boldsymbol{W}_k^{(q)} - \mathbf{I}_n$ where $\boldsymbol{W}_k^{(q)} = ((\boldsymbol{W}_k^{(q)})_{ij})_{i,j\in[n]}$ with

$$(\boldsymbol{W}_k^{(q)})_{ij} = Q_k^{(d_q)}(\langle \overline{\boldsymbol{x}}_i^{(q)}, \overline{\boldsymbol{x}}_j^{(q)}\rangle).$$

Then, defining $\gamma_q \equiv \gamma/\eta_q$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{k\geqslant 2\gamma_q+3} \|\boldsymbol{W}_k^{(q)} - \mathbf{I}_n\|_{\mathrm{op}}^2\right] &\leqslant \mathbb{E}\left[\sum_{k\geqslant 2\gamma_q+3} \|\boldsymbol{W}_k^{(q)} - \mathbf{I}_n\|_F^2\right] \\ &= n(n-1)\sum_{k\geqslant 2\gamma_q+3} \mathbb{E}[Q_k^{(d_q)}(\langle \overline{\boldsymbol{x}}^{(q)}, \overline{\boldsymbol{y}}^{(q)}\rangle)^2] = n(n-1)\sum_{k\geqslant 2\gamma_q+3} B(d_q,k)^{-1}.\end{aligned}$$

For $d$ sufficiently large, there exists $C > 0$ such that for any $p \geqslant m \equiv \lceil 2\gamma_q + 3\rceil$:

$$\begin{aligned}\frac{B(d_q,m)}{B(d_q,p)} = \prod_{k=m}^{p-1} \frac{(2k+d_q-2)}{(2k+d_q)} \cdot \frac{(k+1)}{(k+d_q-2)} &\leqslant \prod_{k=m}^{p-1} \frac{1}{1+(d_q-3)/(k+1)} \\ &\leqslant \prod_{k=m}^{p-1} \mathrm{e}^{-\frac{m+1}{d_q-2+m}\cdot\frac{d_q-2}{k+1}} \leqslant \frac{C}{p^2}.\end{aligned}$$

https://doi.org/10.1088/1742-5468/ac3a81

Hence, there exists constant $C'$, such that for large $d$, we have

$$\sum_{k\geqslant 2\gamma_q+3} B(d_q,k)^{-1} \leqslant C'\cdot B(d_q,m)^{-1}.$$

Recalling that $B(d_q,m)=\Theta_d(d^{\eta_q m})=\omega_d(d^{2\gamma})$, and $n=o_d(d^{\gamma})$, we deduce

$$\mathbb{E}\left[\sup_{k\geqslant 2\gamma_q+3}\|\boldsymbol{W}_k^{(q)}-\mathbf{I}_n\|_{\mathrm{op}}^2\right]=o_d(1). \tag{172}$$

Let us now consider $\boldsymbol{\Delta}=\boldsymbol{W}_{\boldsymbol{k}}-\mathbf{I}_n$. We will denote $\boldsymbol{\Delta}^{(q)}=\boldsymbol{W}_{k_q}^{(q)}-\mathbf{I}_n$. Then it is easy to check (recall the diagonal elements of $\boldsymbol{W}_{k_q}^{(d_q)}$ are equal to one) that for any $q\in[Q]$

$$\boldsymbol{\Delta}=\left(\bigodot_{q'\neq q}\boldsymbol{W}_{k_{q'}}^{(q')}\right)\odot\boldsymbol{\Delta}^{(q)}$$

where $\boldsymbol{A}\odot\boldsymbol{B}$ denotes the Hadamard product, or entrywise product, $(\boldsymbol{A}\odot\boldsymbol{B})_{i,j\in[n]}=(A_{ij}B_{ij})_{i,j\in[n]}$. We recall the following inequality on the operator norm of Hadamard product of two matrices, with $\boldsymbol{A}$ positive definite:

$$\|\boldsymbol{A}\odot\boldsymbol{B}\|_{\mathrm{op}}\leqslant\left(\max_{ij}\boldsymbol{A}_{ij}\right)\|\boldsymbol{B}\|_{\mathrm{op}}.$$

Hence, in particular

$$\|\boldsymbol{\Delta}\|_{\mathrm{op}}\leqslant\left(\prod_{q'\neq q}\max_{ij}[(\boldsymbol{W}_{k_{q'}}^{(q')})_{ij}]\right)\|\boldsymbol{\Delta}^{(q)}\|_{\mathrm{op}}.$$

Consider $\mathcal{I}=[0,2\gamma_1+3[\times\cdots\times[0,2\gamma_Q+3[\cap\mathbb{Z}_{\geqslant 0}^Q$. Then, from equation (172), we get directly

$$\sup_{\boldsymbol{k}\in\mathcal{I}^c}\|\boldsymbol{W}_{\boldsymbol{k}}-\mathbf{I}_n\|_{\mathrm{op}}=o_{d,\mathbb{P}}(1). \tag{173}$$

Furthermore, $\mathcal{I}\cap\mathcal{Q}$ is finite and from proposition 5, we directly get

$$\sup_{\boldsymbol{k}\in\mathcal{I}\cap\mathcal{Q}}\|\boldsymbol{W}_{\boldsymbol{k}}-\mathbf{I}_n\|_{\mathrm{op}}=o_{d,\mathbb{P}}(1). \tag{174}$$

Combining bounds (173) and (174) yields the result. □

### K.1. Proof of proposition 5

The proof follows closely the proof of the uniform case presented in [21]. For completeness, we copy here the relevant lemmas.

**Step 1. Bounding operator norm by moments.**

Denote $\boldsymbol{\Delta} = \boldsymbol{W} - \mathbf{I}_n$. We define for each $q \in [Q]$, $\boldsymbol{W}_{k_q}^{(d_q)} = (Q_{k_q}^{(d_q)}(\langle \overline{\boldsymbol{x}}_i^{(q)}, \overline{\boldsymbol{x}}_j^{(q)} \rangle))_{ij\in[n]}$ and $\boldsymbol{\Delta}^{(q)} = \boldsymbol{W}_{k_q}^{(d_q)} - \mathbf{I}_n$. Then it is easy to check (recall the diagonal elements of $\boldsymbol{W}_{k_q}^{(d_q)}$ are equal to one)

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}^{(1)} \odot \ldots \odot \boldsymbol{\Delta}^{(Q)},$$

where $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the Hadamard product, or entrywise product, $(\boldsymbol{A} \odot \boldsymbol{B})_{i,j\in[n]} = (A_{ij}B_{ij})_{i,j\in[n]}$. For any sequence of integers $p = p(d)$, we have

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] \leqslant \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})^{1/(2p)}] \leqslant \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})]^{1/(2p)}. \tag{175}$$

To prove the proposition, it suffices to show that for any sequence $A_d \to \infty$, we have

$$\lim_{d,n\to\infty, n=O_d(d^{\gamma}\,\mathrm{e}^{-A_d\sqrt{\log\, d}})} \mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})]^{1/(2p)} = 0. \tag{176}$$

In the following, we calculate $\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})]$. We have

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})] &= \sum_{\boldsymbol{i}=(i_1,\ldots,i_{2p})\in[n]^{2p}} \mathbb{E}[\Delta_{i_1i_2}\Delta_{i_2i_3}\ldots\Delta_{i_{2p}i_1}]\\ &= \sum_{\boldsymbol{i}=(i_1,\ldots,i_{2p})\in[n]^{2p}} \prod_{q\in[Q]} \mathbb{E}[\Delta_{i_1i_2}^{(q)}\Delta_{i_2i_3}^{(q)}\ldots\Delta_{i_{2p}i_1}^{(q)}],\end{aligned}$$

where we used that $\overline{\boldsymbol{x}}^{(q)}$ and $\overline{\boldsymbol{x}}^{(q')}$ are independent for $q \neq q'$.

We will denote for any $\boldsymbol{i} = (i_1, \ldots, i_k) \in [n]^k$, define for each $q \in [Q]$

$$M_{\boldsymbol{i}}^{(q)} = \begin{cases} \mathbb{E}[\Delta_{i_1i_2}^{(q)}\ldots\Delta_{i_ki_1}^{(q)}] & k \geqslant 2,\\ 1 & k = 1.\end{cases}$$

Similarly, we define $M_{\boldsymbol{i}}$ associated to $\boldsymbol{\Delta}$,

$$M_{\boldsymbol{i}} = \prod_{q\in[Q]} M_{\boldsymbol{i}}^{(q)}.$$

To calculate these quantities, we will apply repeatedly the following identity, which is an immediate consequence of equation (21). For any $i_1, i_2, i_3$ distinct, we have

$$\mathbb{E}_{\boldsymbol{\theta}_{i_2}}[\Delta_{i_1i_2}^{(q)}\Delta_{i_2i_3}^{(q)}] = \frac{1}{B(d_q, k_q)}\Delta_{i_1i_3}^{(q)}.$$

Throughout the proof, we will denote by $C, C', C''$ constants that may depend on $k$ but not on $p, d, n$. The value of these constants is allowed to change from line to line.

**Step 2. The induced graph and equivalence of index sequences.**

For any index sequence $\boldsymbol{i} = (i_1, i_2, \ldots, i_{2p}) \in [n]^{2p}$, we defined an undirected multigraph $G_{\boldsymbol{i}} = (V_{\boldsymbol{i}}, E_{\boldsymbol{i}})$ associated to index sequence $\boldsymbol{i}$. The vertex set $V_{\boldsymbol{i}}$ is the set of distinct elements in $i_1, \ldots, i_{2p}$. The edge set $E_{\boldsymbol{i}}$ is formed as follows: for any $j \in [2p]$ we add an edge between $i_j$ and $i_{j+1}$ (with convention $2p + 1 \equiv 1$). Notice that this could be

https://doi.org/10.1088/1742-5468/ac3a81

a self-edge, or a repeated edge: $G_{\boldsymbol{i}} = (V_{\boldsymbol{i}}, E_{\boldsymbol{i}})$ will be—in general—a multigraph. We denote $v(\boldsymbol{i}) = |V_{\boldsymbol{i}}|$ to be the number of vertices of $G_{\boldsymbol{i}}$, and $e(\boldsymbol{i}) = |E_{\boldsymbol{i}}|$ to be the number of edges (counting multiplicities). In particular, $e(\boldsymbol{i}) = k$ for $\boldsymbol{i} \in [n]^k$. We define

$$\mathcal{T}_{\star}(p) = \{\boldsymbol{i} \in [n]^{2p} : G_{\boldsymbol{i}}\,\text{does not have self edge}\}.$$

For any two index sequences $\boldsymbol{i}_1, \boldsymbol{i}_2$, we say they are equivalent $\boldsymbol{i}_1 \asymp \boldsymbol{i}_2$, if the two graphs $G_{\boldsymbol{i}_1}$ and $G_{\boldsymbol{i}_2}$ are isomorphic, i.e. there exists an edge-preserving bijection of their vertices (ignoring vertex labels). We denote the equivalent class of $\boldsymbol{i}$ to be

$$\mathcal{C}(\boldsymbol{i}) = \{\boldsymbol{j} : \boldsymbol{j} \asymp \boldsymbol{i}\}.$$

We define the quotient set $\mathcal{Q}(p)$ by

$$\mathcal{Q}(p) = \{\mathcal{C}(\boldsymbol{i}) : \boldsymbol{i} \in [n]^{2p}\}.$$

The following lemma was proved in proposition 3 in [21].

**Lemma 21.** *The following properties holds for all sufficiently large $n$ and $d$:*

*(a) For any equivalent index sequences $\boldsymbol{i} = (i_1, \dots, i_{2p}) \asymp \boldsymbol{j} = (j_1, \dots, j_{2p})$, we have $M_{\boldsymbol{i}}^{(q)} = M_{\boldsymbol{j}}^{(q)}$.*

*(b) For any index sequence $\boldsymbol{i} \in [n]^{2p} \setminus \mathcal{T}_{\star}(p)$, we have $M_{\boldsymbol{i}} = 0$.*

*(c) For any index sequence $\boldsymbol{i} \in \mathcal{T}_{\star}(p)$, the degree of any vertex in $G_{\boldsymbol{i}}$ must be even.*

*(d) The number of equivalent classes $|\mathcal{Q}(p)| \leqslant (2p)^{2p}$.*

*(e) Recall that $v(\boldsymbol{i}) = |V_{\boldsymbol{i}}|$ denotes the number of distinct elements in $\boldsymbol{i}$. Then, for any $\boldsymbol{i} \in [n]^{2p}$, the number of elements in the corresponding equivalence class satisfies $|\mathcal{C}(\boldsymbol{i})| \leqslant v(\boldsymbol{i})^{v(\boldsymbol{i})} \cdot n^{v(\boldsymbol{i})} \leqslant p^p n^{v(\boldsymbol{i})}$.*

In view of property (a) in the last lemma, given an equivalence class $\mathcal{C} = \mathcal{C}(\boldsymbol{i})$, we will write $M_{\mathcal{C}} = M_{\boldsymbol{i}}$ for the corresponding value.

**Step 3. The skeletonization process.**

For multi-graph $G$, we say that one of its vertices is *redundant*, if it has degree 2. For any index sequence $\boldsymbol{i} \in \mathcal{T}_{\star}(p) \subset [n]^{2p}$ (i.e. such that $G_{\boldsymbol{i}}$ does not have self-edges), we denote by $r(\boldsymbol{i}) \in \mathbb{N}_+$ to be the redundancy of $\boldsymbol{i}$, and by $\mathrm{sk}(\boldsymbol{i})$ to be the skeleton of $\boldsymbol{i}$, both defined by the following skeletonization process. Let $\boldsymbol{i}_0 = \boldsymbol{i} \in [n]^{2p}$. For any integer $s \geqslant 0$, if $G_{\boldsymbol{i}_s}$ has no redundant vertices then stop and set $\mathrm{sk}(\boldsymbol{i}) = \boldsymbol{i}_s$. Otherwise, select a redundant vertex $\boldsymbol{i}_s(\ell)$ arbitrarily (the $\ell$th element of $\boldsymbol{i}_s$). If $\boldsymbol{i}_s(\ell - 1) \neq \boldsymbol{i}_s(\ell + 1)$, then remove $\boldsymbol{i}_s(\ell)$ from the graph (and from the sequence), together with its adjacent edges, and connect $\boldsymbol{i}_s(\ell - 1)$ and $\boldsymbol{i}_s(\ell + 1)$ with an edge, and denote $\boldsymbol{i}_{s+1}$ to be the resulting index sequence, i.e. $\boldsymbol{i}_{s+1} = (\boldsymbol{i}_s(1), \dots, \boldsymbol{i}_s(\ell - 1), \boldsymbol{i}_s(\ell + 2), \dots, \boldsymbol{i}_s(\mathrm{end}))$. If $\boldsymbol{i}_s(\ell - 1) = \boldsymbol{i}_s(\ell + 1)$, then remove $\boldsymbol{i}_s(\ell)$ from the graph (and from the sequence), together with its adjacent edges, and denote $\boldsymbol{i}_{s+1}$ to be the resulting index sequence, i.e. $\boldsymbol{i}_{s+1} = (\boldsymbol{i}_s(1), \dots, \boldsymbol{i}_s(\ell - 1), \boldsymbol{i}_s(\ell + 1), \boldsymbol{i}_s(\ell + 2), \dots, \boldsymbol{i}_s(\mathrm{end}))$. (Here $\ell + 1$, and $\ell - 1$ have to be interpreted modulo $|\boldsymbol{i}_s|$, the length of $\boldsymbol{i}_s$.) The redundancy of $\boldsymbol{i}$, denoted by $r(\boldsymbol{i})$, is the number of vertices removed during the skeletonization process.

It is easy to see that the outcome of this process is independent of the order in which we select vertices.

**Lemma 22.** *For the above skeletonization process, the following properties hold*

(a) *If* $\boldsymbol{i} \asymp \boldsymbol{j} \in [n]^p$, *then* $\mathrm{sk}(\boldsymbol{i}) \asymp \mathrm{sk}(\boldsymbol{j})$. *That is, the skeletons of equivalent index sequences are equivalent.*

(b) *For any* $\boldsymbol{i} = (i_1, \ldots, i_k) \in [n]^k$, *and* $q \in [Q]$, *we have*

$$M_{\boldsymbol{i}}^{(q)} = \frac{M_{sk(\boldsymbol{i})}^{(q)}}{B(d_q, k_q)^{r(\boldsymbol{i})}}.$$

(c) *For any* $\boldsymbol{i} \in \mathcal{T}_\star(p) \subset [n]^{2p}$, *its skeleton is either formed by a single element, or an index sequence whose graph has the property that every vertex has degree greater or equal to 4.*

Given an index sequence $\boldsymbol{i} \in \mathcal{T}_\star(p) \subset [n]^{2p}$, we say $\boldsymbol{i}$ is of type 1, if $\mathrm{sk}(\boldsymbol{i})$ contains only one index. We say $\boldsymbol{i}$ is of type 2 if $\mathrm{sk}(\boldsymbol{i})$ is not empty (so that by lemma 22, $G_{\mathrm{sk}(\boldsymbol{i})}$ can only contain vertices with degree greater or equal to 4). Denote the class of type 1 index sequence (respectively type 2 index sequence) by $\mathcal{T}_1(p)$ (respectively $\mathcal{T}_2(p)$). We also denote by $\widetilde{\mathcal{T}}_a(p)$, $a \in \{1, 2\}$ the set of equivalence classes of sequences in $\mathcal{T}_a(p)$. This definition makes sense since the equivalence class of the skeleton of a sequence only depends on the equivalence class of the sequence itself.

**Step 4. Type 1 index sequences.**

Recall that $v(\boldsymbol{i})$ is the number of vertices in $G_{\boldsymbol{i}}$, and $e(\boldsymbol{i})$ is the number of edges in $G_{\boldsymbol{i}}$ (which coincides with the length of $\boldsymbol{i}$). We consider $\boldsymbol{i} \in \mathcal{T}_1(p)$. Since for $\boldsymbol{i} \in \mathcal{T}_1(p)$, every edge of $G_{\boldsymbol{i}}$ must be at most a double edge. Indeed, if $(u_1, u_2)$ had multiplicity larger than 2 in $G_{\boldsymbol{i}}$, neither $u_1$ nor $u_2$ could be deleted during the skeletonization process, contradicting the assumption that $\mathrm{sk}(\boldsymbol{i})$ contains a single vertex. Therefore, we must have $\min_{\boldsymbol{i}\in\mathcal{T}_1} v(\boldsymbol{i}) = p + 1$. According the lemma 22(b), for every $\boldsymbol{i} \in \mathcal{T}_1(p)$, we have

$$M_{\boldsymbol{i}} = \prod_{q\in[Q]} M_{\boldsymbol{i}}^{(q)} = \prod_{q\in[Q]} 1/B(d_q, k_q)^{v(\boldsymbol{i})-1} = \frac{1}{B(\boldsymbol{d}, \boldsymbol{k})^{v(\boldsymbol{i})-1}}.$$

Note by lemma 21(e), the number of elements in the equivalence class of $\boldsymbol{i}$ is $|\mathcal{C}(\boldsymbol{i})| \leqslant p^p \cdot n^{v(\boldsymbol{i})}$. Hence we get

$$\max_{\boldsymbol{i}\in\mathcal{T}_1(p)} [|\mathcal{C}(\boldsymbol{i})||M_{\boldsymbol{i}}|] \leqslant \sup_{\boldsymbol{i}\in\mathcal{T}_1(p)} \left[p^p n^{v(\boldsymbol{i})}/B(\boldsymbol{d}, \boldsymbol{k})^{v(\boldsymbol{i})-1}\right] = p^p n^{p+1}/B(\boldsymbol{d}, \boldsymbol{k})^p. \tag{177}$$

Therefore, denoting $K = \sum_{q\in[Q]} \eta_q k_q$,

$$\sum_{\boldsymbol{i}\in\mathcal{T}_1(p)} M_{\boldsymbol{i}} = \sum_{\mathcal{C}\in\widetilde{\mathcal{T}}_1(p)} |\mathcal{C}||M_{\mathcal{C}}| \tag{178}$$

$$\leqslant |\mathcal{Q}(p)| p^p \frac{n^{p+1}}{B(\boldsymbol{d}, \boldsymbol{k})^p} \leqslant (Cp)^{3p} n^{p+1} d^{-Kp}, \tag{179}$$

where in the last step we used lemma 21 and the fact that for $q \in [Q]$, $B(d_q, k_q) \geqslant C_0 d_q^{k_q}$ for some $C_0 > 0$.

**Step 5. Type 2 index sequences.**

We have the following simple lemma bounding $M_{\boldsymbol{i}}$, copied from proposition 3 in [21]. This bound is useful when $\boldsymbol{i}$ is a skeleton.

**Lemma 23.** *For any $q \in [Q]$, there exists constants $C$ and $d_0$ depending uniquely on $k_q$ such that, for any $d \geqslant d_0(k_q)$, and any index sequence $i \in [n]^m$ with $2 \leqslant m \leqslant d_q/(4k_q)$, we have*

$$|M_{\boldsymbol{i}}^{(q)}| \leqslant \left(Cm^{k_q} \cdot d_q^{-k_q}\right)^{m/2}.$$

Suppose $\boldsymbol{i} \in \mathcal{T}_2(p)$, and denote $v(i)$ to be the number of vertices in $G_{\mathrm{i}}$. We have, for a sequence $p = o_d(d)$, and each $q \in [Q]$

$$\begin{aligned}
|M_{\boldsymbol{i}}^{(q)}| &\overset{(1)}{=} \frac{|M_{\mathrm{sk}(\boldsymbol{i})}^{(q)}|}{B(d_q, k_q)^{r(\boldsymbol{i})}} \\
&\overset{(2)}{\leqslant} \left(\frac{Ce(\mathrm{sk}(\boldsymbol{i}))}{d_q}\right)^{k_q \cdot e(\mathrm{sk}(\boldsymbol{i}))/2} (C' d_q)^{-r(\boldsymbol{i})k_q} \\
&\overset{(3)}{\leqslant} \left(\frac{Cp}{d_q}\right)^{k_q \cdot e(\mathrm{sk}(\boldsymbol{i}))/2} (C' d_q)^{-r(\boldsymbol{i})k_q} \\
&\overset{(4)}{\leqslant} \left(\frac{Cp}{d_q}\right)^{k_q \cdot v(\mathrm{sk}(\boldsymbol{i}))} (C' d_q)^{-r(\boldsymbol{i})k_q} \\
&\overset{(5)}{\leqslant} C^{v(\boldsymbol{i})} p^{k_q \cdot v(\mathrm{sk}(\boldsymbol{i}))} d_q^{-(v(\mathrm{sk}(\boldsymbol{i}))+r(\boldsymbol{i}))\cdot k_q} \\
&\overset{(6)}{\leqslant} (Cp)^{k_q \cdot v(\boldsymbol{i})} d_q^{-v(\boldsymbol{i})k_q}.
\end{aligned}$$

Here (1) holds by lemma 22(b); (2) by lemma 23, and the fact that $\mathrm{sk}(i) \in [n]^{e(\mathrm{sk}(i))}$, together by $B(d_q, k_q) \geqslant C_0 d_q^{k_q}$; (3) because $e(\mathrm{sk}(i)) \leqslant 2p$; (4) by lemma 22(c), implying that for $i \in \mathcal{T}_2(p)$, each vertex of $G_{\mathrm{sk}(i)}$ has degree greater or equal to 4, so that $v(\mathrm{sk}(\mathbf{i})) \leqslant e(\mathrm{sk}(i))/2$ (notice that for $d \geqslant d_0(k_q)$ we can assume $Cp/d_q < 1$). Finally, (5) follows since $r(i), v(\mathrm{sk}(i)) \leqslant v(i)$, and (6) the definition of $r(i)$ implying $r(i) = v(i) - v(\mathrm{sk}(i))$.

Hence we get

$$|M_{\boldsymbol{i}}| \leqslant \prod_{q \in [Q]} (Cp)^{k_q \cdot v(\boldsymbol{i})} d_q^{-v(\boldsymbol{i})k_q}.$$

Note by lemma 21(e), the number of elements in equivalent class $|\mathcal{C}(\boldsymbol{i})| \leqslant p^{v(\boldsymbol{i})} \cdot n^{v(\boldsymbol{i})}$. Since $v(\mathbf{i})$ depends only on the equivalence class of $i$, we will write, with a slight abuse of notation $v(\boldsymbol{i}) = v(\mathcal{C}(\boldsymbol{i}))$. Notice that the number of equivalence classes with $v(\mathcal{C}) = v$ is upper bounded by the number multi-graphs with $v$ vertices and $2p$ edges, which is at

https://doi.org/10.1088/1742-5468/ac3a81

most $v^{4p}$. Denoting $\alpha = \ \max_{q\in[Q]}\{1/\eta_q\}$, we have

$$\begin{aligned}\sum_{\boldsymbol{i}\in\mathcal{T}_2(p)} M_{\boldsymbol{i}} &\leqslant \sum_{\mathcal{C}\in\widetilde{\mathcal{T}}_2(p)} |\mathcal{C}||M_{\mathcal{C}}| &(180)\\ &\leqslant \sum_{\mathcal{C}\in\widetilde{\mathcal{T}}_2(p)} (Cp^{\alpha})^{(K+1)v(\mathcal{C})}\left(\frac{n}{d^K}\right)^{v(\mathcal{C})} &(181)\\ &\leqslant \sum_{v=2}^{2p} v^{4p}\left(\frac{Cnp^{\alpha(K+1)}}{d^K}\right)^{v}. &(182)\end{aligned}$$

Define $\varepsilon = Cnp^{\alpha(K+1)}/d^K$. We will assume hereafter that $p$ is selected such that

$$2p \leqslant -\log\left(\frac{Cnp^{\alpha(K+1)}}{d^K}\right). \tag{183}$$

By calculus and condition (183), the function $F(v) = v^{4p}\varepsilon^{v}$ is maximized over $v \in [2, 2p]$ at $v = 2$, whence

$$\sum_{\boldsymbol{i}\in\mathcal{T}_2(p)} M_{\boldsymbol{i}} \leqslant 2pF(2) \leqslant C^p\left(\frac{n}{d^K}\right)^2. \tag{184}$$

**Step 6. Concluding the proof.**

Using equations (179) and (184), we have, for any $p = o_d(d)$ satisfying equation (183), we have

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Delta}^{2p})] &= \sum_{\boldsymbol{i}=(i_1,\dots,i_{2p})\in[N]^{2p}} M_{\boldsymbol{i}} = \sum_{\boldsymbol{i}\in\mathcal{T}_1(p)} M_{\boldsymbol{i}} + \sum_{\boldsymbol{i}\in\mathcal{T}_2(p)} M_{\boldsymbol{i}} &(185)\\ &\leqslant (Cp)^{3p}\frac{n^{p+1}}{d^{Kp}} + C^p\left(\frac{n}{d^K}\right)^2. &(186)\end{aligned}$$

Form equation (175), we obtain

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] \leqslant C\left\{p^{3/2}n^{1/(2p)}\sqrt{\frac{n}{d^K}} + \left(\frac{n}{d^K}\right)^{1/p}\right\}. \tag{187}$$

Finally setting $n = d^K\,\mathrm{e}^{-2A\sqrt{\log\,d}}$ and $p = (K/A)\sqrt{\log\,d}$, this yields

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] \leqslant C\left\{\mathrm{e}^{-\frac{A}{4}\sqrt{\log\,d}} + \mathrm{e}^{-2A^2/K}\right\}. \tag{188}$$

Therefore, as long as $A \to \infty$, we have $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] \to 0$. It is immediate to check that the above choice of $p$ satisfies the required conditions $p = o_d(d)$ and equation (183) for all $d$ large enough.

## Appendix L. Technical lemmas

We put here one technical lemma that is used in the proof of theorem 7(a).

**Lemma 24.** *Let $\boldsymbol{D} = (\boldsymbol{D}^{qq'})_{q,q' \in [Q]} \in \mathbb{R}^{DN \times DN}$ be a symmetric $Q$ by $Q$ block matrix with $\boldsymbol{D}^{qq'} \in \mathbb{R}^{d_q N \times d_{q'} N}$. Denote $\mathbf{B} = \mathbf{D}^{-1}$. Assume that $\mathbf{D}$ satisfies the following properties:*

*(a) For any $q \in [Q]$, there exists $c_q, C_q > 0$ such that we have with high probability*

$$0 < \frac{r_q^2}{d_q} c_q = d^{\kappa_q} c_q \leqslant \lambda_{\min}(\boldsymbol{D}^{qq}) \leqslant \lambda_{\max}(\boldsymbol{D}^{qq}) \leqslant \frac{r_q^2}{d_q} C_q = d^{\kappa_q} C_q < \infty,$$

*as $d \to \infty$.*

*(b) For any $q \neq q' \in [Q]$, we have $\sigma_{\max}(\boldsymbol{D}^{qq'}) = o_{d,\mathbb{P}}(r_q r_{q'}/\sqrt{d_q d_{q'}}) = o_{d,\mathbb{P}}(d^{(\kappa_q + \kappa_{q'})/2})$.*

*Then for any $q \neq q' \in [Q]$, we have*

$$\begin{aligned} \|\boldsymbol{B}^{qq}\|_{\mathrm{op}} &= O_{d,\mathbb{P}}\left(\frac{d_q}{r_q^2}\right) = O_{d,\mathbb{P}}(d^{-\kappa_q}), \\ \|\boldsymbol{B}^{qq'}\|_{\mathrm{op}} &= o_{d,\mathbb{P}}\left(\frac{\sqrt{d_q d_{q'}}}{r_q r_{q'}}\right) = o_{d,\mathbb{P}}(d^{-(\kappa_q + \kappa_{q'})/2}). \end{aligned} \tag{189}$$

**Proof of lemma 24.** Let us show the result recursively on the integer $Q$. Note that the case $Q = 1$ is direct.

Consider $\boldsymbol{D} = (\boldsymbol{D}^{qq'})_{q,q' \in [Q]}$. Denote $\tilde{D} = D - d_Q$, $\boldsymbol{A} = (\boldsymbol{D}^{qq'})_{q,q' \in [Q-1]} \in \mathbb{R}^{\tilde{D}N \times \tilde{D}N}$ and $\boldsymbol{C} = [(\boldsymbol{D}^{1Q})^{\mathsf{T}}, \ldots, (\boldsymbol{D}^{(Q-1)Q})^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{d_Q N \times \tilde{D} N}$ such that

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathsf{T}} & \boldsymbol{D}^{QQ} \end{bmatrix}.$$

Assume that $\boldsymbol{A}^{-1}$ verifies equation (189). Denote

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ \boldsymbol{T}^{\mathsf{T}} & \boldsymbol{B}^{QQ} \end{bmatrix}.$$

From the two by two blockmatrix inversion, we have:

$$\begin{aligned} \boldsymbol{B}^{QQ} &= (\boldsymbol{D}^{QQ} - \boldsymbol{C}^{\mathsf{T}} \boldsymbol{A}^{-1} \boldsymbol{C})^{-1}, \\ \boldsymbol{T} &= -\boldsymbol{A}^{-1} \boldsymbol{C} \boldsymbol{B}^{QQ}. \end{aligned}$$

We have

$$\begin{aligned}\left\|\boldsymbol{C}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{C}\right\|_{\mathrm{op}} &\leqslant \sum_{q,q'\in[Q-1]} \left\|(\boldsymbol{D}^{qQ})^{\mathsf{T}}(\boldsymbol{A}^{-1})_{qq'}\boldsymbol{D}^{q'Q}\right\|_{\mathrm{op}}\\ &= \sum_{q,q'\in[Q-1]} o_{d,\mathbb{P}}\left(\frac{r_q r_Q}{\sqrt{d_q d_Q}}\right)\cdot O_{d,\mathbb{P}}\left(\frac{\sqrt{d_q d_{q'}}}{r_q r_{q'}}\right)\cdot o_{d,\mathbb{P}}\left(\frac{r_{q'} r_Q}{\sqrt{d_{q'} d_Q}}\right)\\ &= o_{d,\mathbb{P}}(r_Q^2/d_Q),\end{aligned}$$

where we used in the second line the properties on $\boldsymbol{D}$ and our assumption on $\boldsymbol{A}^{-1}$. Hence $\boldsymbol{D}^{QQ} - \boldsymbol{C}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{C} \preceq (r_q^2/d_q)(c_q - o_{d,\mathbb{P}}(1))\mathbf{I}$ and $\|\boldsymbol{B}^{QQ}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d_q/r_q^2)$.

Furthermore, for $q < Q$,

$$\boldsymbol{B}^{qQ} = -\sum_{q'\in[Q-1]} (\boldsymbol{A}^{-1})_{qq'}\boldsymbol{C}_{q'}\boldsymbol{B}^{QQ}.$$

Hence

$$\begin{aligned}\left\|\boldsymbol{B}^{qQ}\right\|_{\mathrm{op}} &\leqslant \sum_{q'\in[Q-1]} \left\|(\boldsymbol{A}^{-1})_{qq'}\boldsymbol{D}^{q'Q}\boldsymbol{B}^{QQ}\right\|_{\mathrm{op}}\\ &= \sum_{q,q'\in[Q-1]} O_{d,\mathbb{P}}\left(\frac{\sqrt{d_q d_{q'}}}{r_q r_{q'}}\right)\cdot o_{d,\mathbb{P}}\left(\frac{r_q r_Q}{\sqrt{d_{q'} d_Q}}\right)\cdot O_{d,\mathbb{P}}\left(\frac{d_Q}{r_Q^2}\right)\\ &= o_{d,\mathbb{P}}\left(\frac{\sqrt{d_q d_Q}}{r_q r_Q}\right),\end{aligned}$$

which finishes the proof. □

### L.1. Useful lemmas from [21]

For completeness, we reproduce in this section lemmas proven in [21].

**Lemma 25.** *The number $B(d,k)$ of independent degree-$k$ spherical harmonics on $\mathbb{S}^{d-1}$ is non-decreasing in $k$ for any fixed $d \geqslant 2$.*

**Lemma 26.** *For any fixed $k$, let $Q_k^{(d)}(x)$ be the kth Gegenbauer polynomial. We expand*

$$Q_k^{(d)}(x) = \sum_{s=0}^{k} p_{k,s}^{(d)} x^s.$$

*Then we have*

$$p_{k,s}^{(d)} = O_d(d^{-k/2-s/2}).$$

**Lemma 27.** *Let* $N = o_d(d^{\ell+1})$ *for a fixed integer* $\ell$. *Let* $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ *independently. Then as* $d \to \infty$, *we have*

$$\max_{i\neq j}|\langle \boldsymbol{w}_i, \boldsymbol{w}_j\rangle| = O_{d,\mathbb{P}}((\log\, d)^{k/2} d^{-k/2}).$$

**Proposition 6 (Bound on the Gram matrix).** *Let* $N \leqslant d^k/\mathrm{e}^{A_d\sqrt{\log\, d}}$ *for a fixed integer* $k$ *and any* $A_d \to \infty$. *Let* $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ *independently, and* $Q_k^{(d)}$ *be the* $k$*'th Gegenbauer polynomial with domain* $[-d, d]$. *Consider the random matrix* $\boldsymbol{W} = (\boldsymbol{W}_{ij})_{i,j\in[N]} \in \mathbb{R}^{N\times N}$, *with* $\boldsymbol{W}_{ij} = Q_k^{(d)}(\langle \boldsymbol{\theta}_i, \boldsymbol{\theta}_j\rangle)$. *Then we have*

$$\lim_{d,N\to\infty} \mathbb{E}[\|\boldsymbol{W} - \mathbf{I}_d\|_{\mathrm{op}}] = 0.$$

## References

[1] Rahimi A and Recht B 2008 Random features for large-scale kernel machines *Advances in Neural Information Processing Systems* pp 1177–84

[2] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *Advances in Neural Information Processing Systems* pp 8571–80

[3] Du S S, Zhai X, Poczos B and Singh A 2019 Gradient descent provably optimizes over-parameterized neural networks *Int. Conf. on Learning Representations*

[4] Du S, Lee J, Li H, Wang L and Zhai X Gradient descent finds global minima of deep neural networks *Proc. 36th Int. Conf. on Machine Learning, vol 97 of Proc. of Machine Learning Research, PMLR* ed K Chaudhuri and R Salakhutdinov (Long Beach, California, USA, 9–15 June 2019) pp 1675–85

[5] Allen-Zhu Z, Li Y and Song Z A convergence theory for deep learning via over-parameterization *Proc. 36th Int. Conf. on Machine Learning, vol 97 of Proc. of Machine Learning Research, PMLR* ed K Chaudhuri and R Salakhutdinov (Long Beach, California, USA, 9–15 June 2019) pp 242–52

[6] Zou D, Cao Y, Zhou D and Gu Q 2018 Stochastic gradient descent optimizes over-parameterized deep relu networks (arXiv:1811.08888)

[7] Oymak S and Soltanolkotabi M 2020 Towards moderate overparameterization: global convergence guarantees for training shallow neural networks *IEEE Journal on Selected Areas in Information Theory* **1** 84–105

[8] Chizat L, Oyallon E and Bach F 2019 On lazy training in differentiable programming *Advances in Neural Information Processing Systems* pp 2933–43

[9] Arora S, Du S S, Li Z, Salakhutdinov R, Wang R and Yu D 2020 Harnessing the power of infinitely wide deep nets on small-data tasks *Int. Conf. on Learning Representations*

[10] Li Z, Wang R, Yu D, Du S S, Hu W, Salakhutdinov R and Arora S 2019 Enhanced convolutional neural tangent kernels (arXiv:1911.00809)

[11] Lee J, Xiao L, Schoenholz S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2019 Wide neural networks of any depth evolve as linear models under gradient descent *Advances in Neural Information Processing Systems* pp 8570–81

[12] Novak R, Xiao L, Bahri Y, Lee J, Yang G, Abolafia D A, Pennington J and Sohl-dickstein J 2019 Bayesian deep convolutional networks with many channels are Gaussian processes *Int. Conf. on Learning Representations*

[13] Lee J, Sohl-dickstein J, Pennington J, Novak R, Schoenholz S and Bahri Y 2018 Deep neural networks as Gaussian processes *Int. Conf. on Learning Representations*

[14] De Matthews A G G, Hron J, Rowland M, Turner R E and Ghahramani Z 2018 Gaussian process behaviour in wide deep neural networks *6th Int. Conf. on Learning Representations, ICLR 2018-Conf. Track Proc.*

[15] Garriga-Alonso A, Rasmussen C E and Aitchison L 2019 Deep convolutional networks as shallow Gaussian processes *Int. Conf. on Learning Representations*

[16] Shankar V, Fang A, Guo W, Fridovich-Keil S, Schmidt L, Ragan-Kelley J and Recht B 2020 Neural kernels without tangents (arXiv:2003.02237)

[17] Arora S, Du S S, Hu W, Li Z, Salakhutdinov R R and Wang R 2019 On exact computation with an infinitely wide neural net *Advances in Neural Information Processing Systems* pp 8139–48

[18] Geiger M, Spigler S, Jacot A and Wyart M 2019 Disentangling feature and lazy learning in deep neural networks: an empirical study (arXiv:1906.08034)

[19] Yehudai G and Shamir O 2019 On the power and limitations of random features for understanding neural networks *Advances in Neural Information Processing Systems* pp 6594–604

[20] Bach F 2017 Breaking the curse of dimensionality with convex neural networks *J. Mach. Learn. Res.* **18** 629–81

[21] Ghorbani B, Mei S, Misiakiewicz T and Montanari A 2019 Linearized two-layers neural networks in high dimension (arXiv:1904.12191)

[22] Ghorbani B, Mei S, Misiakiewicz T and Montanari A 2019 Limitations of lazy training of two-layers neural network *Advances in Neural Information Processing Systems* pp 9108–18

[23] Allen-Zhu Z and Li Y 2019 What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems* pp 9017–28

[24] Allen-Zhu Z and Li Y 2020 Backward feature correction: how deep learning performs deep learning (arXiv:2001.04413)

[25] Mei S, Bai Y and Montanari A 2018 The landscape of empirical risk for nonconvex losses *Ann. Stat.* **46** 2747–74

[26] Mei S, Montanari A and Nguyen P-M 2018 A mean field view of the landscape of two-layer neural networks *Proc. Natl Acad. Sci. USA* **115** E7665–71

[27] Chizat L and Bach F 2018 On the global convergence of gradient descent for over-parameterized models using optimal transport *Advances in Neural Information Processing Systems* pp 3036–46

[28] Rotskoff G M and Vanden-Eijnden E 2018 Neural networks as interacting particle systems: asymptotic convexity of the loss landscape and universal scaling of the approximation error (arXiv:1805.00915)

[29] Sirignano J and Spiliopoulos K 2018 Mean field analysis of neural networks (arXiv:1805.01053)

[30] Chizat L and Bach F 2020 Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss (arXiv:2002.04486)

[31] Donoho D L and Johnstone I M 1995 Adapting to unknown smoothness via wavelet shrinkage *J. Am. Stat. Assoc.* **90** 1200–24

[32] Page D 2018 (Myrtle.ai) How to train your ResNet 4: architecture (https://myrtle.ai/how-to-train-your-resnet-4-architecture/)

[33] Tsybakov A B 2008 *Introduction to Nonparametric Estimation* (Berlin: Springer)

[34] Yin D, Lopes R G, Shlens J, Cubuk E D and Gilmer J 2019 A Fourier perspective on model robustness in computer vision *Advances in Neural Information Processing Systems* pp 13255–65

[35] Nguyen P-M 2019 Mean field limit of the learning dynamics of multilayer neural networks (arXiv:1902.02880)

[36] Nguyen P-M and Pham H T 2020 A rigorous framework for the mean field limit of multilayer neural networks (arXiv:2001.11443)

[37] Abadi M *et al* 2016 Tensorflow: a system for large-scale machine learning *12th (USENIX) Symp. on Operating Systems Design and Implementation ((OSDI) 16)* pp 265–83

[38] Ghorbani B, Krishnan S and Xiao Y 2019 An investigation into neural net optimization via hessian eigenvalue density *Int. Conf. on Machine Learning* pp 2232–41

[39] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[40] Bradbury J, Frostig R, Hawkins P, Johnson M J, Leary C, Maclaurin D and Wanderman-Milne S 2018 JAX: composable transformations of Python+NumPy programs (http://github.com/google/jax)

[41] Novak R, Xiao L, Hron J, Lee J, Alemi A A, Sohl-Dickstein J and Schoenholz S S 2020 Neural tangents: fast and easy infinite neural networks in python *Int. Conf. on Learning Representations*

[42] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Int. Conf. on Machine Learning* pp 448–56