Appl. Comput. Harmon. Anal. 59 (2022) 3–84

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration

Song Mei [a], Theodor Misiakiewicz [b,*], Andrea Montanari [b,c]

[a] *Department of Statistics, University of California, Berkeley, United States of America*
[b] *Department of Statistics, Stanford University, United States of America*
[c] *Department of Electrical Engineering, Stanford University, United States of America*

ARTICLE INFO

Article history:
Received 19 July 2021
Accepted 13 December 2021
Available online 17 December 2021
Communicated by David Donoho

Keywords:
Random features
Kernel methods
Generalization error
High dimensional limit

ABSTRACT

Consider the classical supervised learning problem: we are given data $(y_i, \boldsymbol{x}_i)$, $i \le n$, with y_i a response and $\boldsymbol{x}_i \in \mathcal{X}$ a covariates vector, and try to learn a model $\hat{f} : \mathcal{X} \to \mathbb{R}$ to predict future responses. Random feature methods map the covariates vector $\boldsymbol{x}_i$ to a point $\boldsymbol{\phi}(\boldsymbol{x}_i)$ in a higher dimensional space $\mathbb{R}^N$, via a random featurization map $\boldsymbol{\phi}$. We study the use of random feature methods in conjunction with ridge regression in the feature space $\mathbb{R}^N$. This can be viewed as a finite-dimensional approximation of kernel ridge regression (KRR), or as a stylized model for neural networks in the so called lazy training regime.

We define a class of problems satisfying certain spectral conditions on the underlying kernels, and a hypercontractivity assumption on the associated eigenfunctions. These conditions are verified by classical high-dimensional examples. Under these conditions, we prove a sharp characterization of the error of random feature ridge regression. In particular, we address two fundamental questions: (1) What is the generalization error of KRR? (2) How big N should be for the random feature approximation to achieve the same error as KRR?

In this setting, we prove that KRR is well approximated by a projection onto the top ℓ eigenfunctions of the kernel, where ℓ depends on the sample size n. We show that the test error of random feature ridge regression is dominated by its approximation error and is larger than the error of KRR as long as $N \le n^{1-\delta}$ for some $\delta > 0$. We characterize this gap. For $N \ge n^{1+\delta}$, random features achieve the same error as the corresponding KRR, and further increasing N does not lead to a significant change in test error.

Contents

* Corresponding author.
E-mail address: misiakie@stanford.edu (T. Misiakiewicz).

https://doi.org/10.1016/j.acha.2021.12.003
1063-5203/

1. Introduction

1.1. Background

Consider the supervised learning problem in which we are given i.i.d. samples $(y_i, \boldsymbol{x}_i)$, $i \le n$, from a common probability distribution on $\mathbb{R} \times \mathcal{X}$. Here $\boldsymbol{x}_i \in \mathcal{X}$ is a vector of covariates, and y_i is a response variable. We are interested in learning a model $\hat{f} : \mathcal{X} \to \mathbb{R}$ which, given a new point $\boldsymbol{x}_{\text{test}}$, predicts the corresponding response y_{test} via $\hat{f}(\boldsymbol{x}_{\text{test}})$.

A number of statistical learning methods can be viewed as a combination of two steps: featurization and training. Featurization maps sample points into a convenient 'feature space' $\mathcal{H}$ (a vector space) via a featurization map $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$, $\boldsymbol{x}_i \mapsto \boldsymbol{\phi}(\boldsymbol{x}_i)$. Training fits a model that is linear in the feature space: $\hat{f}(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle_{\mathcal{H}}$. In this paper we will be concerned with a relatively simple method for training, ridge regression:

$$\hat{\boldsymbol{a}}(\lambda) := \arg\min_{\boldsymbol{a}} \Big\{ \sum_{i=1}^{n} \big(y_i - \langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle_{\mathcal{H}} \big)^2 + \lambda \|\boldsymbol{a}\|_{\mathcal{H}}^2 \Big\} . \tag{1}$$

Here it is implicitly assumed that $\mathcal{H}$ is an Hilbert space, and therefore $\boldsymbol{a} \in \mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $\| \cdot \|_{\mathcal{H}}$ are the scalar product and norm in $\mathcal{H}$.

It is useful to discuss a few examples of this paradigm, some of which will play a role in what follows (we refer to Section 2.1 for formal definitions).

Feature engineering. We use this term to refer to the classical approach of crafting a set of N features $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \dots, \phi_N(\boldsymbol{x})) \in \mathcal{H} = \mathbb{R}^N$ for a specific application, by leveraging domain expertise. This has been the standard approach to computer vision for a long time [28,10], and is still the state of the art in most of applied statistics [24].

Kernel methods. In this case $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) defined implicitly via a positive definite kernel $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [15]. Rather than manually constructing features, the statistician/data analyst only needs to encode in $H(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2) \rangle_{\mathcal{H}}$ a suitable notion of similarity in the input space $\mathcal{X}$. The resulting model only depends on the kernel H, and a crucial role is played by its eigenvalue decomposition $H(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{\ell=1}^{\infty} \lambda_\ell^2 \psi_\ell(\boldsymbol{x}_1) \psi_\ell(\boldsymbol{x}_2)$. Ridge regression with RKHS featurization is referred to as kernel ridge regression (KRR). Formally, the KRR estimator takes the form:

$$\hat{f}_\lambda(\boldsymbol{x}) = \sum_{\ell=1}^{\infty} \hat{f}_{\lambda,\ell} \psi_\ell(\boldsymbol{x}) , \quad \hat{f}_{\lambda,\ell} = \sum_{\ell'=1}^{\infty} ((\lambda/n) \cdot \mathbf{I} + \boldsymbol{G})^{-1}_{\ell,\ell'} \lambda_\ell \lambda_{\ell'} \langle \psi_{\ell'}, y \rangle_n , \tag{2}$$

$$G_{\ell,\ell'} := \lambda_\ell \lambda_{\ell'} \langle \psi_\ell, \psi_{\ell'} \rangle_n . \tag{3}$$

Here $\langle f, g \rangle_n := n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_i) g(\boldsymbol{x}_i)$ denotes the scalar product with respect to the empirical measure.

For large n, we can imagine to replace the empirical scalar product with its population counterpart, and therefore $G_{\ell,\ell'} \approx \lambda_\ell^2 \mathbf{1}_{\ell=\ell'}$, whence $\hat{f}_{\lambda,\ell} \approx ((\lambda/n) + \lambda_\ell^2)^{-1} \lambda_\ell^2 \langle \psi_\ell, y \rangle_n$. In words, KRR attempts to estimate accurately the projection of $f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$ onto the eigenvectors of the kernel H, corresponding to large eigenvalues λ_ℓ. On the other hand, it shrinks towards 0 the projections of f onto eigenvectors corresponding to smaller eigenvalues.

Random Features (RF). Instead of constructing the featurization map $\boldsymbol{\phi}$ on the basis of domain expertise, or, implicitly, via a kernel, RF methods use a random map $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^N$. We will study a general construction that generalizes the original proposal of [38,7]. We sample N points in a space Ω via $\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_N \sim_{iid} \tau$ (for a certain probability measure τ on Ω), and then define the mapping $\boldsymbol{\phi}$ by letting $\boldsymbol{\phi}(\boldsymbol{x}) = (\sigma(\boldsymbol{x}; \boldsymbol{\theta}_1), \dots, \sigma(\boldsymbol{x}; \boldsymbol{\theta}_N))$. Here $\sigma : \mathcal{X} \times \Omega \to \mathbb{R}$ is a square integrable function. We endow the feature space $\mathcal{H}_N = \mathbb{R}^N$ with the inner product $\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_{\mathcal{H}_N} = \boldsymbol{a}_1^{\mathsf{T}} \boldsymbol{a}_2 / N$.

Because of the connection to two-layer neural networks (see below) we shall refer to N as the 'number of neurons' (although, 'number of parameters' would be more appropriate), and to σ as the 'activation function.' The resulting function $\hat{f}$ takes the form

$$\hat{f}(\boldsymbol{x}; \boldsymbol{a}) := \langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle_{\mathcal{H}} = \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\boldsymbol{x}; \boldsymbol{\theta}_i) . \tag{4}$$

We will refer to the procedure defined by Eq. (1) with $\boldsymbol{\phi}$ the random feature map defined here as 'random feature ridge regression' (RFRR). RFRR is closely related to KRR. First of all, we can view RFRR as an example of KRR, with kernel

$$H_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2) \rangle_{\mathcal{H}} = \frac{1}{N} \sum_{i=1}^{N} \sigma(\boldsymbol{x}_1; \boldsymbol{\theta}_i) \sigma(\boldsymbol{x}_2; \boldsymbol{\theta}_i).$$

Notice however that the kernel H_N has finite rank and is random, because of the random features $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$.

Second, for large N, we can expect H_N to be a good approximation of its expectation

$$\mathbb{E} H_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = H(\boldsymbol{x}_1, \boldsymbol{x}_2) := \int_{\Omega} \sigma(\boldsymbol{x}_1; \boldsymbol{\theta}) \, \sigma(\boldsymbol{x}_2; \boldsymbol{\theta}) \, \tau(\mathrm{d}\boldsymbol{\theta}) \, . \tag{5}$$

Hence, for large N, we expect RFRR to have generalization properties similar to the underlying RKHS, while possibly exhibiting lower complexity because it only operates on $N \times n$ matrices (instead of $n \times n$ matrices as for KRR).

Neural networks in the linear (lazy) regime. The methods described above fit the general paradigm of Eq. (1). Training does not affect the feature map $\boldsymbol{\phi}$. The model $\hat{f}_\lambda(\,\cdot\,)$ is linear in $\boldsymbol{y}$, as a consequence of the fact that the loss is quadratic (see also Eq. (2)). In contrast, neural networks aim at learning the best feature representation of the data. The feature map changes during training, and indeed there is no clear separation between the feature map $\boldsymbol{\phi}(\boldsymbol{x})$ and the coefficients $\boldsymbol{a}$.

Nevertheless a copious line of recent research shows that —under certain training schemes— neural networks are well approximated by their linearization around a random initialization [25,27,19,17,4,3,1,46, 36]. It is useful to recall the basic argument here. Denote by $\boldsymbol{x} \mapsto f(\boldsymbol{x}; \boldsymbol{\theta})$ the neural network, with parameters (weights) $\boldsymbol{\theta} \in \mathbb{R}^N$, and by $\boldsymbol{\theta}_0$ the initialization for gradient-based training. For highly overparametrized networks, a small change in the parameters $\boldsymbol{\theta}$ is sufficient to change significantly the evaluations of f at the data points, i.e., the vector $(f(\boldsymbol{x}_1; \boldsymbol{\theta}), \dots, f(\boldsymbol{x}_n; \boldsymbol{\theta}))$. As a consequence, an empirical risk minimizer can be found in a small neighborhood of the initialization $\boldsymbol{\theta}_0$, and it is legitimate to approximate f by its first order Taylor expansion in the parameters:

$$f(\boldsymbol{x}; \boldsymbol{\theta}_0 + \boldsymbol{a}) \approx f(\boldsymbol{x}; \boldsymbol{\theta}_0) + \langle \boldsymbol{a}, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}_0) \rangle \, . \tag{6}$$

Apart from the zero-th order term $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$ (which has no free parameters, and hence plays the role of an offset), this linearized model takes the same form $\hat{f}(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$. The featurization map is given by $\boldsymbol{\phi}(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}_0)$. We refer to the model $\boldsymbol{x} \mapsto \langle \boldsymbol{a}, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}_0) \rangle$ as the neural tangent (NT) model.

Notice that the NT featurization map is random, because of the random initialization $\boldsymbol{\theta}_0$. However, in general it does not take the form of the RF model, because the entries of $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}_0)$ are not independent. Despite this important difference, we expect key properties of the RF model to generalize to suitable classes of NT models. Examples of this phenomenon were studied recently in [20,34].

The present paper focuses on KRR and RFRR. We introduce a set of assumptions on the data distribution, the choice of activation function, and the probability distribution τ on the $\boldsymbol{\theta}_i$'s, under which we can characterize the large n, N behavior of the generalization (test) error. While our results apply to an abstract input space $\mathcal{X}$, our assumptions aim at capturing the behavior observed when $\mathcal{X}$ is high-dimensional, and the distribution ν on $\mathcal{X}$ satisfies strong concentration properties. For instance, our results apply to $\mathcal{X} = \mathbb{S}^{d-1}$ (the sphere in d dimension) or $\mathcal{X} = \{+1, -1\}^d$, both endowed with the uniform measure.

Our results do not require the true regression function f_* to belong to the associated RKHS and they characterize the test error (with respect to the square loss) pointwise, i.e., for each specific target function f_*. This characterization holds up to error terms that are negligible compared to the null risk $\mathbb{E}\{f_*(\boldsymbol{x})^2\}$.

In particular our results allow to answer in a quantitative way two sets of key questions that emerge from the above discussion:

Q1. How does the test error of KRR depends on the sample size n, on the target function f_*, and on the kernel H? While this question has attracted considerable attention in the past (see Section 1.3 for an overview), a very precise answer can be given in the present setting.
Q2. How does the test error of RFRR depend on the sample size n, and the number of neurons N? In particular, for a given sample size, how big N should be to achieve the same error as for the associated KRR (which corresponds formally to $N = \infty$)?
Q3. How do the answers to the previous questions depend on the regularization parameter λ? In particular, in which cases is the optimal test error achieved by choosing $\lambda \to 0$, i.e. by using the minimum norm interpolator to the training data?

Let us emphasize that the second question is technically more challenging than the first one, because it amounts to studying KRR *with a random kernel*. The setting introduced here is particularly motivated by the objective to address Q2 (and its ramifications in Q3). Indeed, to the best of our knowledge, we provide the first set of results on the optimal choice of the overparametrization N/n under polynomial scalings of N, n, d.

1.2. Summary of main results

Before summarizing our results, it is useful to describe informally our assumptions: we refer to Sections 2.2 and 3.2 for a formal statement of the same assumptions. We consider $(\boldsymbol{x}_i)_{i\le n} \sim_{iid} \nu$ with ν a probability distribution of the covariates space $\mathcal{X}$, and $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, where f is the target function and $\varepsilon_i \sim \mathsf{N}(0, \sigma_\varepsilon^2)$ independent of $\boldsymbol{x}_i$ is noise.

An RFRR problem is specified by $\nu, f_*, \sigma_\varepsilon$ (which determine the data distribution), σ, τ (which determine the RF representation), and the parameters n, N (sample size and number of neurons). The associated kernel problem is obtained by using the kernel (5). It is also useful to introduce a kernel in the $\boldsymbol{\theta}$ space via $U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \mathbb{E}_{\boldsymbol{x}\sim\nu}\{\sigma(\boldsymbol{x};\boldsymbol{\theta}_1)\,\sigma(\boldsymbol{x};\boldsymbol{\theta}_2)\}$.

We will consider sequences of such problems indexed by an integer d, and characterize their behavior as $N, n, d \to \infty$. In applications, d typically corresponds to the dimension of the covariates space $\mathcal{X}$. In this informal summary, we drop any reference to d for simplicity.

We next describe informally our key assumptions, which depends on integers $(\mathsf{m}, \mathsf{M}, u)$, with $u \ge \max(\mathsf{M}, \mathsf{m})$. (For the sake of simplicity, we omit some assumptions of a more technical nature.)

1. *Hypercontractivity.* The top u eigenvectors of H are 'delocalized'. We formalize this condition by requiring that, for any function $g \in \mathrm{span}(\psi_j : j \le u)$, and for any integer k, $\mathbb{E}_\nu\{g(\boldsymbol{x})^{2k}\} \le C_{k,u}\mathbb{E}_\nu\{g(\boldsymbol{x})^2\}^k$. We assume a same condition for the eigenvectors of U.
2. *Concentration of diagonal elements of the kernels.* Denote by $H_{>\mathsf{m}}$ the kernel obtained from H by setting to zero the eigenvalues $\lambda_1, \dots, \lambda_{\mathsf{m}}$. We require that the diagonal elements $\{H_{>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x}_i)\}_{i\le n}$ concentrate around their expectation with respect to the measure ν on $\mathcal{X}$. Analogously, we require the diagonal elements $\{U_{>\mathsf{M}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)\}_{i\le N}$ to concentrate around their expectation.

   This assumption amounts to a condition of symmetry: most points $\boldsymbol{x}$ in the support of ν are roughly equivalent, in the sense of having the same value of $H_{>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x})$, and similarly for most $\boldsymbol{\theta}$ in the support of ν.
3. *Spectral gap.* Recall that the $(\lambda_j^2)_{j\ge1}$ denote the eigenvalues of the kernel H in decreasing order. We then assume one of the following two conditions to hold:

Undeparametrized regime. We have $N \ll n$ and

$$\frac{1}{\lambda_{\mathsf{M}}^2} \sum_{k=\mathsf{M}+1}^{\infty} \lambda_k^2 \ll N \ll \frac{1}{\lambda_{\mathsf{M}+1}^2} \sum_{k=\mathsf{M}+1}^{\infty} \lambda_k^2 \,, \tag{7}$$

Overparametrized regime. We have $n \ll N$ and

$$\frac{1}{\lambda_{\mathsf{m}}^2} \sum_{k=\mathsf{m}+1}^{\infty} \lambda_k^2 \ll n \ll \frac{1}{\lambda_{\mathsf{m}+1}^2} \sum_{k=\mathsf{m}+1}^{\infty} \lambda_k^2 \,. \tag{8}$$

This assumption ensures a clear separation between the subspace of $\mathcal{D}_d$ which is estimated accurately (spanned by the eigenfunction of H corresponding to the top eigenvalues) and the subspace that is estimated trivially by 0 (corresponding to the low eigenvalues of H.) As we will see, a spectral gap condition holds for classical high-dimensional examples. On the other hand, we believe it should be possible to avoid this condition at the price of a more complicated characterization of the risk, and indeed we do not require it for KRR.

As explained above, KRR attempts to estimate accurately the projection of the target function f_* onto the top eigenvectors of the kernel H, and shrinks to zero its other components. RFRR behaves similarly, except that it only constructs a finite rank approximation of the kernel H. How many components of the target function are estimated accurately? There are of course two limiting factors: the statistical error, which depends on the sample size n; and the approximation error, which depends on the number of neurons N.

It turns out that, in the present setting, the interplay between n and N takes a particularly simple form. In a nutshell, what matters is the smaller of n and N. If $n \ll N$, then the statistical error dominates, and ridge regression estimates correctly the projection of f_* onto the top m eigenfunctions of H (where m is defined per Eq. (8)). If on the other hand $N \ll n$, then the approximation error dominates and ridge regression estimates correctly the projection of f_* onto the top M eigenfunctions of H (where M is defined per Eq. (7)).

In formulas, we denote by $R_{\mathrm{RF}}(f_*; \lambda) = \mathbb{E}\{(f_*(\boldsymbol{x}) - f_\lambda(\boldsymbol{x}))^2\}$ the test error of RFRR (for square loss) when the target function is f_* and the regularization parameter equals λ. Our main result establishes that for all $\lambda \in [0, \lambda_*]$ (with a suitable choice of λ_*), in a certain asymptotic sense, the following hold:

$$R_{\mathrm{RF}}(f_*; \lambda) = \begin{cases} \mathbb{E}\{(\mathsf{P}_{>\mathsf{m}} f_*(\boldsymbol{x}))^2\} + o(1) \cdot \mathbb{E}\{f_*(\boldsymbol{x})^2\} & \text{if } n \ll N, \\ \mathbb{E}\{(\mathsf{P}_{>\mathsf{M}} f_*(\boldsymbol{x}))^2\} + o(1) \cdot \mathbb{E}\{f_*(\boldsymbol{x})^2\} & \text{if } n \gg N, \end{cases} \tag{9}$$

where $\mathsf{P}_{>\ell}$ is the projector onto the span of the eigenfunctions $\{\psi_j : j > \ell\}$. This statement also applies to KRR, if we interpret the latter as the $N = \infty$ case of RFRR. Further, no kernel machine achieves a smaller error.

This characterization implies a relatively simple answer to questions Q1, Q2, and Q3, which we posed in the previous section. We summarize some of the insights that follow from this result.

KRR acts as a projection. As mentioned above, Eq. (9) can be restated as saying that (for the special case $N = \infty$), $\hat{f}_\lambda(\boldsymbol{x}) \approx \mathsf{P}_{\le \mathsf{m}} f_*(\boldsymbol{x})$. Indeed, we will prove a stronger result, which does not require the spectral gap assumption of Eq. (8). The KRR estimator $\hat{f}_\lambda$ is well approximated by the KRR estimator for the population problem ($n = \infty$), but with a larger value of the ridge regularization $\gamma > \lambda$. In other words KRR acts as a shrinkage operator along the eigenfunctions of the kernel.

Effects of overparametrization. In random feature models, we are free to choose the number of neurons N. Equation (9) indicates that any choice of N has roughly the same test error (which is also the test error of KRR) as long as $N \gg n$. This is interesting in both directions. First, the test error does not deteriorate as the number of parameters increases far beyond the sample size. This contrasts with a naive measure of the model complexity: indeed, counting the number of parameters would naively suggest that $N \gg n$ might hurt generalization. Second, the error does not improve with overparametrization either, as long as $N \gg n$.

Optimal overparametrization. At what level of overparametrization should we operate? In view of the previous point, it is sufficient to use a model with a number of parameters much larger than the sample size (formally, $N \geq n^{1+\delta}$ for some $\delta > 0$, although this specific condition is mainly dictated by our proof technique). Further overparametrization does not improve the statistical behavior.

Let us also note that —as proven in [31]— choosing $N/n =: \psi = O(1)$ can lead to sub-optimal test error, with the suboptimality vanishing if $\psi \to \infty$ *after* $N, n \to \infty$.

Optimality of interpolation. Finally, the above phenomena are obtained for all $\lambda \in [0, \lambda_*]$. The case $\lambda = 0$ corresponds to minimum norm interpolators. We also prove that the risk of any kernel machine is lower bounded by $\mathbb{E}\{(\mathsf{P}_{>\mathsf{m}} f_*(\boldsymbol{x}))^2\} + o(1) \cdot \mathbb{E}\{f_*(\boldsymbol{x})^2\}$. We therefore conclude that, in the overparametrized regime $N \gg n$, min-norm interpolators are optimal among all kernel methods.

1.3. Related literature

The test error of KRR was studied by a number of authors in the past [16,26], [43, Theorem 13.17]. In particular, [16] establishes that KRR achieves minimax optimal rates over certain subclasses of the associated RKHS. However these results require a strictly positive ridge regularizer (and hence do not cover interpolation) and characterize the decay of the error as $n \to \infty$ in fixed dimension d. In contrast our focus is on the case in which both d and n grow simultaneously. Further, we provide upper and lower bounds that hold pointwise (for each given target function f_*) while earlier work mostly establish pointwise upper bound and minimax lower bounds (for the worst case f_*). The recent work [26] also derived pointwise upper and lower bounds for kernel ridge regression (but with strictly positive ridge regularizer), which is very similar to our Theorem 5. However, these results are based on a universality assumption whose validity is unclear in specific settings.

Recently, the ridge-less (interpolation) limit of KRR was studied by Liang, Rakhlin and Zhai [29,30]. Again, these authors provide minimax upper bounds that hold within the RKHS, holding for inner product kernel, when the feature vectors $\boldsymbol{x}$ have independent coordinates. Their results are related but not directly comparable to ours.

The complexity of training a kernel machine scales at least quadratically in the sample size. This has motivated the development of randomized techniques to lower the complexity of training and testing. While our focus is on random feature methods, alternative approaches are based on subsampling the columns-rows of the empirical kernel matrix, see e.g. [5,2,37]. In particular, [37] compares the prediction errors using the sketched and the full kernel matrices, and shows that —for a fixed RKHS— it is sufficient to use a number of rows/columns of the order of the square root of the sample size in order to achieve the minimax rate over that RKHS.

The generalization properties of random feature methods have been studied in a smaller number of papers [39,40,33]. Rahimi and Recht [39] proved an upper bound of the order $1/\sqrt{N} + 1/\sqrt{n}$ on the generalization error. The insight provided by this bound is similar to one of our points: about $N \asymp n$ neurons are sufficient for the error to be of the same order as for $N \to \infty$. On the other hand, [39] proves only a minimax upper bound, it is limited to Lipschitz losses, and, crucially, requires the coefficients $\max_{i \leq N} |a_i| \leq C$ so that $\|\boldsymbol{a}\|_2^2 = O(N)$. In contrast, in the present setting, we typically have $\|\boldsymbol{a}\|_2^2 = \Theta(nN)$ The case of square loss was considered earlier by Rudi and Rosasco [40] who proved that, for a target function f_* in the

RKHS, $N = C\sqrt{n}\log n$ is sufficient to learn a random feature model with test error of order $1/\sqrt{n}$. These authors interpret this finding as implying that roughly $\sqrt{n}$ random features are sufficient: we will discuss the difference between their setting and ours in Section 2.3.

Finally, [6] studies optimized distributions for sampling the random features, while [45] provides a comparison between random feature approaches and subsampling of the kernel matrix.

As pointed out above, we find that taking $\lambda \to 0$ yields nearly optimal test error, within our setting. Optimality of minimum norm interpolators has attracted considerable attention recently [11,14,23,12,41]. In particular our results point in the same direction as the general analysis of ridge regression in [12, 41]. Note however that the general results of [12,41] do not apply to the present setting because they require subgaussian features $\boldsymbol{\phi}(\boldsymbol{x}_i)$. Further, they only provide upper and lower bounds that match up to factors depending on the condition number of a certain random matrix. In contrast, our characterization is specialized to the random feature setting, does not require subgaussianity, and holds up to additive errors that are negligible compared to the null risk.

The present paper solves a number of open problems that were left open in our earlier work [20]. First of all, [20] only considered the cases $n = \infty$ (approximation error of random feature models) or $N = \infty$ (generalization error of KRR). Here instead we establish the complete picture for both n and N finite. Second, [20] assumed a special data distribution (ν was the uniform distribution over the d-dimensional sphere), a special structure for the kernel (inner product kernels), and a special type of activation functions (depending on the inner product $\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle$)). The present paper considers general data distribution, kernel, and activation function, under a set of assumptions that covers the previous example as a special case. Finally, the proofs of [20] made use of the moment method, which is difficult to generalize beyond special examples. Here we use a decoupling approach and matrix concentration methods which are significantly more flexible.

The results of [20] were generalized to certain anisotropic distributions in [21]. For the inner product activation functions on the sphere, the precise asymptotics (for $N, n, d \to \infty$ with $N/d \to \psi_1$, $n/d \to \psi_2$, $\psi_1, \psi_2 \in (0, \infty)$) of generalization error of random feature models was calculated in [31].

1.4. Notations

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

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

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

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

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

2. Generalization error of random feature ridge regression

In this section, we present our results on the generalization error of random feature models. We begin in Section 2.1 by introducing the general abstract setting in which we work, and some of its basic properties. We then state our assumptions in Section 2.2, and state our main theorem (Theorem 1) in Section 2.3.

Finally, Section 2.4 presents applications of our general theorem to (i) the case of covariates vectors uniformly distributed over the sphere $\boldsymbol{x}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and (ii) the case of covariates vectors uniformly distributed over the Hamming cube $\boldsymbol{x}_i \sim \mathrm{Unif}(\{+1,-1\}^d)$. While these applications are 'simple' in the sense that checking the assumptions of our general theorem is straightforward, they are in themselves quite interesting. In particular, our result for the uniform distribution on the sphere (cf. Proposition 2) closes the main problem left unsolved in [20].

2.1. Random feature models, kernels, and their spectral decomposition

We consider two sequences of Polish probability spaces $(\mathcal{X}_d, \nu_d)$ and (Ω_d, τ_d), indexed by an integer d. We denote by $L^2(\mathcal{X}_d) = L^2(\mathcal{X}_d, \nu_d)$ the space of square integrable functions on $(\mathcal{X}_d, \nu_d)$, and by $L^2(\Omega_d) = L^2(\Omega_d, \tau_d)$ the space of square integrable functions on (Ω_d, τ_d). Since $(\mathcal{X}_d, \nu_d)$ and (Ω_d, τ_d) are standard probability spaces [18, Theorem 13.1.1], it follows that $L^2(\mathcal{X}_d)$ and $L^2(\Omega_d)$ are separable.

More generally for $p \geq 1$, we denote $\|f\|_{L^p(\mathcal{X})} = \mathbb{E}_{\boldsymbol{x}\sim\nu}[|f(\boldsymbol{x})|^p]^{1/p}$ the L^p norm of f. We will sometimes omit $\mathcal{X}$ and write directly $\|f\|_{L^2}$ and $\|f\|_{L^p}$ when clear from context.

Given two closed linear subspaces $\mathcal{D}_d \subseteq L^2(\mathcal{X}_d)$, $\mathcal{V}_d \subseteq L^2(\Omega_d)$, and the activation function $\sigma_d \in L^2(\mathcal{X}_d \times \Omega_d, \nu_d \otimes \tau_d)$, we define a Fredholm integral operator $\mathbb{T}_d : \mathcal{D}_d \to \mathcal{V}_d$ via

$$\mathbb{T}_d g(\boldsymbol{\theta}) \equiv \int_{\mathcal{X}_d} \sigma_d(\boldsymbol{x}, \boldsymbol{\theta}) g(\boldsymbol{x}) \nu_d(\mathrm{d}\boldsymbol{x}). \tag{10}$$

Note that $\mathbb{T}_d$ is a compact operator by construction. We will assume that $\mathbb{T}_d g \neq 0$ for any $g \in \mathcal{D}_d \setminus \{0\}$. Also, without loss of generality, we can assume $\mathcal{V}_d = \mathrm{Im}(\mathbb{T}_d)$ (which is closed since $\mathbb{T}_d$ is bounded). With an abuse of notation, we will sometimes denote by $\mathbb{T}_d$ the extension of this operator obtained by setting $\mathbb{T}_d g = 0$ for $g \in \mathcal{D}_d^{\perp}$. Notice that we can choose the kernel σ_d so that $\int_{\mathcal{X}_d} \sigma_d(\boldsymbol{x}, \boldsymbol{\theta}) g(\boldsymbol{x}) \nu_d(\mathrm{d}\boldsymbol{x}) = 0$ for any $g \in \mathcal{D}_d^{\perp}$: we will assume such a choice hereafter.

While in simple examples we might assume $\mathcal{D}_d = L^2(\mathcal{X}_d)$, the extra flexibility afforded by a general subspace $\mathcal{D}_d \subseteq L^2(\mathcal{X}_d)$ allows to model some important applications (see Section 2.5 and [32]).

The adjoint operator $\mathbb{T}_d^* : \mathcal{V}_d \to \mathcal{D}_d$ has kernel representation

$$\mathbb{T}_d^* f(\boldsymbol{x}) = \int_{\Omega_d} \sigma_d(\boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{\theta}) \tau_d(\mathrm{d}\boldsymbol{\theta}).$$

As before, we will sometimes extend $\mathbb{T}_d^*$ to $\mathcal{L}^2(\Omega_d)$ by setting $\mathrm{Ker}(\mathbb{T}_{\mathrm{d}}^*) = \mathcal{V}_{\mathrm{d}}^{\perp}$.

The operator $\mathbb{T}_d$ induces two compact self-adjoint positive definite operators: $\mathbb{U}_d = \mathbb{T}_d \mathbb{T}_d^* : \mathcal{V}_d \to \mathcal{V}_d$, and $\mathbb{H}_d = \mathbb{T}_d^* \mathbb{T}_d : \mathcal{D}_d \to \mathcal{D}_d$. These operators admit the kernel representations:

$$\mathbb{U}_d f(\boldsymbol{\theta}) = \int_{\Omega_d} U_d(\boldsymbol{\theta}, \boldsymbol{\theta}') f(\boldsymbol{\theta}') \tau_d(\mathrm{d}\boldsymbol{\theta}'), \tag{11}$$

$$\mathbb{H}_d g(\boldsymbol{x}) = \int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \nu_d(\mathrm{d}\boldsymbol{x}'), \tag{12}$$

where $U_d : \Omega_d \times \Omega_d \to \mathbb{R}$ and $H_d : \mathcal{X}_d \times \mathcal{X}_d \to \mathbb{R}$ are two measurable functions, satisfying respectively $\int_{\Omega_d} U_d(\boldsymbol{\theta}, \boldsymbol{\theta}') f(\boldsymbol{\theta}')\, \tau_d(\mathrm{d}\boldsymbol{\theta}') = 0$ for $f \in \mathcal{V}_d^{\perp}$, and $\int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}')\, \nu_d(\mathrm{d}\boldsymbol{x}') = 0$ for $g \in \mathcal{D}_d^{\perp}$. We immediately have

$$U_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x} \sim \nu_d}[\sigma_d(\boldsymbol{x}, \boldsymbol{\theta}_1)\sigma_d(\boldsymbol{x}, \boldsymbol{\theta}_2)], \tag{13}$$

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{\theta} \sim \tau_d}[\sigma_d(\boldsymbol{x}_1, \boldsymbol{\theta})\sigma_d(\boldsymbol{x}_2, \boldsymbol{\theta})]. \tag{14}$$

By Cauchy-Schwartz inequality, we have $U_d \in L^2(\Omega_d \times \Omega_d)$ and $H_d \in L^2(\mathcal{X}_d \times \mathcal{X}_d)$.

By the spectral theorem of compact operators, there exist two orthonormal bases $(\psi_j)_{j \ge 1}$, $\mathrm{span}(\psi_j, j \ge 1) = \mathcal{D}_d \subseteq L^2(\mathcal{X}_d)$ and $(\phi_j)_{j \ge 1}$, $\mathrm{span}(\phi_j, j \ge 1) = \mathcal{V}_d \subseteq L^2(\Omega_d)$, and eigenvalues $(\lambda_{d,j})_{j \ge 1} \subseteq \mathbb{R}$, with nonincreasing absolute values $|\lambda_{d,1}| \ge |\lambda_{d,2}| \ge \cdots$, and $\sum_{j \ge 1} \lambda_{d,j}^2 < \infty$ such that

$$\mathbb{T}_d = \sum_{j=1}^{\infty} \lambda_{d,j} \psi_j \phi_j^*, \quad \mathbb{U}_d = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \phi_j \phi_j^*, \quad \mathbb{H}_d = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \psi_j \psi_j^*.$$

(Here convergence holds in operator norm.) In terms of the kernel, these identities read

$$\sigma_d(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} \lambda_{d,j} \psi_j(\boldsymbol{x}) \phi_j(\boldsymbol{\theta}), \qquad U_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \phi_j(\boldsymbol{\theta}_1) \phi_j(\boldsymbol{\theta}_2),$$

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \psi_j(\boldsymbol{x}_1) \psi_j(\boldsymbol{x}_2). \tag{15}$$

Here convergence holds in $L^2(\mathcal{X}_d \times \Omega_d)$, $L^2(\Omega_d \times \Omega_d)$, and $L^2(\mathcal{X}_d \times \mathcal{X}_d)$.

Associated to the operator $\mathbb{H}$, we can define a reproducing kernel Hilbert space (RKHS) $\mathcal{H} \subseteq \mathcal{D}_d$ defined as

$$\mathcal{H} = \Big\{ f \in \mathcal{D} : \|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} \lambda_{d,j}^{-2} \langle f, \psi_j \rangle_{L^2}^2 < \infty \Big\},$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the RKHS norm associated to $\mathcal{H}$. In particular, $\mathcal{H}$ is dense in $\mathcal{D}_d$, provided $\lambda_{d,j}^2 > 0$ for all j.

For $S \subseteq \{1, 2, \dots\}$, we denote by P_S the projection operator from $L^2(\mathcal{X}_d)$ onto $\mathcal{D}_{d,S} := \mathrm{span}(\psi_j, j \in S)$. With a little abuse of notations, we also denote by P_S the projection operator from $L^2(\Omega_d)$ onto $\mathcal{V}_{d,S} := \mathrm{span}(\phi_j, j \in S)$. We denote by $\mathbb{T}_{d,S}$ and $\sigma_{d,S}$ the corresponding operator and kernel

$$\mathbb{T}_{d,S} = \sum_{j \in S} \lambda_{d,j} \psi_j \phi_j^*,$$

$$\sigma_{d,S}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j \in S} \lambda_{d,j} \psi_j(\boldsymbol{x}) \phi_j(\boldsymbol{\theta}).$$

We define $\mathbb{U}_{d,S} = \mathbb{T}_{d,S} \mathbb{T}_{d,S}^*$ and $\mathbb{H}_{d,S} = \mathbb{T}_{d,S}^* \mathbb{T}_{d,S}$, and denote by $U_{d,S}$ and $H_{d,S}$ the corresponding kernels. If $S = \{j \in \mathbb{N} : \ j \le \ell\}$ we will write for brevity $\mathbb{T}_{d,\le \ell}$, $\mathbb{U}_{d,\le \ell}$, $\mathbb{H}_{d,\le \ell}$, and similarly for $S = \{j \in \mathbb{N} : \ j > \ell\}$.

Since $\sigma_d \in L^2(\mathcal{X}_d \times \Omega_d)$, it follows that $\mathbb{U}_{d,S}$ is trace class, for any $S \subseteq \mathbb{N}$, with trace given by

$$\mathrm{Tr}(\mathbb{U}_{d,S}) \equiv \sum_{j \in S} \lambda_{d,j}^2 = \mathbb{E}_{\boldsymbol{\theta} \sim \tau_d}[U_{d,S}(\boldsymbol{\theta}, \boldsymbol{\theta})] < \infty\,.$$

Similarly, we have

$$\mathrm{Tr}(\mathbb{H}_{d,S}) \equiv \sum_{j\in S} \lambda_{d,j}^2 = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,S}(\boldsymbol{x},\boldsymbol{x})] < \infty\,.$$

2.2. Assumptions

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{i\in[N]} \sim_{iid} \tau_d$. We define the random feature function class by:

$$\mathcal{F}_{\mathrm{RF},N}(\boldsymbol{\Theta}) = \Big\{ \hat{f}(\boldsymbol{x};\boldsymbol{a}) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma_d(\boldsymbol{x},\boldsymbol{\theta}_i) : a_i \in \mathbb{R}, i\in[N] \Big\}.$$

Note that the factor $1/N$ is immaterial here, and only introduced in order to match the definition of feature map and scalar product in Section 1.1. Note that we use $\hat{f}$ instead of f to indicate that $(\boldsymbol{x},\boldsymbol{a}) \mapsto \hat{f}(\boldsymbol{x};\boldsymbol{a})$ is a specific function $\mathbb{R}^d \times \mathbb{R}^N \to \mathbb{R}$. It is also useful to view $\boldsymbol{a} \mapsto \hat{f}(\,\cdot\,;\boldsymbol{a})$ as a map $\mathbb{R}^N \to L^2(\mathbb{R}^d;\mathbb{P})$.

We observe pairs $(y_i,\boldsymbol{x}_i)_{i\in[n]}$, with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \nu_d$, and $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, $f_* \in L^2(\mathcal{X}_d)$ and $\varepsilon_i \sim \mathsf{N}(0,\sigma_\varepsilon^2)$ independently. We fit the coefficients $(a_i)_{i\le N}$ using ridge regression, cf. Eq. (1) that we reproduce here

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}} \left\{ \sum_{i=1}^{n} \big(y_i - \hat{f}(\boldsymbol{x}_i;\boldsymbol{a})\big)^2 + \frac{\lambda}{N}\|\boldsymbol{a}\|_2^2 \right\}. \tag{16}$$

We allow λ to depend on the dimension parameter d. The test error is given by

$$R_{\mathrm{RF}}(f_*,\boldsymbol{X},\boldsymbol{\Theta},\lambda) := \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x};\hat{\boldsymbol{a}}(\lambda))\Big)^2\Big]. \tag{17}$$

We next state our assumptions on the sequences of probability spaces $(\mathcal{X}_d,\nu_d)$ and (Ω_d,τ_d), and on the activation functions σ_d. The first set of assumptions concerns the concentration properties of the feature map, and are grouped in the next definition. These assumptions are quantified by four sequences of integers $\{(N(d),\mathsf{M}(d),n(d),\mathsf{m}(d))\}_{d\ge1}$, where $N(d)$ and $n(d)$ are, respectively, the number of neurons and the sample size. The integers $\mathsf{M}(d)$ and $\mathsf{m}(d)$ play a minor role in this definition, but will encode the decomposition of $L^2(\Omega_d)$ and $L^2(\mathcal{X}_d)$ (respectively) into the span of the top eigenvectors of $\mathbb{U}_d$ and $\mathbb{H}_d$ (of dimensions $\mathsf{M}(d)$ and $\mathsf{m}(d)$) and their complements.

Assumption 1 *($\{(N(d),\mathsf{M}(d),n(d),\mathsf{m}(d))\}_{d\ge1}$-Feature Map Concentration Property). We say that the sequence of activation functions $\{\sigma_d\}_{d\ge1}$ satisfies the Feature Map Concentration Property (FMCP) with respect to the sequence $\{(N(d),\mathsf{M}(d),n(d),\mathsf{m}(d))\}_{d\ge1}$ if there exists a sequence $\{u(d)\}_{d\ge1}$ with $u(d) \ge \max(\mathsf{M}(d),\mathsf{m}(d))$ such that the following hold.*

(a) (Hypercontractivity of finite eigenspaces)

(i) (Hypercontractivity of finite eigenspaces on $\mathcal{D}_d$.) For any integer $k \ge 1$, there exists C such that, for any $g \in \mathcal{D}_{d,\le u(d)} = \mathrm{span}(\psi_s, 1\le s\le u(d))$, we have

$$\|g\|_{L^{2k}(\mathcal{X}_d)} \le C\cdot\|g\|_{L^2(\mathcal{X}_d)}.$$

(ii) (Hypercontractivity of finite eigenspaces on $\mathcal{V}_d$.) For any integer $k \ge 2$, there exists C' such that, for any $g \in \mathcal{V}_{d,\le u(d)} = \mathrm{span}(\phi_s, 1\le s\le u(d))$, we have

$$\|g\|_{L^{2k}(\Omega_d)} \le C'\cdot\|g\|_{L^2(\Omega_d)}.$$

(b) (Properly decaying eigenvalues.) There exists a fixed $\delta_0 > 0$, such that, for all d large enough

$$\max(N(d), n(d))^{2+\delta_0} \le \frac{\left(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2\right)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4}\,. \tag{18}$$

(c) (Hypercontractivity of the high degree part.) Let $\sigma_{d,>u(d)}$ correspond to the projection on the high degree part of σ_d. Then there exists a fixed $\delta_0 > 0$ and an integer k such that

$$\min(n, N)^{1+2\delta_0} \max(N, n)^{1/k-1} \log(\max(N, n)) = o_d(1),$$

and

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\sigma_{>u(d)}(\boldsymbol{x};\boldsymbol{\theta})^{2k}]^{1/(2k)} = O_d(1) \cdot \min(n, N)^{\delta_0} \cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\sigma_{>u(d)}(\boldsymbol{x};\boldsymbol{\theta})^2]^{1/2}.$$

(d) (Concentration of diagonal elements) For $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$ and $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim_{iid} \tau_d$, we have

$$\sup_{i\in[n(d)]} \left| H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x})] \right| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x})]\,,$$
$$\sup_{i\in[N(d)]} \left| U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}}[U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}, \boldsymbol{\theta})] \right| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_{\boldsymbol{\theta}}[U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}, \boldsymbol{\theta})].$$

This statement formalizes three assumptions. The first one is hypercontractivity (points (a) and (c)). Recall that $\mathcal{D}_{d,\le u(d)}$ is the eigenspace spanned by top eigenvectors of the operator $\mathbb{H}_d$, and $\mathcal{V}_{d,\le u(d)}$ is the eigenspace spanned by top eigenvectors of the operator $\mathbb{U}_d$. We request that functions in these spaces have comparable norms of all orders, which roughly amounts to saying that they take values of the same order as their typical value for most $\boldsymbol{x}$ (or most $\boldsymbol{\theta}$). This typically happens when the functions in the top eigenspaces are delocalized.

The second assumption (assumption (b)) requires that the eigenvalues of kernel operators do not decay too rapidly. If this is not the case, the RKHS will be very close to a low-dimensional space. For instance, if $\lambda_{d,k}^2 \asymp k^{-2\alpha}$, $\alpha > 0$, then this condition holds as long as we take $u(d) \ge \max(N(d), n(d))^{2+\delta_0}$ for some $\delta_0 > 0$.

Finally, assumption (d) concerns the diagonal elements of the kernel matrices. They require the truncated kernel functions $H_{d,>\mathsf{m}(d)}$ and $U_{d,>\mathsf{M}(d)}$ evaluated on covariates and weight vectors to have nearly constant diagonal values.

The second set of assumptions concerns the spectrum of the kernel operator, defined by the sequence of eigenvalues $(\lambda_{d,j}^2)_{j\ge1}$. We require that the spectrum has a gap: the location of this gap dictates the relationship between $N(d)$ and $\mathsf{M}(d)$ and between $n(d)$ and $\mathsf{m}(d)$.

Assumption 2 *(Spectral gap at level $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d))\}_{d\ge1}$). We say that the sequence of activation functions $\{\sigma_d\}_{d\ge1}$ has a spectral gap at level $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d))\}_{d\ge1}$ if one of the following conditions (a), (b) hold for all d large enough.*

(a) (Overparametrized regime.) We have $N(d) \ge n(d)$ and

(i) (Number of samples) There exists fixed $\delta_0 > 0$ such that $\mathsf{m}(d) \le n(d)^{1-\delta_0}$ and

$$\frac{1}{\lambda_{d,\mathsf{m}(d)}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2 \le n(d)^{1-\delta_0} \le n(d)^{1+\delta_0} \le \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2\,. \tag{19}$$

(ii) (Number of features) There exists fixed $\delta_0 > 0$ *such that* $\mathsf{M}(d) \le N(d)^{1-\delta_0}$, $\mathsf{M}(d) \ge \mathsf{m}(d)$ *and*

$$N(d)^{1+\delta_0} \le \frac{1}{\lambda^2_{d,\mathsf{M}(d)+1}} \sum_{k=\mathsf{M}(d)+1}^{\infty} \lambda^2_{d,k}. \tag{20}$$

(b) (Underparametrized regime) We have $n(d) \ge N(d)$ *and*

(i) (Number of features) There exists fixed $\delta_0 > 0$ *such that* $\mathsf{M}(d) \le N(d)^{1-\delta_0}$ *and*

$$\frac{1}{\lambda^2_{d,\mathsf{M}(d)}} \sum_{k=\mathsf{M}(d)+1}^{\infty} \lambda^2_{d,k} \le N(d)^{1-\delta_0} \le N(d)^{1+\delta_0} \le \frac{1}{\lambda^2_{d,\mathsf{M}(d)+1}} \sum_{k=\mathsf{M}(d)+1}^{\infty} \lambda^2_{d,k}\,.$$

(ii) (Number of samples) There exists fixed $\delta_0 > 0$ *such that* $\mathsf{m}(d) \le n(d)^{1-\delta_0}$, $\mathsf{m}(d) \ge \mathsf{M}(d)$ *and*

$$n(d)^{1+\delta_0} \le \cdot \frac{1}{\lambda^2_{d,\mathsf{m}(d)+1}} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda^2_{d,k}\,.$$

The assumption of a spectral gap is useful in that it leads to a clear-cut separation in our main statement below. For instance, in the overparametrized regime $n(d) \ll N(d)$, the projection of the target function onto $\mathcal{D}_{d,\le \mathsf{m}(d)}$ is estimated with negligible error, while the projection onto $\mathcal{D}_{d,>\mathsf{m}(d)}$ is estimated with 0. If there was no spectral gap, the transition would not be as sharp. However, we expect this to affect only target functions with a large projection onto eigenfunctions whose indices are close to $\mathsf{m}(d)$. In this sense, while restrictive, the spectral gap assumption can be in fact a good model for a more generic situation.

2.3. A general theorem

We are now in position to state our main results for random feature ridge regression.

Theorem 1 *(Generalization error of Random Feature Ridge Regression). Let* $\{f_* \in \mathcal{D}_d\}_{d\ge 1}$ *be a sequence of functions,* $\boldsymbol{X} = (\boldsymbol{x}_i)_{i\in[n(d)]}$ *and* $\boldsymbol{\Theta} = (\boldsymbol{\theta}_j)_{j\in[N(d)]}$ *with* $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim \nu_d$ *and* $(\boldsymbol{\theta}_j)_{j\in[N(d)]} \sim \tau_d$ *independently. Let* $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$ *and* $\varepsilon_i \sim_{iid} \mathsf{N}(0, \sigma_\varepsilon^2)$ *for some* $\sigma_\varepsilon > 0$. *Let* $\{\sigma_d\}_{d\ge 1}$ *be a sequence of activation functions satisfying* $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d))\}_{d\ge 1}$*-FMCP (Assumption 1) and spectral gap at level* $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d)))\}_{d\ge 1}$ *(Assumption 2). Then the following hold for the test error of RFRR (see Eq. (17)):*

(a) (Overparametrized regime) If $N(d) \ge d^\delta \cdot n(d)$ *for some* $\delta > 0$, *let* $\lambda_\star$ *be such that* $\lambda_* = O_d(\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}))$. *Then, for any regularization parameter* $\lambda \in [0, \lambda_\star]$, *and any fixed* $\eta > 0$ *and* $\varepsilon > 0$, *with high probability we have*

$$|R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \|\mathsf{P}_{>\mathsf{m}} f_*\|^2_{L^2}| \le \varepsilon \cdot (\|f_*\|^2_{L^2} + \|\mathsf{P}_{>\mathsf{m}} f_*\|^2_{L^{2+\eta}} + \sigma_\varepsilon^2). \tag{21}$$

(b) (Underparametrized regime) If $n(d) \ge d^\delta \cdot N(d)$ *for some* $\delta > 0$, *let* $\lambda_\star$ *be such that* $\lambda_\star = O_d(n/N \cdot \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}))$. *Then, for any regularization parameter* $\lambda \in [0, \lambda_\star]$, *and any fixed* $\eta > 0$ *and* $\varepsilon > 0$, *with high probability we have*

$$|R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \|\mathsf{P}_{>\mathsf{M}} f_*\|^2_{L^2}| \le \varepsilon \cdot (\|f_*\|^2_{L^2} + \|\mathsf{P}_{>\mathsf{M}} f_*\|^2_{L^{2+\eta}} + \sigma_\varepsilon^2). \tag{22}$$

Remark 2.1. The two limits $N = \infty$ and $n = \infty$ play a special role. For $N = \infty$, the random kernel $H_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = N^{-1} \sum_{i=1}^{N} \sigma(\boldsymbol{x}_1; \boldsymbol{\theta}_i)\sigma(\boldsymbol{x}_2; \boldsymbol{\theta}_i)$ converges to its expectation, and we recover KRR. While this case is not technically covered by Theorem 1, we establish the relevant characterization in Theorems 4 and 5.

In the case $n = \infty$ the generalization error vanishes, and we are left with the approximation error. This case is covered separately in Appendix A. In both these limit cases we confirm the result that would have been obtained by naively setting $N = \infty$ or $n = \infty$ in the last theorem.

Notice that the sample size n and the number of neurons N play a nearly symmetric role in this statement, and the smallest of the two determines the test error. An important insight follows: in the present setting, the test error is nearly insensitive to the number of neurons as long as we take $N \gg n$. If we want to minimize computational complexity subject to achieving nearly optimal generalization properties, we should operate, say, at $N \asymp n^{1+\delta}$ for some small $\delta > 0$.

It is instructive to compare this result with [40] which instead suggests $N \asymp \sqrt{n} \log n$. While our setting differs from the one of [40] in a number of technical aspects, we believe that the core difference between the two results lies in the treatment of the target function f_*. Simplifying, the recommendation of [40] is based on two results, the second of which proved in [16] (with an abuse of notation, we indicate the number of neurons and sample size as arguments of $R_{\mathrm{RF}}(f_*) = R_{\mathrm{RF}}(f_*; N, n)$, and use $N = \infty$ to denote the KRR limit case):

$$\sup_{\|f_*\|_{\mathcal{H}} \le r} R_{\mathrm{RF}}(f_*; N_n, n) \le C_1(d) \frac{r^2}{\sqrt{n}}, \qquad \text{for } N_n \asymp \sqrt{n} \log n, \tag{23}$$

$$\sup_{\|f_*\|_{\mathcal{H}} \le r} R_{\mathrm{RF}}(f_*; \infty, n) \le C_2(d) r^2 \Big(\frac{\log n}{n}\Big)^{b/(b+1)}, \tag{24}$$

where $b \in (1, \infty)$ encodes the decay of eigenvalues of the kernel.[1] Now, considering the worst case decay $b \to 1$, the error rate achieved by RFRR, cf. Eq. (23), is of the same order as the one achieved by KRR, cf. Eq. (24).

Note several differences with respect to our results: (*i*) The analysis of [40,16] is minimax, over balls in the RKHS, while our results hold *pointwise*, i.e., for each individual function f_*; (*ii*) Optimality in [40] is established in terms of rates, i.e., up to multiplicative constant, while ours hold up to additive errors (multiplicative constants are exactly characterized); (*iii*) The results of [40,16] apply to a fixed RKHS (in particular, a fixed dimension d), while we study the case in which d is large and N, n, d are polynomially related.

Some of these distinctions are also relevant in comparing our work to other recent results on KRR. In particular points (*i*) and (*ii*) apply when comparing with [29,30].

2.4. Examples: The binary hypercube and the sphere

As examples we consider the case of covariates vectors $\boldsymbol{x}_i$ that are uniformly distributed over the discrete hypercube $\mathscr{Q}^d = \{-1, +1\}^d$ or the sphere $\mathbb{S}^{d-1}(\sqrt{d}) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = d\}$. Namely, letting $\mathcal{A}_d$ to be either $\mathscr{Q}^d$ or $\mathbb{S}^{d-1}(\sqrt{d})$ and $\rho_d = \mathrm{Unif}(\mathcal{A}_d)$, we set $\mathcal{X}_d = \mathcal{A}_d$ and $\nu_d = \rho_d$. We further choose the $\boldsymbol{\theta}_i$'s to be distributed as the covariates vectors, namely $\mathcal{V}_d = \mathcal{A}_d$ and $\tau_d = \rho_d$. Apart from simplifying our analysis, this is a sensible choice: since the covariates vectors do not align along any preferred direction, it is reasonable for the $\boldsymbol{\theta}_i$'s to be isotropic as well.

[1] The results of [16,40] assume the weaker condition that $\inf_{g \in \mathcal{H}} \|f_* - g\|_{L^2}$ is achieved in $\mathcal{H}$: since $\mathcal{H}$ is dense in $L^2(\mathcal{X}_d)$ (provided the kernel is strictly positive definite), this is equivalent to $f_* \in \mathcal{H}$.

Given a function $\bar{\sigma}_d : \mathbb{R} \to \mathbb{R}$ (which we allow to depend on the dimension d), we define the activation function $\sigma_d : \mathcal{A}_d \times \mathcal{A}_d \to \mathbb{R}$ by

$$\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}) = \bar{\sigma}_d(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d}). \tag{25}$$

We denote by $\mathcal{E}_{d,\leq \ell}$ the subspace of $L^2(\mathcal{A}_d, \rho_d)$ spanned by polynomials of degree less or equal to ℓ and by $\overline{\mathsf{P}}_{\leq \ell}$ the orthogonal projection on $\mathcal{E}_{d,\leq \ell}$ in $L^2(\mathcal{A}_d, \rho_d)$. The projectors $\overline{\mathsf{P}}_\ell$ and $\overline{\mathsf{P}}_{>\ell}$ are defined analogously (see Appendix E for more details). Let us emphasize that the projectors $\overline{\mathsf{P}}_{\leq \ell}$ are related but distinct from the $\mathsf{P}_{\leq \mathsf{m}}$: while $\overline{\mathsf{P}}_{\leq \ell}$ projects onto eigenspaces of polynomials of degree at most ℓ, $\mathsf{P}_{\leq \mathsf{m}}$ projects onto the top m-eigenfunctions.[2]

In order to apply Theorem 1, we make the following assumption about $\bar{\sigma}_d$.

Assumption 3 *(Assumptions on $\mathcal{A}_d$ at level $(\mathsf{s}, \mathsf{S}) \in \mathbb{N}^2$). For $\{\bar{\sigma}_d\}_{d \geq 1}$ a sequence of functions $\bar{\sigma}_d : \mathbb{R} \to \mathbb{R}$, we assume the following conditions to hold.*

(a) There exists an integer k and constants $c_1 < 1$ and $c_0 > 0$, $\delta_0 > 1/k$ such that $n \leq N^{1-\delta_0}$ or $N \leq n^{1-\delta_0}$ and $|\bar{\sigma}_d(x)| \leq c_0 \exp(c_1 x^2/(4k))$.

(b) We have

$$\min_{k \leq \mathsf{s}} d^{\mathsf{s}-k} \|\overline{\mathsf{P}}_k \bar{\sigma}_d(\langle \boldsymbol{e}, \cdot \rangle)\|^2_{L^2(\mathcal{A}_d, \rho_d)} = \Omega_d(1), \tag{26}$$

$$\min_{k \leq \mathsf{S}} d^{\mathsf{S}-k} \|\overline{\mathsf{P}}_k \bar{\sigma}_d(\langle \boldsymbol{e}, \cdot \rangle)\|^2_{L^2(\mathcal{A}_d, \rho_d)} = \Omega_d(1), \tag{27}$$

$$\|\overline{\mathsf{P}}_{>2\max(\mathsf{s},\mathsf{S})+1} \bar{\sigma}_d(\langle \boldsymbol{e}, \cdot \rangle)\|_{L^2(\mathcal{A}_d, \rho_d)} = \Omega_d(1), \tag{28}$$

where $\boldsymbol{e} \in \mathcal{A}_d$ is a fixed vector (it is easy to see that these quantities do not depend on $\boldsymbol{e}$).

(c) If $\mathcal{A}_d = \mathscr{Q}^d$, we have, for all d large enough

$$\max_{k \leq 2\max(\mathsf{s},\mathsf{S})+2} d^{-k} \|\overline{\mathsf{P}}_{d-k} \bar{\sigma}_d(\langle \boldsymbol{e}, \cdot \rangle)\|^2_{L^2(\mathcal{A}_d, \rho_d)} \leq d^{-2\max(\mathsf{s},\mathsf{S})-2}. \tag{29}$$

Assumption (a) requires n, N to be well separated and a technical integrability condition. The latter is necessary for the hypercontractivity condition in Assumption 1.(c) to make sense.

Equations (26) and (27) (Assumption (b)) are a quantitative version of a universality condition: if $\overline{\mathsf{P}}_k \bar{\sigma}_d(\langle \boldsymbol{e}, \cdot \rangle / \sqrt{d}) = 0$ for some k, then linear combinations of $\bar{\sigma}_d$ can only span a linear subspace of $L^2(\mathcal{A}_d, \rho_d)$. Equation (28) (Assumption (b)) requires the high degree part of $\bar{\sigma}_d$ to be non-vanishing (and therefore induce a non-zero regularization from the high degree non-linearity).

For $\mathcal{A}_d = \mathscr{Q}^d$, we further require Assumption (c), namely that the last eigenvalues of $\bar{\sigma}_d$ decrease sufficiently fast. This is a necessary conditions to avoid pathological sequences $\{\bar{\sigma}_d\}_{d \geq 1}$ which are very rapidly oscillating.

Remark 2.2. If $\bar{\sigma}_d = \bar{\sigma}$ is independent of the dimension, then Assumptions (b), (c) are easy to check:

- The first two parts of Assumption (b) (Eqs. (26) and (27)) are satisfied if we require $\mathbb{E}\{\bar{\sigma}(G)\, p(G)\} \neq 0$ for all non-vanishing polynomials p of degree at most $\max(\mathsf{s}, \mathsf{S})$ (expectation being taken with respect to $G \sim \mathsf{N}(0, 1)$.) This is in turn equivalent to $\mathbb{E}\{\bar{\sigma}(G)\, \mathrm{He}_k(G)\} \neq 0$ for all $k \leq \max(\mathsf{s}, \mathsf{S})$, where He_k is the k-th Hermite polynomial.

[2] The two coincide if $\mathsf{m} = \sum_{\ell' \leq \ell} B(\mathcal{A}_d; \ell')$, with $B(\mathcal{A}_d; \ell')$ the dimension of the space of degree-ℓ' polynomials and the top m eigenvalues verify $\lambda^2_{d,j} = \Omega_d(d^{-\ell})$, see Appendix D.

- The third part of Assumption (b) (Eq. (28)) amounts to requiring $\bar{\sigma}$ not to be a degree-$(2\max(\mathsf{s},\mathsf{S})+1)$ polynomial.
- In Appendix D.2 we check that Assumption (c) holds if $\bar{\sigma}$ is smooth and there exists $c_0 > 0$ and $c_1 < 1$ constants such that the $(2\max(\mathsf{s},\mathsf{S})+2)$-th derivative verifies $|\bar{\sigma}^{(2\max(\mathsf{s},\mathsf{S})+2)}(x)| \le c_0 \exp(c_1 x^2/4)$.

As an example, the shifted ReLu $\bar{\sigma}_d(x) = (x-c)_+$ with a generic $c \in \mathbb{R}\setminus\{0\}$ verifies Assumption 3. (The case $c = 0$ violates Eq. (26), since $\mathbb{E}\{\bar{\sigma}(G)\mathrm{He}_k(G)\} = 0$ for $k \ge 3$ odd. This is not a limitation of our result: the unshifted ReLU is not universal in the present setting.)

Theorem 2 *(Generalization error of RFRR on the sphere and hypercube). Let $\{f_* \in L^2(\mathcal{A}_d,\rho_d)\}_{d\ge1}$ be a sequence of functions. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{i\in[N]}$ with $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \rho_d$ independently and $\boldsymbol{X} = (\boldsymbol{x}_i)_{i\in[n]}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim \rho_d$ independently. Let $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$ and $\varepsilon_i \sim_{iid} \mathsf{N}(0,\sigma_\varepsilon^2)$ for some $\sigma_\varepsilon > 0$. Assume $d^{\mathsf{s}+\delta_0} \le n \le d^{\mathsf{s}+1-\delta_0}$ and $d^{\mathsf{S}+\delta_0} \le N \le d^{\mathsf{S}+1-\delta_0}$ for fixed integers s,S and for some $\delta_0 > 0$. Let $\{\bar{\sigma}_d\}_{d\ge1}$ satisfy Assumption 3 at level (s,S). Then the following hold for the test error of RFRR (see Eq. (17)):*

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

$$|R_{\mathrm{RF}}(f_*,\boldsymbol{X},\boldsymbol{\Theta},\lambda) - \|\overline{\mathsf{P}}_{>\mathsf{s}} f_*\|_{L^2}^2| \le \varepsilon\cdot(\|f_*\|_{L^2}^2 + \|\overline{\mathsf{P}}_{>\mathsf{s}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{30}$$

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

$$|R_{\mathrm{RF}}(f_*,\boldsymbol{X},\boldsymbol{\Theta},\lambda) - \|\overline{\mathsf{P}}_{>\mathsf{S}} f_*\|_{L^2}^2| \le \varepsilon\cdot(\|f_*\|_{L^2}^2 + \|\overline{\mathsf{P}}_{>\mathsf{S}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{31}$$

As mentioned in the introduction, [20] proves this theorem in the cases $n = \infty$ (RF approximation error) and $N = \infty$ (generalization error of KRR), for the uniform measure on the sphere. The general case follows here as a consequence of Theorem 1.

To see the connection between Theorem 1 and the results given here for the sphere and hypercube cases (see Appendix D for details), notice that the integral operator $\mathbb{T}_d$ associated to the inner product activation function (25) is in this case symmetric, and commutes with rotations in $\mathrm{SO}(d)$ (for the sphere) or with the action of $(\mathbb{Z}_2)^d$ (for the hypercube[3]). Hence, the eigenvectors of $\mathbb{T}_d$ (which is self-adjoint by construction) are given by the spherical harmonics of degree ℓ (for the sphere) or the homogeneous polynomials of degree ℓ (for the hypercube). The spaces spanned by the low degree spherical harmonics and homogeneous polynomials verify the hypercontractivity condition of Assumption 1.(a) (see Appendix E.3). The corresponding distinct eigenvalues are $\xi_{d,\ell}$, with degeneracy

$$B(\mathbb{S}^{d-1};\ell) = \frac{d-2+2\ell}{d-2}\binom{d-3+\ell}{\ell}, \qquad B(\mathscr{Q}^d;\ell) = \binom{d}{\ell}. \tag{32}$$

Notice that in both cases $B(\mathcal{A}_d;\ell) = (d^\ell/\ell!)(1+o_d(1))$ and, hence $\xi_{d,\ell} \lesssim d^{-\ell/2}$ (by construction $\mathrm{Tr}(\mathbb{H}_d)$ is bounded uniformly). Indeed, by Assumption 3.(a), we have $\xi_{d,\ell} \asymp d^{-\ell/2}$.

As a consequence, if we set $\mathsf{m} = \sum_{\ell\le\mathsf{s}} B(\mathcal{A}_d;\ell)$, $\mathsf{M} = \sum_{\ell\le\mathsf{S}} B(\mathcal{A}_d;\ell)$, we have $\sum_{k=\ell+1}^{\infty}\lambda_{k,d}^2 = \Theta(1)$ (indeed this sum is $O_d(1)$ because $\mathrm{Tr}(\mathbb{H}_d)$ is bounded uniformly, and it is $\Omega_d(1)$ by Assumption 3.(b)). Therefore, the conditions (7) and (8) (or, more formally, the conditions in Assumption 2) can be rewritten as

[3] In the $\{+1,-1\}^d$ representation, $\boldsymbol{z} \in \{+1,-1\}^d$ acts on $\mathscr{Q}^d$ via $\boldsymbol{x} \mapsto \boldsymbol{D}_{\boldsymbol{z}}\boldsymbol{x}$, where $\boldsymbol{D}_{\boldsymbol{z}}$ is the diagonal matrix with $\mathrm{diag}(\boldsymbol{D}_{\boldsymbol{z}}) = \boldsymbol{z}$.

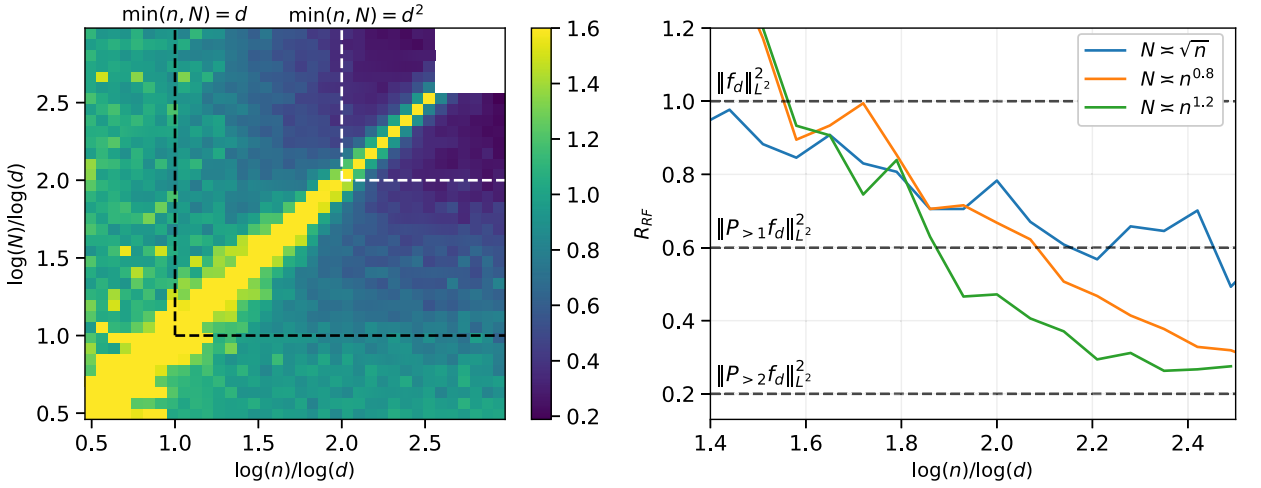

Fig. 1. Learning a polynomial f_* (cf. Eq. (35)) over the d-dimensional sphere, $d = 50$, using a random feature model and min-norm interpolation. We report the test error averaged over 10 realizations. Left: heatmap of the test error as a function of the number of neurons N (y-axis) and number of samples n (x-axis). The dashed lines correspond to $\min(n, N) = d$ (black) and $\min(n, N) = d^2$ (white). Notice the blow-up at the interpolation threshold $N \approx n$, and the symmetry around this line. Right: decrease of the test error as a function of sample size for scalings of the network size $N = n^\alpha$. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.)

$$d^{\mathsf{S}} \asymp \frac{1}{\xi_{\mathsf{S}}^2} \ll N \ll \frac{1}{\xi_{\mathsf{S}+1}^2} \asymp d^{\mathsf{S}+1}\,, \tag{33}$$

$$d^{\mathsf{s}} \asymp \frac{1}{\xi_{\mathsf{s}}^2} \ll n \ll \frac{1}{\xi_{\mathsf{s}+1}^2} \asymp d^{\mathsf{s}+1}\,, \tag{34}$$

which matches the assumptions in Theorem 2.

Fig. 1 provides an illustration of Theorem 2, for the case of the uniform distribution over the sphere $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$. We fix $d = 50$, and generate data $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$ with no noise $\sigma_\varepsilon = 0$. We use the target function

$$f_*(\boldsymbol{x}) = g_d(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\,, \tag{35}$$

where $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and g is a fourth-order polynomial: $g(z) = \frac{2}{\sqrt{5}}\widehat{Q}_1(z) + \frac{2}{\sqrt{5}}\widehat{Q}_2(z) + \frac{1}{\sqrt{10}}\widehat{Q}_3(z) + \frac{1}{\sqrt{10}}\widehat{Q}_4(z)$ (here $\widehat{Q}_\ell$ is the ℓ-th Gegenbauer polynomial, normalized so that $\|\widehat{Q}_\ell(\langle \boldsymbol{v}, \cdot\rangle)\|_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = 1$). While the precise form of f_* does not really matter here, we note that $\|\overline{\mathsf{P}}_1 f_*\|_{L^2}^2 = \|\overline{\mathsf{P}}_2 f_*\|_{L^2}^2 = 0.4$, $\|\overline{\mathsf{P}}_3 f_*\|_{L^2}^2 = \|\overline{\mathsf{P}}_4 f_*\|_{L^2}^2 = 0.1$ and $\|\overline{\mathsf{P}}_{>4} f_*\|_{L^2}^2 = 0$. We plot the test error of RFRR using $\sigma(x) = \max(x - 0.5, 0)$ (shifted ReLu), and $\lambda = 0+$ (min-norm interpolation). We repeat this calculation for a grid of values of n, N, and for each point in the grid report the average risk over 10 realizations.

We plot the observed average risk in the sample-size/number-of-parameters plane with axes $\log n/\log d$ and $\log N/\log d$ (corresponding to the exponents in the polynomial relation between n and d, and between N and d). Several prominent features of this plot are worth of note:

- The risk has a large peak for $N \approx n$. This phenomenon was characterized precisely in the proportional regime $N \asymp d$, $n \asymp d$ in [23,31].
- The plot appears completely symmetric under exchange of N and n: the number of parameters and sample size plays the same role in limiting the generalization abilities, as anticipated by Theorem 1 and Theorem 2.
- The risk is bounded away from zero even for $N, n \asymp d^3$. Indeed, Theorem 2 implies that consistent estimation would require $N, n \gg d^4$ in this case.

- Finally, for a fixed n, near optimal test error is achieved when $N \asymp n^{1+\delta_*}$, for δ_* a small positive constant.

2.5. Invariant function estimation

Many predictive tasks of practical interest present important symmetry properties. Namely, the 'true' label $f_*(\boldsymbol{x})$ does not change when a certain group of transformations is applied to the covariates vector $\boldsymbol{x}$. The goal is then to construct predictive models that exploit such invariances. As a further application of our general theory, we discuss here the case of invariant random feature and kernel methods. We refer to [32] for a more complete treatment.

We focus on the case of the cyclic group $\mathrm{Cyc}_d = \{g_0, \dots, g_{d-1}\}$, where g_i shifts the covariates vector by i coordinates. Namely, for any $\boldsymbol{x} = (x_1, \dots, x_d)^{\mathsf{T}} \in \mathbb{R}^d$, the action of the group element g_i on $\boldsymbol{x}$ is defined by $g_i \cdot \boldsymbol{x} = (x_{i+1}, x_{i+2}, \dots x_d, x_1, x_2, \dots, x_i)^{\mathsf{T}}$. We take again $\boldsymbol{x}$ uniformly distributed on $\mathcal{A}_d$, the sphere or the discrete hypercube (in particular the action of Cyc_d preserves $(\mathcal{A}_d, \rho_d)$). The goal is to fit data where the target function $f_\star$ is invariant under the action of Cyc_d, i.e., $f_\star$ belongs to the 'cyclic functions' class

$$L^2(\mathcal{A}_d, \mathrm{Cyc}_d) := \left\{ f \in L^2(\mathcal{A}_d, \rho_d) : f(g \cdot \boldsymbol{x}) = f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{A}_d, \forall g \in \mathrm{Cyc}_d \right\}.$$

For example, one can think about an image recognition task where $f_\star \in L^2(\mathcal{A}_d, \mathrm{Cyc}_d)$ is a label that is invariant by translation of the (one-dimensional) image $\boldsymbol{x}$.

We define a cyclic activation function $\sigma_d : \mathcal{A}_d \times \mathcal{A}_d \to \mathbb{R}$ as follows: given a function $\bar{\sigma} : \mathbb{R} \to \mathbb{R}$ (that we take here independent of d),

$$\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{d} \sum_{\ell=0}^{d-1} \bar{\sigma}(\langle g_\ell \cdot \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d}) \,. \tag{36}$$

Following the notations of Section 2.1, we have $\mathcal{D}_d = \mathcal{V}_d = L^2(\mathcal{A}_d, \mathrm{Cyc}_d)$ which is a closed linear subspace of $L^2(\mathcal{A}_d, \rho_d)$. Notice that $\hat{f} \in \mathcal{F}_{\mathrm{RF},N}(\boldsymbol{\Theta})$ can be written as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{d} \sum_{i=1}^{N} a_i \sum_{\ell=1}^{d} \bar{\sigma}(\langle g_\ell \cdot \boldsymbol{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d}) \,.$$

In neural networks jargon, this corresponds to fitting the second layer weights of a two-layer convolutional network with a nonlinear convolution of N filters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N \in \mathbb{R}^d$ with the image $\boldsymbol{x}$ followed by global average pooling. In contrast, the inner-product activation (25) corresponds to fitting the last layer of a two-layer fully connected network.

The following result follows by verifying that the assumptions of Theorem 1 are verified by this invariant model, which is done in [32, Theorem 1].

Theorem 3 *(Generalization error of RFRR with cyclic activation [32]). Let $\{f_* \in L^2(\mathcal{A}_d, \mathrm{Cyc}_d)\}_{d \geq 1}$ be a sequence of cyclic functions. Assume $d^{\mathsf{s}-1+\delta} \leq n \leq d^{\mathsf{s}-\delta}$ and $d^{\mathsf{S}-1+\delta} \leq N \leq d^{\mathsf{S}-\delta}$ for fixed integers s, S and some $\delta > 0$. Let $\bar{\sigma}$ be a function that satisfies some smoothness condition at level (s, S) [32, Assumption 1]. Then the following hold for the test error of RFRR with cyclic activation (36):*

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

$$R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathsf{P}_{>\mathsf{s}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{37}$$

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

$$R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathsf{P}_{>\mathsf{S}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{38}$$

The smoothness assumption on σ is somewhat technical, and we consider it mainly a proof artifact. It is satisfied —for instance— by smooth versions of the ReLU activation, e.g. $\sigma(x) = \mathbb{E}\{(x - b - \varepsilon G)_+\}$, where expectation is over $G \sim \mathsf{N}(0,1)$, $b \neq 0$ is a fixed shift and $\varepsilon > 0$ is a smoothing parameter that can be taken arbitrarily small.

Theorem 3 can be contrasted to Theorem 2: to achieve the same test error, RFRR with inner-product activation function needs $d^{\mathsf{s}+\delta} \leq n \leq d^{\mathsf{s}+1-\delta}$ and $d^{\mathsf{S}+\delta} \leq N \leq d^{\mathsf{S}+1-\delta}$. Hence, we gain a factor d in sample and feature complexity by using a cyclic-invariant activation compared to an inner-product activation function. This gain can be understood using the following observation: the subspaces $V_{d,k}$ of degree-k polynomials (which are eigenspaces of the inner-product activation with eigenvalues $\xi_{d,k}$ and degeneracies $B(\mathcal{A}_d; k) = \dim(V_{d,k})$) are preserved under the cyclic group Cyc_d. Hence the cyclic activation has eigenspaces $V_{d,k}(\mathrm{Cyc}_d)$ (the subspace of cyclic-invariant polynomials of degree-k) with same eigenvalues $\xi_{d,k}$ and new degeneracies $D(\mathcal{A}_d; k) := \dim(V_{d,k}(\mathrm{Cyc}_d)) = \Theta_d(d^{-1}) \cdot B(\mathcal{A}_d; k)$. Setting $\mathsf{m} = \sum_{\ell \leq \mathsf{s}} D(\mathcal{A}_d; \ell)$ and $\mathsf{M} = \sum_{\ell \leq \mathsf{S}} D(\mathcal{A}_d; \ell)$, we have $\sum_{k=r+1}^{\infty} \lambda_{k,d}^2 = \Theta_d(d^{-1})$ with $r = \mathsf{m}$ or M. Injecting these bounds in Assumption 2 yields a factor d improvement in n and N.

In fact, [32] considers more general invariance groups $\mathcal{G}_d$ called of 'degeneracy α', with $\alpha \leq 1$, and shows that RFRR with $\mathcal{G}_d$-invariant activations gains a d^{α} factor in sample size and number of features with respect to RFRR with inner-product activations. The cyclic group Cyc_d is an example of a degeneracy 1 group, and so is the group of shifts on 2-dimensional images.

3. Generalization error of kernel machines

Formally, kernel ridge regression (KRR) corresponds to the limit $N \to \infty$ of random feature ridge regression. Despite this, we cannot apply directly Theorem 1 with $N = \infty$. We state therefore a separate theorem for kernel methods. As a side benefit, we establish somewhat stronger results in this case. In particular:

- We simplify the set of assumptions (in particular, the assumptions concern only H_d and not the activation function σ_d, as they should).
- We prove a risk lower bound, Theorem 4, that holds for general kernel methods, not only KRR.
- Crucially, we remove the spectral gap assumption. In this more general setting, the risk of KRR is not approximated by the square norm of the projection of f_* orthogonal to the leading eigenfunctions of the kernel. We instead obtain an approximation in terms of a population-level ridge regression problem, with an effective value of the regularization parameter, which we determine.

Throughout this section, the setting is the same as in the previous one: we observe i.i.d. data $(y_i, \boldsymbol{x}_i)_{i \in [n]}$, with feature vectors $\boldsymbol{x}_i$ from the probability space $(\mathcal{X}_d, \nu_d)$. Responses are given by $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, $f_* \in \mathcal{D}_d$ and $\varepsilon_i \sim \mathsf{N}(0, \sigma_\varepsilon^2)$ independently of $\boldsymbol{x}_i$.

We introduce some general background in Section 3.1, then state our assumptions in Section 3.2, and formally state our results in Sections 3.3 and 3.4.

3.1. Background on kernel methods

We consider a general RKHS defined on the probability space $(\mathcal{X}_d, \nu_d)$, via $\mathbb{H}_d$ a compact self-adjoint positive definite operators: $\mathbb{H}_d : \mathcal{D}_d \to \mathcal{D}_d$ with kernel representation

$$\mathbb{H}_d g(\boldsymbol{x}_1) = \int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \nu_d(\mathrm{d}\boldsymbol{x}') \,,$$

where $H_d : \mathcal{X}_d \times \mathcal{X}_d \to \mathbb{R}$ is a square integrable function $H_d \in L^2(\mathcal{X}_d \times \mathcal{X}_d)$, with the property that $\int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \nu_d(\mathrm{d}\boldsymbol{x}') = 0$ for $g \in \mathcal{D}_d^{\perp}$.

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$ a general kernel method learns the function

$$\hat{f}_\lambda = \arg\min_f \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\} , \tag{39}$$

where $\|f\|_{\mathcal{H}}$ is the RKHS norm associated to H_d. Kernel ridge regression corresponds to the special case $\ell(y, \hat{y}) = (y - \hat{y})^2$. As before, we will evaluate kernel methods via their test error, which we denote as follows in the case of KRR

$$R_{\mathrm{KR}}(f_*, \boldsymbol{X}, \lambda) := \mathbb{E}_{\boldsymbol{x}} \Big[\Big(f_*(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x}) \Big)^2 \Big] . \tag{40}$$

As mentioned above, any kernel method can be seen as the $N \to \infty$ limit of a RF model. To see this, note any positive semidefinite kernel can be written in the form $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{\theta} \sim \tau_d}[\sigma_d(\boldsymbol{x}_1, \boldsymbol{\theta}) \sigma_d(\boldsymbol{x}_2, \boldsymbol{\theta})]$, for some activation function σ_d, and some probability space (Ω_d, τ_d). This is akin to taking the square root of a matrix and —as in the finite-dimensional case— the square root is not unique. For instance, we can let σ_d be the symmetric square root $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ replacing $\lambda_{d,j}^2$ by $\lambda_{d,j}$ in Eq. (15).

Given a choice of this square root, we can rewrite the estimator (39) as $\hat{f}_\lambda(\boldsymbol{x}) = f(\boldsymbol{x}; \hat{a}_\lambda)$, where $\hat{a}_\lambda \in L^2(\Omega_d; \nu_d)$ and

$$\hat{a}_\lambda = \arg\min_a \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \right\} , \tag{41}$$

$$f(\boldsymbol{x}; a) := \int \sigma_d(\boldsymbol{x}; \boldsymbol{\theta})\, a(\boldsymbol{\theta})\, \tau_d(\mathrm{d}\boldsymbol{\theta}) \,. \tag{42}$$

This can be informally seen as the $N \to \infty$ limit of Eq. (16) if we choose the square loss function.

3.2. Assumptions on the kernel

As for the case of RFRR, we collect our assumptions in two groups. The first one is mainly concerned with the concentration properties of the kernel, which are quantified in terms of the sequences of integers $n(d)$, $\mathsf{m}(d)$.

Assumption 4 *($\{n(d), \mathsf{m}(d)\}_{d \ge 1}$-Kernel Concentration Property). We say that the sequence of operators $\{\mathbb{H}_d\}_{d \ge 1}$ satisfies the Kernel Concentration Property (KCP) with respect to the sequence $\{(n(d), \mathsf{m}(d))\}_{d \ge 1}$ if there exists a sequence of integers $\{u(d)\}_{d \ge 1}$ with $u(d) \ge \mathsf{m}(d)$ such that the following conditions hold.*

(a) (Hypercontractivity of finite eigenspaces.) For any fixed $q \ge 1$, there exists a constant C such that, for any $h \in \mathcal{D}_{d, \le u(d)} = \mathrm{span}(\psi_s, 1 \le s \le u(d))$, we have

$$\|h\|_{L^{2q}} \le C \cdot \|h\|_{L^2}. \tag{43}$$

(b) (Properly decaying eigenvalues.) There exists fixed $\delta_0 > 0$, such that, for all d large enough,

$$n(d)^{2+\delta_0} \le \frac{\left(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4\right)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^8}, \tag{44}$$

$$n(d)^{2+\delta_0} \le \frac{\left(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2\right)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4}. \tag{45}$$

(c) (Concentration of diagonal elements of kernel) For $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$, we have:

$$\max_{i\in[n(d)]} \left|\mathbb{E}_{\boldsymbol{x}\sim\nu_d}\left[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x})^2\right] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}\left[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x}')^2\right]\right| = o_{d,\mathbb{P}}(1)\cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}\left[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x}')^2\right], \tag{46}$$

$$\max_{i\in[n(d)]} \left|H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x})]\right| = o_{d,\mathbb{P}}(1)\cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x})]. \tag{47}$$

In the last definition, assumptions (a) and (c) have an interpretation that is similar to the one for RFRR. Namely, assumption (a) requires that the top eigenvectors of $\mathbb{H}_d$ are delocalized, and assumption (c) requires that 'most points' in the sample space $\mathcal{X}_d$ behave similarly, in the sense of having similar values of the kernel diagonal $H_d(\boldsymbol{x},\boldsymbol{x})$. Condition (b) is very mild in high dimension, and concerns the tail of eigenvalues of $\mathbb{H}_d$.

The next condition essentially connects the sample size $n(d)$ to the eigenvalue index $\mathsf{m}(d)$, via the eigenvalue sequence.

Assumption 5 *(Eigenvalue condition at level $\{(n(d),\mathsf{m}(d))\}_{d\ge1}$). We say that the sequence of Kernel operators $\{\mathbb{H}_d\}_{d\ge1}$ satisfies the Eigenvalue Condition at level $\{(n(d),\mathsf{m}(d))\}_{d\ge1}$ if the following conditions hold for all d large enough.*

(a) There exists fixed $\delta_0 > 0$, such that

$$n(d)^{1+\delta_0} \le \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^4} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^4, \tag{48}$$

$$n(d)^{1+\delta_0} \le \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2. \tag{49}$$

(b) There exists fixed $\delta_0 > 0$, such that

$$\mathsf{m}(d) \le n(d)^{1-\delta_0}.$$

Unlike in the case of RFRR, we do not require the existence of a spectral gap, but we assume two different upper bounds $n(d)$ to hold simultaneously. In many cases of interest, the right hand sides of (48) and (49) have roughly the same value, which is given by the number of eigenvalues between $\lambda_{d,\mathsf{m}(d)+1}$ and $c_0\lambda_{d,\mathsf{m}(d)+1}$ for a small c_0 (counting degeneracy). The technical requirement (b) is mild and we do not know of any interesting counterexample.

3.3. Lower bound for general kernel methods

Consider any regression method of the form (39). By the representer theorem, there exist coefficients $\hat{\zeta}_1, \dots, \hat{\zeta}_n$ such that

$$\hat{f}_\lambda(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\zeta}_i \, H_d(\boldsymbol{x}, \boldsymbol{x}_i) \, . \tag{50}$$

We are therefore led to define the following data-dependent prediction risk function for kernel methods

$$R_H(f_*, \boldsymbol{X}) := \min_{\boldsymbol{\zeta}} \mathbb{E}_{\boldsymbol{x}} \Big\{ \Big(f_*(\boldsymbol{x}) - \sum_{i=1}^{n} \zeta_i H_d(\boldsymbol{x}_i, \boldsymbol{x}) \Big)^2 \Big\}. \tag{51}$$

This is a lower bound on the prediction error of any kernel methods of the form (39).

The next theorem provides a lower bound on the generalization of kernel methods that is a consequence of the approximation bound in Theorem 6.(a) derived for the random feature model, in Appendix A.

Theorem 4. *Let $\{f_* \in \mathcal{D}_d\}_{d \ge 1}$ be a sequence of functions, $(\boldsymbol{x}_i)_{i \in [n(d)]} \sim \nu_d$ independently, $\{\mathbb{H}_d\}_{d \ge 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, n(d), \mathsf{m}(d))\}_{d \ge 1}$ satisfies Eqs. (43), (44), (46), and (48). Then we have (cf. Eq. (51))*

$$\Big| R_H(f_*, \boldsymbol{X}) - R_H(\mathsf{P}_{\le \mathsf{m}(d)} f_*, \boldsymbol{X}) - \|\mathsf{P}_{> \mathsf{m}(d)} f_*\|_{L^2}^2 \Big| \le o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2} \|\mathsf{P}_{> \mathsf{m}(d)} f_*\|_{L^2}. \tag{52}$$

Proof. This follows immediately from Theorem 6 (a) stated in Appendix A. Indeed, setting $\sigma_d(\boldsymbol{x}, \boldsymbol{x}') = H_d(\boldsymbol{x}, \boldsymbol{x}')$, we obtain $R_H(f_*, \boldsymbol{X}) = R_{\rm RF}(f_*, \boldsymbol{X})$, whence the claim follows by applying Eq. (60). □

Notice that $R_H(\mathsf{P}_{\le \mathsf{m}(d)} f_*, \boldsymbol{X}) \ge 0$ by construction and therefore this theorem immediately implies a lower bound on the test error of kernel ridge regression (cf. Eq. (40))

$$R_{\rm KR}(f_*; \boldsymbol{X}, \lambda) \ge R_H(f_*, \boldsymbol{X}) \ge \|\mathsf{P}_{> \mathsf{m}(d)} f_*\|_{L^2}^2 - o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2} \|\mathsf{P}_{> \mathsf{m}(d)} f_*\|_{L^2}. \tag{53}$$

In words, if we neglect the error term $o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2} \|\mathsf{P}_{> \mathsf{m}(d)} f_*\|_{L^2}$, no kernel method can achieve non-trivial accuracy on the projection of f_* onto eigenvectors beyond the first $\mathsf{m}(d)$.

3.4. The risk of kernel ridge regression

Kernel ridge regression is one specific way of selecting the coefficients $\hat{\boldsymbol{\zeta}}$ in Eq. (50), namely by using $\ell(\hat{y}, y) = (\hat{y} - y)^2$ in Eq. (39). Solving for the coefficients yields

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

where the kernel matrix $\boldsymbol{H} = (H_{ij})_{ij \in [n]}$ is given by $H_{ij} = H_d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\boldsymbol{y} = (y_1, \dots, y_n)^{\mathsf{T}}$.

It is convenient to state our main results in terms of an effective ridge regression estimator

$$\hat{f}_\gamma^{\rm eff} = \arg\min_f \Big\{ \|f_* - f\|_{L^2}^2 + \frac{\gamma}{n} \|f\|_{\mathcal{H}}^2 \Big\} \, . \tag{54}$$

This amounts to replacing the empirical risk in Eq. (39) by its population counterpart $\|f_* - f\|_{L^2}^2 = \mathbb{E}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\}$. Also note that the regularization parameter does not coincide with λ: its precise value will be specified below.

The solution of the population ridge problem (54) can be explicitly written in terms of a shrinkage operator in the basis of eigenfunctions of $\mathbb{H}_d$:

$$f(\boldsymbol{x}) = \sum_{\ell=1}^{\infty} c_\ell \psi_{d,\ell}(\boldsymbol{x}) \quad \mapsto \quad \hat{f}_\gamma^{\rm eff}(\boldsymbol{x}) = \sum_{\ell=1}^{\infty} \frac{\lambda_{d,\ell}^2}{\lambda_{d,\ell}^2 + \frac{\gamma}{n}} c_\ell \psi_{d,\ell}(\boldsymbol{x}) \,. \tag{55}$$

Theorem 5. *Let $\{f_* \in \mathcal{D}_d\}_{d\ge 1}$ be a sequence of functions, $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim \nu_d$ independently, $\{\mathbb{H}_d\}_{d\ge 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, n(d), \mathsf{m}(d))\}_{d\ge 1}$ satisfies $\{n(d), \mathsf{m}(d)\}_{d\ge 1}$-KPCP (Assumption 4) and eigenvalue condition at level $\{n(d), \mathsf{m}(d)\}_{d\ge 1}$ (Assumption 5). Define the effective regularization*

$$\gamma^{\rm eff} := \lambda + \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}(d)}) \,. \tag{56}$$

Then, for any regularization parameter $\lambda \in [0, \lambda_\star]$ where $\lambda_\star = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}(d)})$, any $\eta > 0$, we have (cf. Eq. (40))

$$\left| R_{\rm KR}(f_*, \boldsymbol{X}, \lambda) - \|f_* - \hat{f}_{\gamma^{\rm eff}}^{\rm eff}\|_{L^2} \right| = o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{57}$$

Further, the ridge regression estimator $\hat{f}_\lambda$ is close to the effective estimator $\hat{f}_{\gamma^{\rm eff}}^{\rm eff}$, namely

$$\left\| \hat{f}_\lambda - \hat{f}_{\gamma^{\rm eff}}^{\rm eff} \right\|_{L^2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{58}$$

The proof of Theorem 5 is deferred to Appendix C.

In words, KRR behaves as ridge regression with respect to the population risk, except that the regularization parameter is increased by $\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$. The underlying mechanism is quite simple. The empirical kernel matrix is decomposed as $\boldsymbol{H} = \boldsymbol{H}_{\le \mathsf{m}} + \boldsymbol{H}_{>\mathsf{m}}$, and the second component can be approximated by a multiple of the identity: $\boldsymbol{H}_{>\mathsf{m}} \approx \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) \cdot \mathbf{I}_n$. This term acts as an additional ridge regularizer.

As mentioned above, we do not assume here any eigenvalue gap condition. However, formulas simplify if we assume an eigenvalue gap, e.g.:

$$n(d) = \omega_d(1) \cdot \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2 \,.$$

Under this additional assumption, Theorem 5 implies the following simplified formula for the test error:

$$\left| R_{\rm KR}(f_*, \boldsymbol{X}, \lambda) - \|\mathsf{P}_{>\mathsf{m}(d)} f_*\|_{L^2}^2 \right| = o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

As anticipated, this coincides with the risk of RFRR, if we heuristically set $N = \infty$ in Theorem 1.

Acknowledgments

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

Appendix A. Approximation error of random feature model

In this section, we consider the approximation error of the random feature function class. Formally, the approximation error can be seen as the generalization error of random feature ridge regression for finite

number of neurons $N < \infty$ and infinite data $n = \infty$. However, we cannot apply directly Theorem 1 with $n = \infty$. We therefore state a separate theorem. This is also used to prove the lower bound of Theorem 4 on the generalization error of general kernel methods.

In Section A.1, we state our assumptions and theorem. Sections A.2 and A.3 provide a proof of the theorem, while Section A.4 gathers key technical concentration results that will also be used in the proofs of Theorem 1 and Theorem 5.

A.1. Assumptions and theorem

Recall the definition of the random feature function class (see Section 2.1): let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{i\in[N]} \sim_{iid} \tau_d$,

$$\mathcal{F}_{\mathrm{RF},N}(\boldsymbol{\Theta}) = \Big\{\hat{f}(\boldsymbol{x};\boldsymbol{a}) = \sum_{i=1}^{N} a_i \sigma_d(\boldsymbol{x};\boldsymbol{\theta}_i) : a_i \in \mathbb{R}, i \in [N]\Big\}.$$

We define the approximation error of the random feature function class for a target function $f_* \in L^2(\mathcal{X}_d)$ as

$$R_{\mathrm{App}}(f_*, \boldsymbol{\Theta}) := \inf_{\hat{f}\in\mathcal{F}_{\mathrm{RF},N}(\boldsymbol{\Theta})} \mathbb{E}_{\boldsymbol{x}\sim\tau_d}[(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2]. \tag{59}$$

Similarly to Sections 2.2 and 3.2, we will quantify our assumptions on the sequences of probability spaces $(\mathcal{X}_d, \nu_d)$ and (Ω_d, τ_d), and on the activation functions $\sigma_d \in L^2(\mathcal{X}_d \times \Omega_d)$, in terms of the sequences of integers $N(d), \mathsf{M}(d)$. We state the assumptions in two groups: Assumption 6 and Assumption 7 deal respectively with the concentration properties and the spectrum of the sequence of feature kernel operators $\{U_d\}_{d\geq 1}$ defined as

$$U_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[\sigma_d(\boldsymbol{x};\boldsymbol{\theta}_1)\sigma_d(\boldsymbol{x};\boldsymbol{\theta}_2)].$$

Assumption 6 *(Feature kernel concentration at level $\{(N(d), \mathsf{M}(d))\}_{d\geq 1}$). The sequences of spaces $\{\mathcal{V}_d\}_{d\geq 1}$, operators $\{\mathbb{U}_d\}_{d\geq 1}$ and numbers of neurons $\{N(d)\}_{d\geq 1}$ satisfy feature kernel concentration at level $\{\mathsf{M}(d)\}_{d\geq 1}$ if there exists a sequence $\{u(d)\}_{d\geq 1}$ with $u(d) \geq \mathsf{M}(d)$, such that the following hold.*

(a) (Hypercontractivity of finite eigenspaces.) For any fixed $q \geq 1$, there exists C such that, for any $g \in \mathcal{V}_{d,\leq u(d)} = \mathrm{span}(\phi_s, 1 \leq s \leq u(d))$, we have

$$\|g\|_{L^{2q}(\Omega_d)} \leq C \cdot \|g\|_{L^2(\Omega_d)}.$$

(b) (Properly decaying eigenvalues.) There exists a fixed $\delta_0 > 0$, such that

$$N(d)^{2+\delta_0} \leq \frac{\Big(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2\Big)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4}.$$

(c) (Upper bound on the diagonal elements of the kernel) For $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim_{iid} \tau_d$ and any $\delta > 0$, we have

$$\max_{i\in[N(d)]} U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}}[U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}, \boldsymbol{\theta})].$$

(d) (Lower bound on the diagonal elements of the kernel) For $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim_{iid} \tau_d$ and any $\delta > 0$, we have

$$\min_{i\in[N(d)]} U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) = \Omega_{d,\mathbb{P}}(N(d)^{-\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}}[U_{d,>\mathsf{M}(d)}(\boldsymbol{\theta}, \boldsymbol{\theta})].$$

Assumption 7 *(Spectral gap at level $\{(N(d), \mathsf{M}(d))\}_{d\geq 1}$). The sequence of operators $\{\mathbb{U}_d\}_{d\geq 1}$ has a spectral gap at level $\{(N(d), \mathsf{M}(d))\}_{d\geq 1}$ if the following hold.*

(a) There exists a fixed $\delta_0 > 0$, such that

$$N(d)^{1+\delta_0} \leq \frac{1}{\lambda^2_{d,\mathsf{M}(d)+1}} \sum_{j=\mathsf{M}(d)+1}^{\infty} \lambda^2_{j,d}\,.$$

(b) There exists a fixed $\delta_0 > 0$, such that $\mathsf{M}(d) \leq N(d)^{1-\delta_0}$ and

$$N(d)^{1-\delta_0} \geq \frac{1}{\lambda^2_{d,\mathsf{M}(d)}} \sum_{j=\mathsf{M}(d)+1}^{\infty} \lambda^2_{j,d}\,.$$

Remark A.1. In Assumption 6.(*c*), we can replace $U_{d,>\mathsf{M}(d)}$ by $U_{d,>u(d)}$ (see Lemma 7).

We are now in position to state our theorem on the approximation error of the random feature function class. We state the lower and upper bounds and their assumptions separately.

Theorem 6 *(Approximation error of the random feature function class). Let $\{f_* \in \mathcal{D}_d\}_{d\geq 1}$ be a sequence of functions and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{i\in[N(d)]}$ with $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim \tau_d$ independently. Let $\{\sigma_d\}_{d\geq 1}$ be a sequence of activation functions satisfying Assumptions 6.(a) and 6.(b) at level $\{\mathsf{M}(d)\}_{d\geq 1}$. Then the following hold for the approximation error of the random feature class (see Eq. (59)):*

(a) (Lower bound) If $\{\sigma_d\}_{d\geq 1}$ satisfies further Assumptions 6.(d) and 7.(a), then we have

$$\left|R_{\rm App}(f_*, \boldsymbol{\Theta}) - R_{\rm App}(\mathsf{P}_{\leq \mathsf{M}(d)} f_*, \boldsymbol{\Theta}) - \|\mathsf{P}_{>\mathsf{M}(d)} f_*\|^2_{L^2}\right| \leq o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2} \|\mathsf{P}_{>\mathsf{M}(d)} f_*\|_{L^2}. \tag{60}$$

(b) (Upper bound) If $\{\sigma_d\}_{d\geq 1}$ satisfies further Assumptions 6.(c) and 7, then we have

$$\left|R_{\rm App}(\mathsf{P}_{\leq \mathsf{M}(d)} f_*, \boldsymbol{\Theta})\right| \leq o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2} \|\mathsf{P}_{\leq \mathsf{M}(d)} f_*\|_{L^2}. \tag{61}$$

Point (*a*) is proved in Section A.2, while point (*b*) is proved in Section A.3.

The lower bound on general kernel methods in Theorem 4 is obtained as a direct consequence of Theorem 6.(a), by taking $\sigma_d(\boldsymbol{x}, \boldsymbol{x}') = H_d(\boldsymbol{x}, \boldsymbol{x}')$. Indeed, it is easy to check that Eqs. (43) and (45) imply Assumptions 6.(a) and 6.(*b*), Eq. (47) implies Assumptions 6.(*c*) and 6.(*d*), and Eq. (49) implies Assumption 7.(a).

A.2. Proof of Theorem 6.(a): lower bound on the approximation error

Let $\mathbb{E}_{\boldsymbol{\theta}}$ denote the expectation operator with respect to $\boldsymbol{\theta} \sim \tau_d$, $\mathbb{E}_{\boldsymbol{x}}$ to be the expectation operator with respect to $\boldsymbol{x} \sim \nu_d$. We will denote $\mathsf{M} = \mathsf{M}(d)$ and $N = N(d)$.

Define the random vector $\boldsymbol{V} = (V_1, \dots, V_N)^{\mathsf{T}}$ and its low- and high- degree parts given respectively by $\boldsymbol{V}_{\leq \mathsf{M}} = (V_{1,\leq \mathsf{M}}, \dots, V_{N,\leq \mathsf{M}})^{\mathsf{T}}$ and $\boldsymbol{V}_{>\mathsf{M}} = (V_{1,>\mathsf{M}}, \dots, V_{N,>\mathsf{M}})^{\mathsf{T}}$, with

$$\begin{aligned}
V_{i,\leq \mathsf{M}} &\equiv \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[[\mathsf{P}_{\leq \mathsf{M}} f_*](\boldsymbol{x}) \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)],\\
V_{i,>\mathsf{M}} &\equiv \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[[\mathsf{P}_{>\mathsf{M}} f_*](\boldsymbol{x}) \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)],\\
V_i &\equiv \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[f_*(\boldsymbol{x}) \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)] = V_{i,\leq \mathsf{M}} + V_{i,>\mathsf{M}}.
\end{aligned}$$

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

$$U_{ij} = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[\sigma_d(\boldsymbol{x};\boldsymbol{\theta}_i)\sigma_d(\boldsymbol{x};\boldsymbol{\theta}_j)]. \tag{62}$$

In what follows, we write $R_{\rm App}(f_*) = R_{\rm App}(f_*, \boldsymbol{\Theta})$ for the approximation error of the random feature model, omitting the dependence on the weights $\boldsymbol{\Theta}$. By definition and a simple calculation, we have

$$\begin{aligned}
R_{\rm App}(f_*) &= \min_{\boldsymbol{a}\in\mathbb{R}^N}\Big\{\mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})^2] - 2\langle \boldsymbol{a}, \boldsymbol{V}\rangle + \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\Big\} = \mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})^2] - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V},\\
R_{\rm App}(\mathsf{P}_{\le \mathsf{M}} f_*) &= \min_{\boldsymbol{a}\in\mathbb{R}^N}\Big\{\mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_{\le \mathsf{M}} f_*(\boldsymbol{x})^2] - 2\langle \boldsymbol{a}, \boldsymbol{V}_{\le \mathsf{M}}\rangle + \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\Big\} = \mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_{\le \mathsf{M}} f_*(\boldsymbol{x})^2] - \boldsymbol{V}_{\le \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le \mathsf{M}}.
\end{aligned}$$

By orthogonality, we have

$$\mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})^2] = \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{\le \mathsf{M}} f_*](\boldsymbol{x})^2] + \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{> \mathsf{M}} f_*](\boldsymbol{x})^2],$$

which gives

$$\begin{aligned}
&\Big|R_{\rm App}(f_*) - R_{\rm App}(\mathsf{P}_{\le \mathsf{M}} f_*) - \mathbb{E}_{\boldsymbol{x}}[[\mathsf{P}_{> \mathsf{M}} f_*](\boldsymbol{x})^2]\Big|\\
=&\Big|\boldsymbol{V}_{\le \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le \mathsf{M}} - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}\Big| = \Big|\boldsymbol{V}_{\le \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le \mathsf{M}} - (\boldsymbol{V}_{\le \mathsf{M}} + \boldsymbol{V}_{> \mathsf{M}})^{\mathsf{T}}\boldsymbol{U}^{-1}(\boldsymbol{V}_{\le \mathsf{M}} + \boldsymbol{V}_{> \mathsf{M}})\Big|\\
=&\Big|2\boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{> \mathsf{M}} - \boldsymbol{V}_{> \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{> \mathsf{M}}\Big| \le 2\|\boldsymbol{U}^{-1/2}\boldsymbol{V}_{> \mathsf{M}}\|_2\|\boldsymbol{U}^{-1/2}\boldsymbol{V}\|_2 + \|\boldsymbol{U}^{-1}\|_{\rm op}\|\boldsymbol{V}_{> \mathsf{M}}\|_2^2\\
\le&2\|\boldsymbol{U}^{-1/2}\|_{\rm op}\|\boldsymbol{V}_{> \mathsf{M}}\|_2\|f_*\|_{L^2} + \|\boldsymbol{U}^{-1}\|_{\rm op}\|\boldsymbol{V}_{> \mathsf{M}}\|_2^2,
\end{aligned} \tag{63}$$

where the last inequality used the fact that

$$0 \le R_{\rm App}(f_*) = \|f_*\|_{L^2}^2 - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V},$$

so that

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

By Eq. (63), to prove Theorem 6.(a), we need to bound $\|\boldsymbol{U}^{-1}\|_{\rm op}\|\boldsymbol{V}_{> \mathsf{M}}\|_2^2$. This is achieved in the two following propositions.

Proposition 1 *(Expected norm of $\boldsymbol{V}$). Let $\{f_* \in \mathcal{D}_d\}$ be a sequence of target functions. Define $\mathcal{E}_{> \mathsf{M}}$ by*

$$\mathcal{E}_{> \mathsf{M}} \equiv \mathbb{E}_{\boldsymbol{\theta}}\Big[\Big(\mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_{> \mathsf{M}} f_*(\boldsymbol{x})\sigma_d(\boldsymbol{x};\boldsymbol{\theta})]\Big)^2\Big].$$

Then we have

$$\mathcal{E}_{> \mathsf{M}} \le \|\mathbb{U}_{d,> \mathsf{M}}\|_{\rm op}\cdot\|\mathsf{P}_{> \mathsf{M}} f_*\|_{L^2}^2.$$

Proof of Proposition 1. We have

$$\begin{aligned}
\mathcal{E}_{> \mathsf{M}} &\equiv \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[\langle \mathsf{P}_{> \mathsf{M}} f_*, \sigma_d(\,\cdot\,,\boldsymbol{\theta})\rangle_{L^2(\mathcal{X}_d)}^2]\\
&= \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\nu_d}[\mathsf{P}_{> \mathsf{M}} f_*(\boldsymbol{x}_1)\sigma_d(\boldsymbol{x}_1,\boldsymbol{\theta})\sigma_d(\boldsymbol{x}_2,\boldsymbol{\theta})\mathsf{P}_{> \mathsf{M}} f_*(\boldsymbol{x}_2)]\\
&= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\nu_d}[\mathsf{P}_{> \mathsf{M}} f_*(\boldsymbol{x}_1)\mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[\sigma_d(\boldsymbol{x}_1,\boldsymbol{\theta})\sigma_d(\boldsymbol{x}_2,\boldsymbol{\theta})]\mathsf{P}_{> \mathsf{M}} f_*(\boldsymbol{x}_2)]
\end{aligned}$$

$$= \langle \mathsf{P}_{>\mathsf{M}} f_*, \mathbb{H}_d \, \mathsf{P}_{>\mathsf{M}} f_* \rangle_{L^2} = \langle \mathsf{P}_{>\mathsf{M}} f_*, \mathbb{H}_{d,>\mathsf{M}} \mathsf{P}_{>\mathsf{M}} f_* \rangle_{L^2}$$
$$\le \|\mathbb{H}_{d,>\mathsf{M}}\|_{\rm op} \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^2}^2 = \|\mathbb{U}_{d,>\mathsf{M}}\|_{\rm op} \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^2}^2.$$

This proves the proposition. □

Proposition 2 *(Lower bound on the kernel matrix). Let $\{\sigma_d\}_{d\ge 1}$ be a sequence of activation functions satisfying Assumptions 6.(a), 6.(b) and 7.(a) at level $\{(N(d), \mathsf{M}(d))\}_{d\ge 1}$. Let $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \tau_d$ independently and let $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ be the kernel matrix defined by Eq. (62). Then, we have*

$$\boldsymbol{U} \succeq \kappa_{>\mathsf{M}} (\boldsymbol{\Lambda} + \boldsymbol{\Delta}), \tag{64}$$

with $\boldsymbol{\Lambda} = \mathrm{diag}((U_{d,>\mathsf{M}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)/\kappa_{>\mathsf{M}})_{i\in[N]})$, $\kappa_{>\mathsf{M}} = \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})$, and $\boldsymbol{\Delta}$ is such that there exists some $\delta' > 0$, such that

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}] = O_d(N^{-\delta'}).$$

Proof of Proposition 2. This is a direct consequence of Theorem 7.(a). □

By Proposition 1, we have

$$\mathbb{E}[\|\boldsymbol{V}_{>\mathsf{M}}\|_2^2] = N\mathcal{E}_{>\mathsf{M}} \le N \cdot \|\mathbb{U}_{d,>\mathsf{M}}\|_{\rm op} \cdot \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^2}^2. \tag{65}$$

Next, it follows by Proposition 2 and Assumption 6.(d), that for any fixed $\delta > 0$ with $\delta < \delta'$,

$$\|\boldsymbol{U}^{-1}\|_{\rm op} \cdot \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) \le \Big[\min_{i\in[N]} U_{d,>\mathsf{M}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)/\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) - O_{d,\mathbb{P}}(N^{-\delta'}) \Big]^{-1} \le O_{d,\mathbb{P}}(N^{\delta}),$$

and hence by Markov's inequality, we obtain

$$\frac{\|\boldsymbol{U}^{-1}\|_{\rm op} \|\boldsymbol{V}_{>\mathsf{M}}\|_2^2}{\|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^2}^2} \le O_{d,\mathbb{P}}(N^{\delta}) \cdot N \cdot \frac{\|\mathbb{U}_{d,>\mathsf{M}}\|_{\rm op}}{\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})}. \tag{66}$$

By Assumption 7.(a), we have $N \cdot \|\mathbb{U}_{d,>\mathsf{M}}\|_{\rm op}/\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) = O_d(N^{-\delta_0})$ for some $\delta_0 > 0$. Plugging this equation into Eq. (66) and choosing $\delta < \delta_0$ yield

$$\|\boldsymbol{U}^{-1}\|_{\rm op} \|\boldsymbol{V}_{>\mathsf{M}}\|_2^2 = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^2}^2. \tag{67}$$

Combining Eq. (67) with Eq. (63) proves Theorem 6.(a).

A.3. Proof of Theorem 6.(b): upper bound on the approximation error

In the following, we would like to calculate the quantity $R_{\rm App}(\mathsf{P}_{\le\mathsf{M}} f_*, \boldsymbol{\Theta})$. We have

$$R_{\rm App}(\mathsf{P}_{\le\mathsf{M}} f_*, \boldsymbol{\Theta}) = \|\mathsf{P}_{\le\mathsf{M}} f_*\|_{L_2}^2 - \boldsymbol{V}_{\le\mathsf{M}}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{\le\mathsf{M}},$$

where $\boldsymbol{V}_{\le\mathsf{M}} = (V_{\le\mathsf{M},1}, \dots, V_{\le\mathsf{M},N})^{\mathsf{T}}$ and $\boldsymbol{U} = (U_{ij})_{ij\in[N]}$ with

$$V_{\le\mathsf{M},i} = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[\mathsf{P}_{\le\mathsf{M}} f_*(\boldsymbol{x}) \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)],$$
$$U_{ij} = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i) \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_j)].$$

Recall that $(\psi_k)_{k\ge 1}$ is the orthonormal eigenbasis of $\mathbb{H}_d$. We denote the decomposition of $\mathsf{P}_{\le \mathsf{M}} f_*$ in this basis by

$$\mathsf{P}_{\le \mathsf{M}} f_*(\boldsymbol{x}) = \sum_{k=1}^{\mathsf{M}} \langle f_*, \psi_k\rangle_{L^2} \psi_k(\boldsymbol{x}) \equiv \sum_{k=1}^{\mathsf{M}} \hat{f}_k \psi_k(\boldsymbol{x})\,.$$

Recall the decomposition of σ_d

$$\sigma_d(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \lambda_{d,k}\psi_k(\boldsymbol{x})\phi_k(\boldsymbol{\theta}).$$

By orthonormality of the $(\psi_k)_{k\ge 1}$, we have

$$V_{\le \mathsf{M},i} = \sum_{k=1}^{\mathsf{M}} \hat{f}_k \lambda_{d,k}\phi_k(\boldsymbol{\theta}_i).$$

Define

$$\begin{aligned}
\hat{\boldsymbol{f}} &= (\hat{f}_1,\dots,\hat{f}_{\mathsf{M}})^{\mathsf{T}} \in \mathbb{R}^{\mathsf{M}},\\
\boldsymbol{D} &= \mathrm{diag}(\lambda_{d,1},\dots,\lambda_{d,\mathsf{M}}) \in \mathbb{R}^{\mathsf{M}\times\mathsf{M}},\\
\boldsymbol{\Phi} &= (\phi_k(\boldsymbol{\theta}_i))_{i\in[N],k\in[\mathsf{M}]} \in \mathbb{R}^{N\times\mathsf{M}},\\
\boldsymbol{L} &= \boldsymbol{\Phi}\boldsymbol{D} \in \mathbb{R}^{N\times\mathsf{M}}.
\end{aligned}$$

Then we have

$$\boldsymbol{V}_{\le \mathsf{M}} = \Big(\sum_{k=1}^{\mathsf{M}} \hat{f}_k \lambda_{d,k}\phi_k(\boldsymbol{\theta}_i)\Big)_{i\in[N]} = \boldsymbol{\Phi}\boldsymbol{D}\hat{\boldsymbol{f}} = \boldsymbol{L}\hat{\boldsymbol{f}}.$$

By Eq. (70) in Theorem 7, there exists $\boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$ such that

$$\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{D}^2\boldsymbol{\Phi}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\boldsymbol{\Lambda}+\boldsymbol{\Delta}) = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\boldsymbol{\Lambda}+\boldsymbol{\Delta}),$$

where $\kappa_{>\mathsf{M}} = \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})$, $\boldsymbol{\Lambda} = \mathrm{diag}((U_{d,>\mathsf{M}}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)/\kappa_{>\mathsf{M}})_{i\in[N]})$. By simple algebra,

$$\boldsymbol{V}_{\le \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le \mathsf{M}} = \hat{\boldsymbol{f}}^{\mathsf{T}}\boldsymbol{S}\hat{\boldsymbol{f}},$$

where

$$\boldsymbol{S} = \boldsymbol{L}^{\mathsf{T}}(\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\boldsymbol{\Lambda}+\boldsymbol{\Delta}))^{-1}\boldsymbol{L}.$$

Therefore, we obtain

$$\begin{aligned}
R_{\mathrm{App}}(\mathsf{P}_{\le \mathsf{M}} f_*, \boldsymbol{W}) =& \|\mathsf{P}_{\le \mathsf{M}} f_*\|_{L_2}^2 - \boldsymbol{V}_{\le \mathsf{M}}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le \mathsf{M}} = \|\hat{\boldsymbol{f}}\|_2^2 - \langle \hat{\boldsymbol{f}}, \boldsymbol{S}\hat{\boldsymbol{f}}\rangle\\
\le& \|\mathbf{I}_{\mathsf{M}} - \boldsymbol{S}\|_{\mathrm{op}}\|\hat{\boldsymbol{f}}\|_2^2 = o_{d,\mathbb{P}}(1)\cdot \|\mathsf{P}_{\le \mathsf{M}} f_*\|_{L^2}^2.
\end{aligned}$$

The last equation follows from Lemma 1 which is stated and proved below. This proves the theorem.

Lemma 1 *(Concentration of $\boldsymbol{S}$). Let Assumptions 6.(a), 6.(b), 6.(c) and 7 hold. Then we have*

$$\|\mathbf{I}_{\mathsf{M}} - \boldsymbol{S}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Proof of Lemma 1. Applying the Sherman-Morrison-Woodbury formula produces the identity

$$\mathbf{I}_{\mathsf{M}} - \boldsymbol{S} = \mathbf{I}_{\mathsf{M}} - \boldsymbol{L}^{\mathsf{T}}(\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta}))^{-1}\boldsymbol{L} = (\mathbf{I}_{\mathsf{M}} + \boldsymbol{L}^{\mathsf{T}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta})^{-1}\boldsymbol{L}/\kappa_{>\mathsf{M}})^{-1},$$

so that

$$\|\mathbf{I}_{\mathsf{M}} - \boldsymbol{S}\|_{\mathrm{op}} \le 1/\lambda_{\min}(\boldsymbol{L}^{\mathsf{T}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta})^{-1}\boldsymbol{L}/\kappa_{>\mathsf{M}}). \tag{68}$$

Note that we have

$$\begin{aligned} &\lambda_{\min}(\boldsymbol{L}^{\mathsf{T}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta})^{-1}\boldsymbol{L})/\kappa_{>\mathsf{M}} = \lambda_{\min}(\boldsymbol{D}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta})^{-1}\boldsymbol{\Phi}\boldsymbol{D})/\kappa_{>\mathsf{M}} \\ \ge\ &\lambda_{\min}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/N) \cdot [N \cdot \lambda_{\min}(\boldsymbol{D}^2)/\kappa_{>\mathsf{M}}]/\|\boldsymbol{\Lambda} + \boldsymbol{\Delta}\|_{\mathrm{op}} \\ =\ &\lambda_{\min}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/N) \cdot [N \cdot \lambda_{\min}(\mathbb{U}_{d,\le\mathsf{M}})/\kappa_{>\mathsf{M}}]/\|\boldsymbol{\Lambda} + \boldsymbol{\Delta}\|_{\mathrm{op}}. \end{aligned} \tag{69}$$

By Theorem 7.(b), we know that

$$\lambda_{\min}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/N) = \Theta_{d,\mathbb{P}}(1).$$

By Assumption 6.(c), we have $\|\boldsymbol{\Lambda}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(N^{\delta})$ for any $\delta > 0$. Therefore, by Theorem 7.(a), for any $\delta > 0$, we obtain

$$\|\boldsymbol{\Lambda} + \boldsymbol{\Delta}\|_{\mathrm{op}} \le \|\boldsymbol{\Lambda}\|_{\mathrm{op}} + \|\boldsymbol{\Delta}\|_{\mathrm{op}} = O_d(N^{\delta}).$$

By Assumption 7.(b), there exists $\delta_0 > 0$, such that

$$[N \cdot \lambda_{\min}(\mathbb{U}_{d,\le\mathsf{M}})/\kappa_{>\mathsf{M}}] = \Omega_d(N^{\delta_0}).$$

Combining the above equalities with Eq. (69) and choosing δ such that $0 < \delta < \delta_0$, lead to

$$\lambda_{\min}(\boldsymbol{L}^{\mathsf{T}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta})^{-1}\boldsymbol{L}/\kappa_{>\mathsf{M}}) = \omega_{d,\mathbb{P}}(1).$$

Combining with Eq. (68) proves the lemma. □

A.4. Structure of the empirical kernel matrix

In this section, we present a key theorem describing the structure of the empirical kernel matrix $\boldsymbol{U} = (U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j))_{i,j\in[N]} \in \mathbb{R}^{N\times N}$. The proof of this theorem relies on two propositions: Proposition 3 shows that the matrix of the top eigenvectors evaluated on the random weights $(\boldsymbol{\theta}_i)_{i\in[N]}$ is nearly orthogonal and is presented in Section A.4.1; Proposition 4 shows the concentration to zero in operator norm of the off-diagonal part of the matrix $\boldsymbol{U}_{>\mathsf{M}}$ and is presented in Section A.4.2. The proof of Proposition 4 is deferred to Section A.5.

Theorem 7 *(Structure of the empirical kernel matrix). Let Assumptions 6.(a), 6.(b) and 7.(a) hold. Let $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \tau_d$ independently, and define $\boldsymbol{U} = (U_{ij})_{ij\in[N]}$ with*

$$U_{ij} := U_d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\,.$$

(Recall that $U_d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \equiv \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[\sigma_d(\boldsymbol{x}, \boldsymbol{\theta}_i)\sigma_d(\boldsymbol{x}, \boldsymbol{\theta}_j)]$*.) Then, we can rewrite* $\boldsymbol{U}$ *(by choosing* $\boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$*)*

$$\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{D}^2\boldsymbol{\Phi}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\boldsymbol{\Lambda} + \boldsymbol{\Delta}), \tag{70}$$

with $\kappa_{>\mathsf{M}} = \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})$ *and*

$$\boldsymbol{\Phi} = (\phi_k(\boldsymbol{\theta}_i))_{i\in[N],k\in[\mathsf{M}]}, \quad \boldsymbol{D} = \mathrm{diag}(\lambda_{d,1}, \dots, \lambda_{d,\mathsf{M}}), \quad \boldsymbol{\Lambda} = \mathrm{diag}((U_{d,>\mathsf{M}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)/\kappa_{>\mathsf{M}})_{i\in[N]}).$$

The following hold:

(*a*) *There exists a fixed* $\delta' > 0$*, such that*

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] = O_d(N^{-\delta'}).$$

(b) If further we assume $\mathsf{M}(d) \le N(d)^{1-\delta_0}$ *for a fixed* $\delta_0 > 0$*, then we have*

$$\left\|\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/N - \mathbf{I}_{\mathsf{M}}\right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Proof of Theorem 7. For $S \subseteq \{1, 2, 3, \dots\}$, recall that

$$\mathbb{U}_{d,S} \equiv \sum_{s\in S} \lambda_{d,s}^2 \phi_s \phi_s^* ,$$

and let $U_{d,S}$ denote the kernel associated to $\mathbb{U}_{d,S}$. Define $\boldsymbol{Q}_S = (Q_{S,ij})_{i,j\in[N(d)]}$ by

$$Q_{S,ij} = U_{d,S}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\mathbf{1}_{i\neq j}.$$

By decomposing the entries of $\boldsymbol{U}$ in the orthonormal basis $\{\phi_j\}_{j\ge1}$, we can write $\boldsymbol{U} = \boldsymbol{U}_{\le\mathsf{M}} + \boldsymbol{U}_{>\mathsf{M}}$ where

$$\boldsymbol{U}_{\le\mathsf{M}} = \boldsymbol{\Phi}\boldsymbol{D}^2\boldsymbol{\Phi}^{\mathsf{T}},$$
$$\boldsymbol{U}_{>\mathsf{M}} = (U_{d,>\mathsf{M}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j))_{i,j\in[N]}.$$

We begin by part (*b*). By Assumption 6.(*a*) and $\mathsf{M}(d) \le N(d)^{1-\delta_0}$, the assumptions of Proposition 3 are satisfied with $D = \mathsf{M}$ and $(\phi_1, \dots, \phi_{\mathsf{M}})$ the top M eigenvectors of $\mathbb{U}$. Hence, there exists $C = C(q) > 0$, a constant that depends only on q such that

$$\mathbb{E}\left[\left\|\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/N - \mathbf{I}_{\mathsf{M}}\right\|_{\mathrm{op}}\right] \le C\frac{\mathsf{M}\log(N)}{N^{1-1/q}}.$$

Taking $q > 1/\delta_0$, the right hand side becomes $o_d(1)$ and Theorem 7.(*b*) follows by Markov's inequality.

Next, we prove part (*a*), namely that $\boldsymbol{U}_{>\mathsf{M}} = \kappa_{>\mathsf{M}} \cdot (\boldsymbol{\Lambda} + \boldsymbol{\Delta})$ with $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(N^{-\delta'})$ for some $\delta' > 0$.

Letting $\boldsymbol{Q} \in \mathbb{R}^{N\times N}$ be the matrix with entries $Q_{ij} = (\boldsymbol{U}_{>\mathsf{M}})_{ij}\mathbf{1}_{i\neq j}$. Then we have $\boldsymbol{U}_{>\mathsf{M}} = \kappa_{>\mathsf{M}}\boldsymbol{\Lambda} + \boldsymbol{Q}$. We next apply Proposition 4 to the operator $\widehat{\mathbb{U}}_d = \mathbb{U}_{d,>\mathsf{M}}$ and subspace $\widehat{\mathcal{V}}_d = \mathcal{V}_{d,>\mathsf{M}}$. Notice that the assumptions of Proposition 4 are satisfied by Assumptions 6.(*a*), 6.(*b*) and 7.(*a*). We therefore conclude that $\mathbb{E}[\|\boldsymbol{Q}\|_{\mathrm{op}}] = O_d(N^{-\delta'}) \cdot \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}) = O_d(N^{-\delta'}) \cdot \kappa_{>\mathsf{M}}$ for some $\delta' > 0$. This concludes the proof of Theorem 7.(*a*). □

This theorem implies a particularly simple structure of the empirical kernel matrix $\boldsymbol{U}$. Under the additional Assumptions 6.(*c*), 6.(*d*) and 7.(*b*), $\boldsymbol{U}$ can be written as a sum of a 'spike' $\boldsymbol{U}_{\le\mathsf{M}}$ (of rank M and

eigenvalues $\gg \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})$) and a full rank matrix $\boldsymbol{U}_{>\mathsf{M}}$ with eigenvalues of order $\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})$. The 'spike' matrix $\boldsymbol{U}_{\leq\mathsf{M}}$ has the following approximate diagonalization:

$$\boldsymbol{U}_{\leq\mathsf{M}} = \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{D}}^2\tilde{\boldsymbol{\Phi}}^{\mathsf{T}},$$

where $\tilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi}/\sqrt{N} \in \mathbb{R}^{N\times\mathsf{M}}$ is approximately an orthogonal matrix $\|\tilde{\boldsymbol{\Phi}}^{\mathsf{T}}\tilde{\boldsymbol{\Phi}} - \mathbf{I}_{\mathsf{M}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and the diagonal matrix $\tilde{\boldsymbol{D}}^2 = \mathrm{diag}(N\lambda_{d,1}^2, \dots, N\lambda_{d,\mathsf{M}}^2)$ verifies $\tilde{\boldsymbol{D}}^2 \succeq N(d)^{\delta_0}\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})\cdot\mathbf{I}_N$ (by Assumption 7.(*b*)). Furthermore, by Assumptions 6.(*c*), and 6.(*d*), and Theorem 7.(*a*), we have for any $\delta > 0$,

$$\Omega_{d,\mathbb{P}}(N^{-\delta})\cdot\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})\cdot\mathbf{I}_N \preceq \boldsymbol{U}_{>\mathsf{M}} \preceq O_{d,\mathbb{P}}(N^{\delta})\cdot\mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}})\cdot\mathbf{I}_N.$$

A.4.1. Concentration of the top eigenvectors

Here we state and prove a general matrix concentration result. For each $d \geq 1$, let (Ω_d, τ_d) be a (Polish) probability space, and $(\phi_k)_{k\geq 1}$ an orthonormal basis of $L^2(\Omega_d, \tau_d)$. Define $\boldsymbol{\phi}(\boldsymbol{\theta}) \equiv (\phi_1(\boldsymbol{\theta}), \dots, \phi_D(\boldsymbol{\theta}))^{\mathsf{T}} \in \mathbb{R}^D$, and let $(\boldsymbol{\theta}_i)_{i\leq N} \sim_{iid} \tau_d$. The law of large numbers and orthonormality imply that, for any fixed D,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}(\boldsymbol{\theta}_i)\boldsymbol{\phi}(\boldsymbol{\theta}_i)^{\mathsf{T}} = \int_{\Omega_d}\boldsymbol{\phi}(\boldsymbol{\theta})\boldsymbol{\phi}(\boldsymbol{\theta})^{\mathsf{T}}\,\tau_d(\mathrm{d}\boldsymbol{\theta}) = \mathbf{I}_D\,. \tag{71}$$

The next proposition establishes a generalization of this fact for the case in which both D and N diverge.

Proposition 3. *Let $\{\phi_k \in L^2(\Omega,\tau)\}_{k=1}^D$ be orthonormal functions. Let $\{\boldsymbol{\theta}_i\}_{i\in[N]} \sim \tau$ independently. Define $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i) = (\phi_1(\boldsymbol{\theta}_i), \dots, \phi_D(\boldsymbol{\theta}_i))^{\mathsf{T}} \in \mathbb{R}^D$ for $i \in [N]$. We assume that, for any integer $q \geq 2$, there exists $C = C(q)$ such that we have*

$$\sup_{k\in[D]}\|\phi_k\|_{L^{2q}} \leq C(q). \tag{72}$$

Then for any $q \geq 2$, there exists $K = K(q)$ that only depends on $C(q)$, such that denoting $\delta \equiv K(q)D\log(D\vee N)/N^{1-1/q}$, we have

$$\mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathsf{T}} - \mathbf{I}_D\Big\|_{\mathrm{op}} \leq (\delta\vee\sqrt{\delta}).$$

Proof of Proposition 3. By the hypercontractivity assumption, cf. Eq. (72), we have

$$\begin{aligned}\Gamma &:= \mathbb{E}\Big[\max_{i\in[N]}\|\boldsymbol{\phi}_i\|_2^2\Big] \leq \mathbb{E}\Big[\max_{i\in[N]}\|\boldsymbol{\phi}_i\|_2^{2q}\Big]^{1/q} \leq N^{1/q}\cdot\mathbb{E}[\|\boldsymbol{\phi}_i\|_2^{2q}]^{1/q}\\ &= N^{1/q}\cdot\Big\|\sum_{k=1}^{D}\phi_k^2\Big\|_{L^q} \leq N^{1/q}D\cdot\max_{k\in[D]}\|\phi_k\|_{L^{2q}}^2 \leq C(q)^2\cdot N^{1/q}D.\end{aligned}$$

Applying Lemma 2 below proves the proposition. □

Lemma 2 *([42] Theorem 5.45). Let $\{\boldsymbol{a}_i \in \mathbb{R}^D\}_{i\in[N]}$ be independent random vectors with $\mathbb{E}[\boldsymbol{a}_i\boldsymbol{a}_i^{\mathsf{T}}] = \mathbf{I}_D$. Denote $\Gamma \equiv \mathbb{E}[\max_{i\in[N]}\|\boldsymbol{a}_i\|_2^2]$. Then there exists a universal constant C, such that denoting $\delta \equiv C\cdot\Gamma\cdot\log(N\wedge D)/N$, we have*

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathsf{T}} - \mathbf{I}_D\Big\|_{\mathrm{op}}\Big] \leq \delta\vee\sqrt{\delta}.$$

A.4.2. Bounding the off-diagonal part of the matrix $U_{>\mathsf{M}}$

We state a key proposition) whose proof will be presented in Section A.5. The statement and the assumptions are self-contained.

Proposition 4 *(Bound on the off-diagonal part of the matrix $\boldsymbol{U}_{>\mathsf{M}}$). Let $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim_{iid} \tau_d$. Let $\widehat{\mathbb{U}}_d$ be a self-adjoint positive definite operator $\widehat{\mathbb{U}}_d : \widehat{\mathcal{V}}_d \to \widehat{\mathcal{V}}_d$, $\widehat{\mathcal{V}}_d \subseteq L^2(\Omega_d)$ with kernel $\widehat{U}_d \in L^2(\Omega_d \times \Omega_d)$ (see Eq. (11)) satisfying $\int_{\Omega_d} \widehat{U}_d(\boldsymbol{\theta},\boldsymbol{\theta}')f(\boldsymbol{\theta}')\,\tau_d(\mathrm{d}\boldsymbol{\theta}') = 0$ for any $f \in \mathcal{V}_d^{\perp}$. Let $(\hat{\phi}_j)_{j\ge1}$ be an orthonormal basis of eigenfunctions with $\mathrm{span}(\hat{\phi}_j, j \ge 1) = \widehat{\mathcal{V}}_d \subseteq L^2(\Omega_d)$, and eigenvalues $(\hat{\lambda}_{d,j})_{j\ge1} \subseteq \mathbb{R}$ with nonincreasing absolute values $|\hat{\lambda}_{d,1}| \ge |\hat{\lambda}_{d,2}| \ge \cdots$ and $\sum_{j\ge1}\hat{\lambda}_{d,j}^2 < \infty$, such that*

$$\widehat{\mathbb{U}}_d = \sum_{j=1}^{\infty} \hat{\lambda}_{d,j}^2 \hat{\phi}_j \hat{\phi}_j^*, \qquad \widehat{U}_d(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{j=1}^{\infty} \hat{\lambda}_{d,j}^2 \hat{\phi}_j(\boldsymbol{\theta})\hat{\phi}_j(\boldsymbol{\theta}').$$

When $S \subseteq \{1,2,3,\ldots\}$, we denote

$$\widehat{\mathbb{U}}_{d,S} = \sum_{j\in S} \hat{\lambda}_{d,j}^2 \hat{\phi}_j \hat{\phi}_j^*, \qquad \widehat{U}_{d,S}(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{j\in S} \hat{\lambda}_{d,j}^2 \hat{\phi}_j(\boldsymbol{\theta})\hat{\phi}_j(\boldsymbol{\theta}').$$

We make the following assumptions:

(A1) *There exists a sequence $\{v(d)\}_{d\ge1}$, such that for any fixed $q \ge 1$, there exists $C = C(q, \{v(d)\}_{d\ge1})$ such that, for any $f_* \in \widehat{\mathcal{V}}_{d,\le v(d)} \equiv \mathrm{span}(\hat{\phi}_s, 1 \le s \le v(d))$, we have*

$$\|f_*\|_{L^{2q}} \le C \cdot \|f_*\|_{L^2}.$$

(A2) *For the same sequence $\{v(d)\}_{d\ge1}$ as in (A1), there exists fixed $\delta_0 > 0$, such that*

$$\mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)}^2) \cdot N(d)^{2+\delta_0} = O_d(1) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)})^2.$$

(A3) *There exists $\delta_0 > 0$, such that*

$$N(d)^{1+\delta_0} \cdot \|\widehat{\mathbb{U}}_d\|_{\mathrm{op}} = O_d(1) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_d). \tag{73}$$

Consider the random matrix $\boldsymbol{Q} = (Q_{ij})_{i,j\in[N(d)]} \in \mathbb{R}^{N\times N}$, with

$$Q_{ij} = \widehat{U}_d(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\mathbf{1}_{i\ne j}.$$

Then there exists $\delta' > 0$, such that

$$\mathbb{E}[\|\boldsymbol{Q}\|_{\mathrm{op}}] = O_d(N^{-\delta'}) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_d).$$

A.5. Proof of Proposition 4

We begin by stating two key estimates which are used in the proof of Proposition 4. The notations of Lemma 3 follow the notations of Proposition 4. The notations and assumptions of Proposition 5 are self-contained. We collect a number of technical lemmas in Section A.5.1.

Lemma 3. *Consider the same setup as Proposition 4. Let $\{N(d)\}_{d\ge1}$ and $\{v(d)\}_{d\ge1}$ be two sequences, and assume that there exists $\delta_0 > 0$ such that (this is Assumption (A2) in Proposition 4)*

$$N(d)^2 \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)}^2) = O_d(N^{-\delta_0}) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)})^2. \tag{74}$$

Consider the random matrix $\boldsymbol{Q}_{>v(d)} = (Q_{>v(d),ij})_{i,j\in[N(d)]} \in \mathbb{R}^{N\times N}$, *with*

$$Q_{>v(d),ij} = \widehat{U}_{d,>v(d)}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\mathbf{1}_{i\neq j}.$$

Then we have

$$\mathbb{E}[\|\boldsymbol{Q}_{>v(d)}\|_{\mathrm{op}}^2]^{1/2} = O_d(N^{-\delta_0}) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)}).$$

Proposition 5 *(Vanishing off-diagonal). Let* $\overline{\mathbb{U}}$ *be a compact self-adjoint positive definite operator on a closed subspace* $\overline{\mathcal{V}} \subseteq L^2(\Omega,\tau)$, $\overline{\mathbb{U}} : \overline{\mathcal{V}} \to \overline{\mathcal{V}}$, *with corresponding kernel* $\overline{U} \in L^2(\Omega\times\Omega)$, *satisfying* $\int_\Omega \overline{U}(\boldsymbol{\theta},\boldsymbol{\theta}')\, f(\boldsymbol{\theta}')\, \tau(\mathrm{d}\boldsymbol{\theta}') = 0$ *for all* $f \in \overline{\mathcal{V}}^{\perp}$. *For any* $q \geq 1$, *we assume that there exists* $C(q)$ *such that*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\sim\tau}[|\overline{U}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)|^{2q}]^{1/(2q)} &\leq C(q)\cdot \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\sim\tau}[\overline{U}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)^2]^{1/2}, \\ \mathbb{E}_{\boldsymbol{\theta}\sim\tau}[|\overline{U}(\boldsymbol{\theta},\boldsymbol{\theta})|^{q}]^{1/q} &\leq C(q)\cdot \mathbb{E}_{\boldsymbol{\theta}\sim\tau}[\overline{U}(\boldsymbol{\theta},\boldsymbol{\theta})]. \end{aligned} \tag{75}$$

Moreover, let $\{\boldsymbol{\theta}_i\}_{i\in[N]} \sim_{iid} \tau$ *independently, and consider* $\boldsymbol{\Delta} = (\Delta_{ij})_{i,j\in[N]} \in \mathbb{R}^{N\times N}$, *with*

$$\Delta_{ij} = \overline{U}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\mathbf{1}_{i\neq j}.$$

Then for any integer $p > 0$, *there exists a constant* $K(p)$ *which only depends on the constant* $C(p)$, *such that*

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] \leq K(p)\cdot\Big\{N\|\overline{\mathbb{U}}\|_{\mathrm{op}} + [\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathrm{Tr}(\overline{\mathbb{U}})N^{1+2/p}\log N]^{1/2}\Big\}. \tag{76}$$

We are now in position to prove Proposition 4.

Proof of Proposition 4. We decompose the operator $\widehat{\mathbb{U}}_d = \widehat{\mathbb{U}}_{d,\leq v(d)} + \widehat{\mathbb{U}}_{d,>v(d)}$, and the kernel $\widehat{U}_d = \widehat{U}_{d,\leq v(d)} + \widehat{U}_{d,>v(d)}$. Define $\boldsymbol{Q}_S = (Q_{S,ij})_{i,j\in[N(d)]}$ with

$$Q_{S,ij} = U_{d,S}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\mathbf{1}_{i\neq j}.$$

By Assumption (A2) and by Lemma 3, we have

$$\mathbb{E}[\|\boldsymbol{Q}_{>v(d)}\|_{\mathrm{op}}^2]^{1/2} = O_d(N^{-\delta_0}) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,>v(d)}).$$

By Assumption (A1) and Lemma 6 which is stated in Section A.5.1 below, the assumptions of Proposition 5 are satisfied, in which we take $\boldsymbol{\Delta} = \boldsymbol{Q}_{\leq v(d)}$, $\overline{\mathbb{U}} = \widehat{\mathbb{U}}_{d,\leq v(d)}$, $\overline{U} = \widehat{U}_{d,\leq v(d)}$, and $\overline{\mathcal{V}} \equiv \mathrm{span}(\phi_s :\ 1 \leq s \leq v(d))$. Further by Assumption (A3) as in Eq. (73), we fix some $p > 4/\delta_0$ in Proposition 5, then for $\delta' = \delta_0/4 > 0$, we obtain

$$\mathbb{E}[\|\boldsymbol{Q}_{\leq v(d)}\|_{\mathrm{op}}] = O_d(N^{-\delta'}) \cdot \mathrm{Tr}(\widehat{\mathbb{U}}_{d,\leq v(d)}).$$

Combining the equations in the last two displays proves the proposition. □

We next prove Lemma 3 and Proposition 5.

Proof of Lemma 3. We have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{Q}_{>v(d)}\|_{\rm op}^2] &\le \mathbb{E}[\|\boldsymbol{Q}_{>v(d)}\|_F^2] = N(N-1)\cdot \mathbb{E}[Q_{>v(d),ij}^2]\\ &= N(N-1)\cdot {\rm Tr}(\widehat{\mathbb{U}}_{d,>v(d)}^2) = O_d(N^{-\delta_0})\cdot {\rm Tr}(\widehat{\mathbb{U}}_{d,>v(d)})^2,\end{aligned}$$

where the last equation is by Eq. (74). This proves the lemma. □

Proof of Proposition 5. With a little abuse of notation, we define $\overline{\boldsymbol{U}} = (\overline{U}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j))_{i,j\in[N]} \in \mathbb{R}^{N\times N}$.

Step 1. Bound $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}]$ using matrix decoupling. For $T_1, T_2 \subseteq [N]$, we denote $\boldsymbol{A}_{T_1,T_2} = (A_{ij})_{i\in T_1, j\in T_2}$. By Lemma 4, which is stated in Section A.5.1 below, we have

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}] \le 4 \sup_{T\subseteq[N]} \mathbb{E}[\|\boldsymbol{\Delta}_{TT^c}\|_{\rm op}]. \tag{77}$$

For any $S \subseteq [N]$, we denote $\mathbb{E}_S$ to be the expectation with respect to $\{\boldsymbol{\theta}_i\}_{i\in S}$ and conditional on $\{\boldsymbol{\theta}_j\}_{j\in S^c}$. Fix $T \subseteq [N]$. Using Lemma 5 (which is stated in Section A.5.1 below) conditioning on $\{\boldsymbol{\theta}_j\}_{j\in T^c}$, we get

$$\mathbb{E}_T[\|\boldsymbol{\Delta}_{TT^c}\|_{\rm op}] \le [\Sigma(T)\cdot N]^{1/2} + C\cdot(\Gamma(T)\cdot \log N)^{1/2},$$

where $\Sigma(T) \equiv \|\mathbb{E}_{\boldsymbol{\theta}_u}[\boldsymbol{\Delta}_{T^cu}\boldsymbol{\Delta}_{uT^c}]\|_{\rm op}$ (for some $u \in T$) and $\Gamma(T) \equiv \mathbb{E}_T[\max_{i\in T}\|\boldsymbol{\Delta}_{iT^c}\|_2^2]$. Therefore, by Holder's inequality:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}] &\le 4\sup_{T\subseteq[N]} \mathbb{E}[\|\boldsymbol{\Delta}_{TT^c}\|_{\rm op}] = 4\sup_{T\subseteq[N]} \mathbb{E}_{T^c}\mathbb{E}_T[\|\boldsymbol{\Delta}_{TT^c}\|_{\rm op}]\\ &\le 4\sup_{T\subseteq[N]} \Big\{[\mathbb{E}_{T^c}[\Sigma(T)]\cdot N]^{1/2} + C\cdot(\mathbb{E}_{T^c}[\Gamma(T)]\cdot\log N)^{1/2}\Big\}.\end{aligned} \tag{78}$$

Step 3. Bound $\mathbb{E}_{T^c}[\Sigma(T)]$. By the compactness of operator $\overline{\mathbb{U}}|_{\overline{V}}$, there exists orthogonal basis $\{\phi_k\}_{k\ge1}$ and real numbers $\{\lambda_k\}_{k\ge1}$, such that $\overline{U}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) = \sum_k \lambda_k^2\phi_k(\boldsymbol{\theta}_i)\phi_k(\boldsymbol{\theta}_j)$. Therefore, we have

$$\begin{aligned}\Sigma(T) &= \|\mathbb{E}_{\boldsymbol{\theta}_u}[\boldsymbol{\Delta}_{T^cu}\boldsymbol{\Delta}_{uT^c}]\|_{\rm op} = \sup_{\|\boldsymbol{z}\|_2=1}\sum_{i,j\in T^c}\sum_k \lambda_k^4\phi_k(\boldsymbol{\theta}_i)\phi_k(\boldsymbol{\theta}_j)z_iz_j\\ &\le \|\overline{\mathbb{U}}\|_{\rm op}\cdot\sup_{\|\boldsymbol{z}\|_2=1}\sum_{i,j\in T^c}\sum_k \lambda_k^2\phi_k(\boldsymbol{\theta}_i)\phi_k(\boldsymbol{\theta}_j)z_iz_j\\ &= \|\overline{\mathbb{U}}\|_{\rm op}\cdot\|(\overline{U}_{ij})_{i,j\in T^c}\|_{\rm op} \le \|\overline{\mathbb{U}}\|_{\rm op}\cdot[\|{\rm ddiag}(\overline{\boldsymbol{U}})\|_{\rm op} + \|\boldsymbol{\Delta}\|_{\rm op}],\end{aligned}$$

where we denoted ddiag($\overline{\boldsymbol{U}}$) the diagonal matrix obtained by zeroing the non-diagonal elements of $\overline{\boldsymbol{U}}$. Note by the hypercontractivity assumption as in Eq. (75), we have

$$\mathbb{E}[\|{\rm ddiag}(\overline{\boldsymbol{U}})\|_{\rm op}] \le \mathbb{E}\Big[\sum_{i=1}^N \overline{U}_{ii}^p\Big]^{1/p} \le N^{1/p}\cdot\mathbb{E}[\overline{U}_{ii}^p]^{1/p} \le C(p)N^{1/p}\cdot\mathbb{E}[\overline{U}_{ii}] \le C(p)N^{1/p}\cdot{\rm Tr}(\overline{\mathbb{U}}).$$

This gives

$$\mathbb{E}_{T^c}[\Sigma(T)] \le C(p)N^{1/p}\cdot\|\overline{\mathbb{U}}\|_{\rm op}{\rm Tr}(\overline{\mathbb{U}}) + \|\overline{\mathbb{U}}\|_{\rm op}\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}]. \tag{79}$$

Step 4. Bound $\mathbb{E}_{T^c}[\Gamma(T)]$. By the hypercontractivity assumption as in Eq. (75), we obtain

$$\begin{aligned}\mathbb{E}_{T^c}[\Gamma(T)] &\equiv \mathbb{E}\Big[\max_{i\in T}\|\boldsymbol{\Delta}_{iT^c}\|_2^2\Big] \le N\cdot\mathbb{E}\Big[\max_{i\in T}\max_{j\in T^c}\Delta_{ij}^2\Big]\\ &\le N\cdot\mathbb{E}\Big[\max_{i\in T,j\in T^c}\Delta_{ij}^{2p}\Big]^{1/p} \le N^{1+2/p}\cdot\mathbb{E}[\Delta_{ij}^{2p}]^{1/p}\\ &= C(p)^2N^{1+2/p}\cdot\mathbb{E}[\Delta_{ij}^2] \le C(p)^2N^{1+2/p}\cdot\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathrm{Tr}(\overline{\mathbb{U}}).\end{aligned} \tag{80}$$

The last inequality holds since $\mathbb{E}[\Delta_{ij}^2]=\mathbb{E}\{[\sum_k\lambda_k\phi_k(\boldsymbol{\theta}_i)\phi_k(\boldsymbol{\theta}_j)]^2\}=\sum_k\lambda_k^4\le\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathrm{Tr}(\overline{\mathbb{U}})$.

Step 5. Combining the equations. Combining Eq. (78), (79), and (80):

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}] &\le 4\sup_{T\subseteq[N]}\Big\{[\mathbb{E}_{T^c}[\Sigma(T)]\cdot N]^{1/2}+C\cdot(\mathbb{E}_{T^c}[\Gamma(T)]\cdot\log N)^{1/2}\Big\}\\ &\le K(p)\Big\{\{\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathrm{Tr}(\overline{\mathbb{U}})N^{1+2/p}\log N\}^{1/2}+\{N\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}]\}^{1/2}\Big\}.\end{aligned}$$

Denote $\varepsilon_1=K(p)(N\|\overline{\mathbb{U}}\|_{\mathrm{op}})^{1/2}\ge0$ and $\varepsilon_2=K(p)\{\|\overline{\mathbb{U}}\|_{\mathrm{op}}\mathrm{Tr}(\overline{\mathbb{U}})N^{1+2/p}\log N\}^{1/2}\ge0$, $x=\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}]^{1/2}$. The above inequality implies $x^2-\varepsilon_1x-\varepsilon_2\le0$, which gives $x\le[\varepsilon_1+(\varepsilon_1^2+4\varepsilon_2)^{1/2}]/2\le(\varepsilon_1^2+4\varepsilon_2)^{1/2}$. This concludes the proof. □

A.5.1. Auxiliary lemmas

The following standard decoupling trick follows, for instance, from [42] in Lemma 5.60.

Lemma 4 *(Matrix decoupling). Let $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ be a real symmetric random matrix. For $T_1,T_2\subseteq\{1,2,\dots,N\}$, we denote $\boldsymbol{A}_{T_1,T_2}=(A_{ij})_{i\in T_2,j\in T_2}$. Then we have*

$$\mathbb{E}[\|\boldsymbol{A}-\mathrm{ddiag}(\boldsymbol{A})\|_{\mathrm{op}}]\le4\max_{T\subseteq[N]}\mathbb{E}[\|\boldsymbol{A}_{T,T^c}\|_{\mathrm{op}}].$$

Proof of Lemma 4. Let T be a random subset of $\{1,2,\dots,N\}$, with each element selected with probability $1/2$ independently. For any $\boldsymbol{x}\in\mathbb{S}^{N-1}$, we have

$$\langle\boldsymbol{x},[\boldsymbol{A}-\mathrm{ddiag}(\boldsymbol{A})]\boldsymbol{x}\rangle=4\mathbb{E}_T\Big[\sum_{i\in T,j\in T^c}A_{ij}x_ix_j\Big].$$

By Jensen's inequality we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{A}-\mathrm{ddiag}(\boldsymbol{A})\|_{\mathrm{op}}] &= \mathbb{E}_{\boldsymbol{A}}\Big[\sup_{\boldsymbol{x}\in\mathbb{S}^{N-1}}\langle\boldsymbol{x},[\boldsymbol{A}-\mathrm{ddiag}(\boldsymbol{A})]\boldsymbol{x}\rangle\Big]\le4\mathbb{E}_T\mathbb{E}_{\boldsymbol{A}}\Big[\sup_{\boldsymbol{x}\in\mathbb{S}^{N-1}}\sum_{i\in T,j\in T^c}A_{ij}x_ix_j\Big]\\ &\le4\sup_{T\subseteq[N]}\mathbb{E}[\|A_{TT^c}\|_{\mathrm{op}}].\end{aligned}$$

This completes the proof. □

Lemma 5 *([42] Theorem 5.48). Let $\boldsymbol{A}\in\mathbb{R}^{N\times n}$ with $\boldsymbol{A}^{\mathsf{T}}=[\boldsymbol{a}_1,\dots,\boldsymbol{a}_N]$ where $\boldsymbol{a}_i$ are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\boldsymbol{\Sigma}=\mathbb{E}[\boldsymbol{a}_i\boldsymbol{a}_i^{\mathsf{T}}]$. Let $\Gamma\equiv\mathbb{E}[\max_{i\in[N]}\|\boldsymbol{a}_i\|_2^2]$. Then there exists a universal constant C, such that*

$$\mathbb{E}[\|\boldsymbol{A}\|_{\mathrm{op}}^2]^{1/2}\le(\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\cdot N)^{1/2}+C\cdot(\Gamma\cdot\log(N\wedge n))^{1/2}.$$

Lemma 6. *Let $\{\phi_k\}_{1\le k\le Z}\subseteq L^2(\Omega,\tau)$ be a set of orthonormal functions. We assume that, for any fixed $q\ge1$, there exists $C=C(q)$, such that for any $f\in\mathrm{span}\{\phi_k:1\le k\le Z\}$, we have*

$$\|f\|_{L^{2q}} \le C(q) \cdot \|f\|_{L^2}.$$

For $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Omega$, we denote $\overline{U}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{k=1}^{Z} \lambda_k^2 \phi_k(\boldsymbol{\theta}) \phi_k(\boldsymbol{\theta}')$ where $\{\lambda_k\}_{1 \le k \le Z}$ are fixed real numbers. Then for any $q \ge 1$, we have

$$\mathbb{E}_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim \tau}[\overline{U}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{2q}]^{1/(2q)} \le C(q)^2 \cdot \mathbb{E}_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim \tau}[\overline{U}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^2]^{1/2}, \tag{81}$$

$$\mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\overline{U}(\boldsymbol{\theta}, \boldsymbol{\theta})^q]^{1/q} \le C(q)^2 \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\overline{U}(\boldsymbol{\theta}, \boldsymbol{\theta})]. \tag{82}$$

Proof of Lemma 6. For any $q \ge 1$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim \tau}[\overline{U}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{2q}] &= \mathbb{E}_{\boldsymbol{\theta}_1 \sim \tau}\Big\{ \mathbb{E}_{\boldsymbol{\theta}_2 \sim \tau}\Big\{ \Big[\sum_{k=1}^{Z} \lambda_k^2 \phi_k(\boldsymbol{\theta}_1) \phi_k(\boldsymbol{\theta}_2) \Big]^{2q} \Big| \boldsymbol{\theta}_1 \Big\} \Big\} \\
&\overset{(a)}{\le} C(q)^{2q} \cdot \mathbb{E}_{\boldsymbol{\theta}_1 \sim \tau}\Big\{ \mathbb{E}_{\boldsymbol{\theta}_2 \sim \tau}\Big\{ \Big[\sum_{k=1}^{Z} \lambda_k^2 \phi_k(\boldsymbol{\theta}_1) \phi_k(\boldsymbol{\theta}_2) \Big]^{2} \Big| \boldsymbol{\theta}_1 \Big\}^q \Big\} \overset{(b)}{=} C(q)^{2q} \cdot \mathbb{E}_{\boldsymbol{\theta}_1 \sim \tau}\Big\{ \Big[\sum_{k=1}^{Z} \lambda_k^4 \phi_k(\boldsymbol{\theta}_1)^2 \Big]^q \Big\} \\
&\overset{(c)}{\le} C(q)^{2q} \cdot \Big\{ \sum_{k=1}^{Z} \lambda_k^4 \cdot \mathbb{E}_{\boldsymbol{\theta}_1 \sim \tau}[\phi_k(\boldsymbol{\theta}_1)^{2q}]^{1/q} \Big\}^q \overset{(d)}{\le} C(q)^{2q} \cdot \Big\{ C(q)^2 \sum_{k=1}^{Z} \lambda_k^4 \cdot \mathbb{E}_{\boldsymbol{\theta}_1 \sim \tau}[\phi_k(\boldsymbol{\theta}_1)^2] \Big\}^q \\
&\overset{(e)}{=} C(q)^{4q} \Big[\sum_{k=1}^{Z} \lambda_k^4 \Big]^q \overset{(f)}{=} C(q)^{4q} \cdot \Big\{ \mathbb{E}_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim \tau}[\overline{U}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^2] \Big\}^q.
\end{aligned}$$

Here, inequality (a) follows by applying the hypercontractivity inequality with respect to the function $f(\boldsymbol{\theta}_2) = \sum_{k=1}^{Z} \lambda_k^2 \phi_k(\boldsymbol{\theta}_1) \phi_k(\boldsymbol{\theta}_2)$ (and conditional on $\boldsymbol{\theta}_1$). Equality (b) by the fact that $(\phi_k)_{1 \le k \le Z}$ are orthonormal functions. Inequality (c) is by the Minkowski inequality. Inequality (d) follows by applying the hypercontractivity inequality with respect to $f(\boldsymbol{\theta}_1) = \phi_k(\boldsymbol{\theta}_1)$. Equality (e) holds because $(\phi_k)_{1 \le k \le Z}$ are orthonormal functions. Finally, equality (f) follows by simple calculation. This proves Eq. (81).

For any $q \ge 1$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\overline{U}(\boldsymbol{\theta}, \boldsymbol{\theta})^q] &= \mathbb{E}_{\boldsymbol{\theta} \sim \tau}\Big[\Big(\sum_{k=1}^{Z} \lambda_k^2 \phi_k(\boldsymbol{\theta})^2 \Big)^q \Big] \overset{(a)}{\le} \Big[\sum_{k=1}^{Z} \lambda_k^2 \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\phi_k(\boldsymbol{\theta})^{2q}]^{1/q} \Big]^q \\
&\overset{(b)}{\le} C(q)^{2q} \Big[\sum_{k=1}^{Z} \lambda_k^2 \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\phi_k(\boldsymbol{\theta})^2] \Big]^q \overset{(c)}{=} C(q)^{2q} \Big[\sum_{k=1}^{Z} \lambda_k^2 \Big]^q \overset{(d)}{=} C(q)^{2q} \Big\{ \mathbb{E}_{\boldsymbol{\theta} \sim \tau}[\overline{U}(\boldsymbol{\theta}, \boldsymbol{\theta})] \Big\}^q.
\end{aligned}$$

Here, inequality (a) holds by Minkowski inequality. Inequality (b) follows by applying the hypercontractivity inequality with respect to $f(\boldsymbol{\theta}) = \phi_k(\boldsymbol{\theta})$. Equality (c) holds because $(\phi_k)_{1 \le k \le Z}$ are orthonormal functions, and equality (d) by a simple calculation. This proves Eq. (82). □

Lemma 7 *(Bound on the maximum of diagonal). Consider a sequence of probability spaces (Ω_d, τ_d) with $\{\phi_{d,k}\}_{k \ge 1}$ an orthonormal basis of functions for $\mathcal{D}_d \subseteq L^2(\Omega_d, \tau_d)$. Assume that there exists a sequence of integers $\{u(d)\}_{d \ge 1}$ such that the subspace $\mathcal{D}_{d, \le u(d)} = \operatorname{span}(\phi_{d,k} : 1 \le k \le u(d))$ is hypercontractive, i.e., for any fixed $k \ge 1$, there exists a constant C such that, for any $g \in \mathcal{D}_{d, \le u(d)}$, we have*

$$\|g\|_{L^{2k}(\Omega_d)} \le C \cdot \|g\|_{L^2(\Omega_d)}.$$

Let $\{U_d\}_{d \ge 1}$ be a sequence of positive definite kernels $U_d \in L^2(\Omega_d \times \Omega_d)$ with

$$U_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{j=1}^{\infty} \lambda_{d,k}^2 \phi_{d,k}(\boldsymbol{\theta}_1) \phi_{d,k}(\boldsymbol{\theta}_2).$$

Denote $U_{d,>\ell}$ the kernel function obtained by setting $\lambda_{d,1} = \ldots = \lambda_{d,\ell} = 0$. Letting $(\boldsymbol{\theta}_i)_{i\in[N(d)]} \sim_{iid} \tau_d$, if we assume that for any $\delta > 0$,

$$\max_{i\in[N(d)]} U_{d,>u(d)}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i) = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[U_{d,>u(d)}(\boldsymbol{\theta},\boldsymbol{\theta})], \tag{83}$$

then for any $\delta > 0$,

$$\max_{i\in[N(d)]} U_{d}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i) = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[U_{d}(\boldsymbol{\theta},\boldsymbol{\theta})]. \tag{84}$$

Furthermore, if we assume that for any $\delta > 0$,

$$\max_{i\in[N(d)]} \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[U_{d,>u(d)}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2] = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\sim\tau_d}[U_{d,>u(d)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)^2], \tag{85}$$

then for any $\delta > 0$,

$$\max_{i\in[N(d)]} \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[U_{d}(\boldsymbol{\theta}_i,\boldsymbol{\theta})^2] = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\sim\tau_d}[U_{d}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)^2]. \tag{86}$$

Proof of Lemma 7. Let us decompose U_d in a high and low degree parts, $U_d = U_{d,\le u} + U_{d,>u}$ where

$$\begin{aligned}
U_{d,\le u}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= \sum_{k=1}^{u} \lambda_{d,k}^2 \phi_{d,k}(\boldsymbol{\theta}_1)\phi_{d,k}(\boldsymbol{\theta}_2),\\
U_{d,>u}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= \sum_{k=u+1}^{\infty} \lambda_{d,k}^2 \phi_{d,k}(\boldsymbol{\theta}_1)\phi_{d,k}(\boldsymbol{\theta}_2).
\end{aligned}$$

By Lemma 6, we have for any $q \ge 1$,

$$\begin{aligned}
\mathbb{E}\Big[\max_{i\in[N(d)]} U_{d,\le u}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)\Big] &\le \mathbb{E}\Big[\max_{i\in[N(d)]} U_{d,\le u}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)^q\Big]^{1/q}\\
&\le N^{1/q}\mathbb{E}\Big[U_{d,\le u}(\boldsymbol{\theta},\boldsymbol{\theta})^q\Big]^{1/q}\\
&\le C(q)^2 N^{1/q}\mathbb{E}\Big[U_{d,\le u}(\boldsymbol{\theta},\boldsymbol{\theta})\Big].
\end{aligned}$$

Hence, by Markov's inequality and condition (83), we get for any $\delta > 0$, taking q sufficiently large,

$$\max_{i\in[N(d)]} U_{d}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i) = O_{d,\mathbb{P}}(N(d)^{\delta}) \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\tau_d}[U_{d}(\boldsymbol{\theta},\boldsymbol{\theta})].$$

The proof of Eq. (86) follows from a similar argument. □

Appendix B. Generalization error of random feature model: Proof of Theorem 1

In this section, we prove Theorem 1. The proof in the overparametrized regime is presented in Section B.1. The proof in the underparametrized regime follows from a very similar argument: we will omit it and simply add comments in the overparametrized proof where they differ.

We defer the proofs of some technical results to later sections. Section B.2 proves a key proposition on the structure of the feature matrix $\boldsymbol{Z} = (\sigma_d(\boldsymbol{x}_i;\boldsymbol{\theta}_j))_{i\in[n],j\in[N]}$. Section B.3 gather some technical bounds necessary for the proof of Theorem 1. Finally, Section B.4 contains concentration results on the high degree part of the feature matrix.

B.1. Proof of Theorem 1 in the overparametrized regime

In this section, we prove Theorem 1 in the overparametrized regime. We defer the proofs of some of the technical lemmas and matrix concentration results to Sections B.2, B.3 and B.4. The underparametrized case follows from the same proof with the following mapping $n \leftrightarrow N$, $\mathsf{m} \leftrightarrow \mathsf{M}$ and $\lambda \to \lambda_N = N\lambda/n$. We will add remarks in the proof when a difference arises.

Step 1. Rewrite the y, V, U, Z matrices.

We recall that the random feature ridge regression solution is given by

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}} \Big\{ \sum_{i=1}^{n} \big(y_i - \hat{f}(\boldsymbol{x}_i; \boldsymbol{a})\big)^2 + \frac{\lambda}{N}\|\boldsymbol{a}\|_2^2 \Big\}.$$

Solving for the coefficients yields

$$\hat{\boldsymbol{a}}(\lambda) = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}/N + \lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \dots, y_n)$ and $\boldsymbol{Z} = (Z_{ij})_{i\in[n], j\in[N]} \in \mathbb{R}^{n\times N}$ with $Z_{ij} = \sigma_d(\boldsymbol{x}_i; \boldsymbol{\theta}_j)$. Hence, the prediction function at location $\boldsymbol{x}$ is given by

$$\hat{f}(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda)) = \boldsymbol{y}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}/N + \lambda \mathbf{I}_N)^{-1}\boldsymbol{\sigma}(\boldsymbol{x})/N,$$

where $\boldsymbol{\sigma}(\boldsymbol{x}) = (\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_1), \dots, \sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_N)) \in \mathbb{R}^N$.

Expanding the test error, we get

$$\begin{aligned} R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \equiv& \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_*(\boldsymbol{x}) - \boldsymbol{y}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}/N + \lambda \mathbf{I}_N)^{-1}\boldsymbol{\sigma}(\boldsymbol{x})/N\Big)^2\Big] \\ =& \mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})^2] - 2\boldsymbol{y}^\mathsf{T}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{V}/N + \boldsymbol{y}^\mathsf{T}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}/N^2, \end{aligned}$$

where $\boldsymbol{V} = (V_1, \dots, V_N)^\mathsf{T}$ and $\boldsymbol{U} = (U_{ij})_{ij\in[N]}$ with

$$\begin{aligned} V_i =& \mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)], \\ U_{ij} =& \mathbb{E}_{\boldsymbol{x}}[\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_i)\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}_j)], \end{aligned}$$

and $\hat{\boldsymbol{U}}_\lambda = \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}/N + \lambda \mathbf{I}_N$ is the (rescaled) regularized empirical kernel matrix

$$\hat{U}_{\lambda,ij} = \frac{1}{N}\sum_{k\in[n]} \sigma_d(\boldsymbol{x}_k; \boldsymbol{\theta}_i)\sigma_d(\boldsymbol{x}_k; \boldsymbol{\theta}_j) + \lambda\delta_{ij}.$$

We recall that the eigendecomposition of σ_d is given by

$$\sigma_d(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{\infty} \lambda_{d,k}\psi_k(\boldsymbol{x})\phi_k(\boldsymbol{\theta}).$$

We write the orthogonal decomposition of f_* in this basis as

$$f_*(\boldsymbol{x}) = \sum_{k=1}^{\infty} \hat{f}_{d,k}\psi_k(\boldsymbol{x}),$$

Define

$$\begin{aligned}
\boldsymbol{\psi}_k =&(\psi_k(\boldsymbol{x}_1),\dots,\psi_k(\boldsymbol{x}_n))^{\mathsf{T}} \in \mathbb{R}^n,\\
\boldsymbol{\phi}_k =&(\phi_k(\boldsymbol{\theta}_1),\dots,\phi_k(\boldsymbol{\theta}_N))^{\mathsf{T}} \in \mathbb{R}^N,\\
\boldsymbol{D}_{\le \mathsf{m}} =&\mathrm{diag}(\lambda_{d,1},\lambda_{d,2},\dots,\lambda_{d,\mathsf{m}}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}},\\
\boldsymbol{\psi}_{\le \mathsf{m}} =&(\boldsymbol{\psi}_k(\boldsymbol{x}_i))_{i\in[n],k\in[\mathsf{m}]} \in \mathbb{R}^{n\times\mathsf{m}},\\
\boldsymbol{\phi}_{\le \mathsf{m}} =&(\boldsymbol{\phi}_k(\boldsymbol{\theta}_i))_{i\in[N],k\in[\mathsf{m}]} \in \mathbb{R}^{N\times\mathsf{m}},\\
\hat{\boldsymbol{f}}_{\le \mathsf{m}} =&(\hat{f}_{d,1},\hat{f}_{d,2},\dots,\hat{f}_{d,\mathsf{m}})^{\mathsf{T}} \in \mathbb{R}^{\mathsf{m}}.
\end{aligned} \tag{87}$$

Recall that $\boldsymbol{y} = (y_1,\dots,y_n)^{\mathsf{T}} = \boldsymbol{f} + \boldsymbol{\varepsilon}$ with

$$\begin{aligned}
\boldsymbol{f} =&(f_*(\boldsymbol{x}_1),\dots,f_*(\boldsymbol{x}_n))^{\mathsf{T}}\\
\boldsymbol{\varepsilon} =&(\varepsilon_1,\dots,\varepsilon_n)^{\mathsf{T}}.
\end{aligned}$$

Using the above notations, we can decompose the vectors and matrices $\boldsymbol{f}$, $\boldsymbol{V}$, $\boldsymbol{U}$, and as

$$\begin{aligned}
&\boldsymbol{f} = \boldsymbol{f}_{\le \mathsf{m}} + \boldsymbol{f}_{>\mathsf{m}}, && \boldsymbol{f}_{\le \mathsf{m}} = \boldsymbol{\psi}_{\le \mathsf{m}}\hat{\boldsymbol{f}}_{\le \mathsf{m}}, && \boldsymbol{f}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k}\boldsymbol{\psi}_k,\\
&\boldsymbol{V} = \boldsymbol{V}_{\le \mathsf{m}} + \boldsymbol{V}_{>\mathsf{m}}, && \boldsymbol{V}_{\le \mathsf{m}} = \boldsymbol{\phi}_{\le \mathsf{m}}\boldsymbol{D}_{\le \mathsf{m}}\hat{\boldsymbol{f}}_{\le \mathsf{m}}, && \boldsymbol{V}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k}\lambda_{d,k}\boldsymbol{\phi}_k,\\
&\boldsymbol{U} = \boldsymbol{U}_{\le \mathsf{m}} + \boldsymbol{U}_{>\mathsf{m}}, && \boldsymbol{U}_{\le \mathsf{m}} = \boldsymbol{\phi}_{\le \mathsf{m}}\boldsymbol{D}^2_{\le \mathsf{m}}\boldsymbol{\phi}^{\mathsf{T}}_{\le \mathsf{m}}, && \boldsymbol{U}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda^2_{d,k}\boldsymbol{\phi}_k\boldsymbol{\phi}^{\mathsf{T}}_k,\\
&\boldsymbol{Z} = \boldsymbol{Z}_{\le \mathsf{m}} + \boldsymbol{Z}_{>\mathsf{m}}, && \boldsymbol{Z}_{\le \mathsf{m}} = \boldsymbol{\psi}_{\le \mathsf{m}}\boldsymbol{D}_{\le \mathsf{m}}\boldsymbol{\phi}^{\mathsf{T}}_{\le \mathsf{m}}, && \boldsymbol{Z}_{>\mathsf{m}} = \sum_{k\ge \mathsf{m}+1} \lambda_{d,k}\boldsymbol{\psi}_k\boldsymbol{\phi}^{\mathsf{T}}_k\,.
\end{aligned} \tag{88}$$

Step 2. Decompose the risk.

We decompose the risk with respect to $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$ as follows

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

where

$$\begin{aligned}
T_1 =&\boldsymbol{f}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{V}/N,\\
T_2 =&\boldsymbol{f}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/N^2,\\
T_3 =&\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\varepsilon}/N^2,\\
T_4 =&\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{V}/N,\\
T_5 =&\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/N^2.
\end{aligned} \tag{89}$$

The proof relies on the following key result on the structure of the feature matrix $\boldsymbol{Z}$:

Proposition 6 *(Structure of the feature matrix $\boldsymbol{Z}$). Follow the assumptions and the notations in the proof of Theorem 1 in the overparametrized regime (note in particular that $N \ge n^{1+\delta_0}$ and $n \ge \mathsf{m}^{1+\delta_0}$ for some fixed $\delta_0 > 0$). Consider the singular value decomposition of $\boldsymbol{Z} = (Z_{ij})_{i\in[n],j\in[N]}$ with $Z_{ij} = \sigma_d(\boldsymbol{x}_i;\boldsymbol{\theta}_j)$:*

$$\boldsymbol{Z}/\sqrt{N} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}} = [\boldsymbol{P}_1,\boldsymbol{P}_2]\mathrm{diag}(\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2)[\boldsymbol{Q}_1,\boldsymbol{Q}_2]^{\mathsf{T}} \in \mathbb{R}^{n\times N},$$

where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{Q} \in \mathbb{R}^{N\times n}$, *and* $\boldsymbol{P}_1 \in \mathbb{R}^{n\times \mathsf{m}}$ *and* $\boldsymbol{Q}_1 \in \mathbb{R}^{N\times \mathsf{m}}$ *correspond to the left and right singular vectors associated to the largest* m *singular values* $\boldsymbol{\Lambda}_1$, *while* $\boldsymbol{P}_2 \in \mathbb{R}^{n\times(n-\mathsf{m})}$ *and* $\boldsymbol{Q}_2 \in \mathbb{R}^{N\times(n-\mathsf{m})}$ *correspond to the left and right singular vectors associated to the last* $(n-\mathsf{m})$ *smallest singular values* $\boldsymbol{\Lambda}_2$. *Define* $\kappa_{>\mathsf{m}} = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$.

Then the singular value decomposition has the following properties:

(a) Define $\boldsymbol{\Lambda} = \mathrm{diag}((\sigma_i(\boldsymbol{Z}/\sqrt{N}))_{i\in[n]})$ *the singular values (in non increasing order) of* $\boldsymbol{Z}/\sqrt{N}$. *Then the singular values verify*

$$\sigma_{\min}(\boldsymbol{\Lambda}_1) = \min_{i\in[\mathsf{m}]} \sigma_i(\boldsymbol{Z}/\sqrt{N}) = \kappa_{>\mathsf{m}}^{1/2}\cdot\omega_{d,\mathbb{P}}(1), \tag{90}$$

$$\|\boldsymbol{\Lambda}_2 - \kappa_{>\mathsf{m}}^{1/2}\cdot\mathbf{I}_{n-\mathsf{m}}\|_{\mathrm{op}} = \max_{i=\mathsf{m}+1,\dots,n} \left|\sigma_i(\boldsymbol{Z}/\sqrt{N}) - \kappa_{>\mathsf{m}}^{1/2}\right| = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1). \tag{91}$$

(b) The left and right singular vectors associated to the $(n-\mathsf{m})$ *smallest singular values verify*

$$n^{-1/2}\|\boldsymbol{\psi}_{\le\mathsf{m}}^{\mathsf{T}}\boldsymbol{P}_2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1), \qquad N^{-1/2}\|\boldsymbol{\phi}_{\le\mathsf{m}}^{\mathsf{T}}\boldsymbol{Q}_2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{92}$$

(c) We have

$$N^{-1/2}\|\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{Q}_2\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1). \tag{93}$$

We defer the proof of Proposition 6 to Section B.2.

Remark B.1. Proposition 6 shows that the feature matrix $\boldsymbol{Z} = \boldsymbol{Z}_{\le\mathsf{m}} + \boldsymbol{Z}_{>\mathsf{m}}$ (cf. Eq. (88)) is a spiked matrix, with m spikes with singular values $\boldsymbol{\Lambda}_1$ much larger than $\kappa_{>\mathsf{m}}^{1/2}$ coming from the low-degree part $\boldsymbol{Z}_{\le\mathsf{m}}$ (in particular, Proposition 6.(*b*) shows that the left and right singular vectors of the spikes are approximately spanned by the left and right singular vectors of $\boldsymbol{Z}_{\le\mathsf{m}}$) while the rest of the singular values are approximately constant equal to $\kappa_{>\mathsf{m}}^{1/2}$. The proof of this proposition is based on the following observations:

(*a*) $\boldsymbol{Z}_{\le\mathsf{m}}/\sqrt{N} = \boldsymbol{\psi}_{\le\mathsf{m}}\boldsymbol{D}_{\le\mathsf{m}}\boldsymbol{\phi}_{\le\mathsf{m}}^{\mathsf{T}}/\sqrt{N}$ is a rank m matrix with

  (*i*) $\boldsymbol{\psi}_{\le\mathsf{m}}/\sqrt{n}$ and $\boldsymbol{\phi}_{\le\mathsf{m}}/\sqrt{N}$ are approximately orthogonal matrices (see Eq. (108)).
  (*ii*) $\sqrt{n}|\boldsymbol{D}_{\le\mathsf{m}}| = \mathrm{diag}(\sqrt{n}|\lambda_1|,\dots,\sqrt{n}|\lambda_{\mathsf{m}}|) \succeq \omega_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2})\cdot\mathbf{I}_{\mathsf{m}}$ from condition (19) in Assumption 2.(*a*).

(*b*) The high degree part $\boldsymbol{Z}_{>\mathsf{m}}/\sqrt{N}$ has nearly constant singular values $\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}/N - \kappa_{>\mathsf{m}}\mathbf{I}_n\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}\cdot o_{d,\mathbb{P}}(1)$ and is nearly orthogonal to the span of the right singular vectors of $\boldsymbol{Z}_{\le\mathsf{m}}$, i.e., $\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{\phi}_{\le\mathsf{m}}/N\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1)$ (see Proposition 8 in Section B.4).

Using Proposition 6, we can prove the following list of bounds that will be the main tools for the rest of the proof of Theorem 1.

Proposition 7. *Follow the assumptions and the notations in the proof of Theorem 1 in the overparametrized regime. Then the following bounds hold. (Recall that* $\kappa_{>\mathsf{m}} = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$.*)*

(*a*) *Bounds on* $\hat{\boldsymbol{U}}_\lambda^{-1} = (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}/N + \lambda\mathbf{I}_N)^{-1}$*:*

$$\boldsymbol{\psi}_{\le\mathsf{m}}^{\mathsf{T}}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{\phi}_{\le\mathsf{m}}\boldsymbol{D}_{\le\mathsf{m}}/N = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}, \tag{94}$$

$$\|\boldsymbol{D}_{\leq \mathsf{m}}\boldsymbol{\phi}_{\leq \mathsf{m}}^{\mathsf{T}}\hat{\boldsymbol{U}}_{\lambda}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{m}}/N\|_2 = \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}} \cdot o_{d,\mathbb{P}}(1), \tag{95}$$

$$\sqrt{n}\|\boldsymbol{Z}\hat{\boldsymbol{U}}_{\lambda}^{-1}\boldsymbol{\phi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}/N\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \tag{96}$$

where $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. *Furthermore, we have*

$$\|\boldsymbol{Z}\hat{\boldsymbol{U}}_{\lambda}^{-1}/\sqrt{N}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{-1/2} \cdot O_{d,\mathbb{P}}(1). \tag{97}$$

(*b*) *Bound on* $\boldsymbol{U}_{>\mathsf{m}}$*:*

$$\frac{n}{N}\|\boldsymbol{U}_{>\mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}} \cdot o_{d,\mathbb{P}}(1).$$

(*c*) *Bounds on* $\boldsymbol{f}$*:*

$$\begin{aligned}\|\boldsymbol{f}\|_2 =& \sqrt{n}\|f_*\|_{L^2} \cdot O_{d,\mathbb{P}}(1),\\ \|\boldsymbol{\psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{m}}/n\|_2 =& \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}} \cdot o_{d,\mathbb{P}}(1).\end{aligned}$$

(*d*) *Bound on* $\boldsymbol{V}_{>\mathsf{m}}$*:*

$$\sqrt{\frac{n}{N}}\|\boldsymbol{V}_{>\mathsf{m}}\|_2 = \kappa_{>\mathsf{m}}^{1/2}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1).$$

The proof of Proposition 7 is deferred to Section B.3.

Remark B.2. In the underparametrized case, the proofs and statements of Proposition 6 and Proposition 7.(*a*) and 7.(*c*) are symmetric under the mapping $n \leftrightarrow N$, $\mathsf{m} \leftrightarrow \mathsf{M}$ and $\lambda \to \lambda_N = N\lambda/n$. The bounds in Propositions 7.(*b*) and 7.(*d*) can be easily replaced by

$$\|\boldsymbol{U}_{>\mathsf{M}}\|_{\mathrm{op}} = \kappa_{>\mathsf{M}} \cdot O_{d,\mathbb{P}}(1), \qquad \|\boldsymbol{V}_{>\mathsf{M}}\|_2 = \kappa_{>\mathsf{M}}^{1/2}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1).$$

In order to bound the term T_{22} in Eq. (103), we will further use the following bound

$$\|\hat{\boldsymbol{U}}_{\lambda}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/n\|_{\mathrm{op}} = \kappa_{>\mathsf{M}}^{-1/2} \cdot \|f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1),$$

that we prove in Section B.3.5. It is easy to insert into the new bounds below the aforementioned mapping and check that the underparametrized case follows indeed from the same computation.

The rest of the proof amounts to controlling each term separately using the claims listed in Proposition 7. We will use extensively the following (basic) properties of the operator norm: for $\boldsymbol{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^p$, we have

$$\begin{aligned}\|\boldsymbol{A}\|_{\mathrm{op}} =& \|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\|_{\mathrm{op}}^{1/2} = \|\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}\|_{\mathrm{op}}^{1/2},\\ \|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{op}} \leq& \|\boldsymbol{A}\|_{\mathrm{op}}\|\boldsymbol{B}\|_{\mathrm{op}},\\ \boldsymbol{u}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{v} \leq& \|\boldsymbol{u}\|_2\|\boldsymbol{A}\|_{\mathrm{op}}\|\boldsymbol{v}\|_2.\end{aligned}$$

Step 3. Term T_1.

Let us decompose T_1 into

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

where

$$\begin{aligned}T_{11} =& \boldsymbol{f}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{V}_{\le \mathsf{m}}/N,\\ T_{12} =& \boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{V}_{\le \mathsf{m}}/N,\\ T_{13} =& \boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{V}_{>\mathsf{m}}/N.\end{aligned}$$

Recall that $\boldsymbol{V}_{\le \mathsf{m}} = \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}$ and $\boldsymbol{f}_{\le \mathsf{m}} = \boldsymbol{\psi}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}$. Hence by Eq. (94) in Proposition 7.(a):

$$\begin{aligned}T_{11} =& \hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} (\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}/N) \hat{\boldsymbol{f}}_{\le \mathsf{m}}\\ =& \hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}) \hat{\boldsymbol{f}}_{\le \mathsf{m}}\\ =& \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 \cdot o_{d,\mathbb{P}}(1).\end{aligned} \tag{98}$$

Similarly by Eq. (95) in Proposition 7.(a),

$$\begin{aligned}|T_{12}| =& |\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}/N|\\ \le& \|\boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}_{>\mathsf{m}}/N\|_2 \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2\\ =& \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}} \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1).\end{aligned} \tag{99}$$

Using Proposition 7.(c) and 7.(d) as well as Eq. (97) in Proposition 7.(a), we get

$$\begin{aligned}|T_{13}| = |\boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{V}_{>\mathsf{m}}/N| \le& \|\boldsymbol{f}/\sqrt{n}\|_2 \|(\boldsymbol{Z}/\sqrt{N}) \hat{\boldsymbol{U}}_{\lambda}^{-1}\|_{\mathrm{op}} \cdot \sqrt{n/N} \|\boldsymbol{V}_{>\mathsf{m}}\|_2\\ \le& O_{d,\mathbb{P}}(\|f_*\|_{L^2}) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2}) \cdot o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2} \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2})\\ =& \|f_*\|_{L^2} \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1).\end{aligned} \tag{100}$$

Combining Eqs. (98), (99) and (100) yields

$$T_1 = \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2). \tag{101}$$

Step 4. Term T_2

Recalling $\boldsymbol{U} = \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} + \boldsymbol{U}_{>\mathsf{m}}$, we can decompose T_2 as

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

where

$$\begin{aligned}T_{21} =& (\boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}/N)(\boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}/N),\\ T_{22} =& \boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{U}_{>\mathsf{m}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}/N^2.\end{aligned}$$

From Eqs. (94) and (95) in Proposition 7.(a), we obtain

$$\begin{aligned}\boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}/N =& \boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{\psi}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}/N + \boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_{\lambda}^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}_{>\mathsf{m}}/N\\ =& (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_0) \hat{\boldsymbol{f}}_{\le \mathsf{m}} + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}} \cdot \boldsymbol{\Delta}_1,\end{aligned}$$

where $\|\boldsymbol{\Delta}_1\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, $\|\boldsymbol{\Delta}_2\|_2 = o_{d,\mathbb{P}}(1)$. Hence,

$$T_{21} = \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\delta}}^2) \cdot o_{d,\mathbb{P}}(1). \tag{102}$$

From Eq. (97) in Proposition 7.(a) as well as Proposition 7.(b), 7.(c), the second term is bounded by

$$\begin{aligned} |T_{22}| =& |\boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{U}_{>\mathsf{m}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}/N^2| \\ \le& \|(n/N)\boldsymbol{U}_{>\mathsf{m}}\|_{\mathrm{op}} \|(\boldsymbol{Z}/\sqrt{N})\hat{\boldsymbol{U}}_\lambda^{-1}\|_{\mathrm{op}}^2 \|\boldsymbol{f}/\sqrt{n}\|_2^2 \\ =& o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1}) \cdot O_{d,\mathbb{P}}(\|f_*\|_{L^2}^2) = \|f_*\|_{L^2}^2 \cdot o_{d,\mathbb{P}}(1). \end{aligned} \tag{103}$$

As a result, combining Eqs. (102) and (103) leaves us with

$$T_2 = \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2). \tag{104}$$

Step 5. Terms T_3, T_4 and T_5.

Let us start with the term T_3. Decompose $\boldsymbol{U} = \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} + \boldsymbol{U}_{>\mathsf{m}}$:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\varepsilon}}[T_3]/\sigma_\varepsilon^2 =& \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{U} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2 \\ =& \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2 + \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{U}_{>\mathsf{m}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2. \end{aligned}$$

By Eq. (96) in Proposition 7.(a), and since $\mathsf{m} \le n^{1-\delta_0}$ by Assumption 2.(a), we get

$$\mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2 \le \mathsf{m} \cdot \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}/N\|_{\mathrm{op}}^2 = \frac{\mathsf{m}}{n} \cdot O_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1).$$

By Eq. (97) in Proposition 7.(a) as well as Proposition 7.(b), the second term is bounded by

$$\begin{aligned} \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{U}_{>\mathsf{m}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2 \le& \|(n/N)\boldsymbol{U}_{>\mathsf{m}}\|_{\mathrm{op}} \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-2} \boldsymbol{Z}^{\mathsf{T}}/N\|_{\mathrm{op}}/n \\ =& o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1}) \cdot n^{-1} = o_{d,\mathbb{P}}(1). \end{aligned}$$

Combining these two bounds and using Markov's inequality, we get

$$T_3 = o_{d,\mathbb{P}}(1) \cdot \sigma_\varepsilon^2. \tag{105}$$

Let us consider term T_4. Recall that we can decompose $\boldsymbol{V} = \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} + \boldsymbol{V}_{>\mathsf{m}}$, so

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\varepsilon}}[T_4^2]/\sigma_\varepsilon^2 =& \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}\boldsymbol{V}^{\mathsf{T}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}})/N^2 \\ =& \boldsymbol{V}^{\mathsf{T}} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{Z} \hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}/N^2 \\ \le& 2(\|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}_{\le \mathsf{m}}/N\|_2^2 + \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}_{>\mathsf{m}}/N\|_2^2). \end{aligned}$$

We have by Eq. (96) in Proposition 7.(a),

$$\|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}_{\le \mathsf{m}}/N\|_2 \le \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{\phi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}/N\|_{\mathrm{op}} \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2 = \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1),$$

and by Proposition 7.(d),

$$\begin{aligned} \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1} \boldsymbol{V}_{>\mathsf{m}}/N\|_2 \le& \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}/\sqrt{N}\|_2 \|\boldsymbol{V}_{>\mathsf{m}}/\sqrt{N}\|_2 \\ =& O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2}) \cdot o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2} \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} n^{-1/2}) = \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1). \end{aligned}$$

Combining the two above bounds, we get by Markov's inequality

$$T_4 = o_{d,\mathbb{P}}(1) \cdot \sigma_\varepsilon \|f_*\|_{L^2} = o_{d,\mathbb{P}}(1) \cdot (\sigma_\varepsilon^2 + \|f_*\|_{L^2}^2). \tag{106}$$

Let us consider the last term T_5:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\varepsilon}}[T_5^2]/\sigma_\varepsilon^2 =& \mathrm{tr}(\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}\boldsymbol{f}^\mathsf{T}\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T})/N^4 \\ =& \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}/N^2\|_2^2 \le \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T}\sqrt{n}/N^2\|_{\mathrm{op}}^2 \|\boldsymbol{f}/\sqrt{n}\|_2^2.\end{aligned}$$

By Eq. (95) in Proposition 7.(a), and Proposition 7.(b),

$$\begin{aligned}\|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{U}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^\mathsf{T}\sqrt{n}/N^2\|_{\mathrm{op}} \le& \sqrt{n} \cdot \|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{\phi}_{\le \mathsf{m}}\boldsymbol{D}_{\le \mathsf{m}}/N\|_{\mathrm{op}}^2 + \|\sqrt{n/N^2}\boldsymbol{U}_{>\mathsf{m}}\|_{\mathrm{op}}\|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-2}\boldsymbol{Z}^\mathsf{T}/N\|_{\mathrm{op}} \\ =& o_{d,\mathbb{P}}(1).\end{aligned}$$

Hence, by Proposition 7.(c),

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[T_5^2]/\sigma_\varepsilon^2 = o_{d,\mathbb{P}}(1) \cdot \|\boldsymbol{f}/\sqrt{n}\|_2^2 = \|f_*\|_2^2 \cdot o_{d,\mathbb{P}}(1),$$

which gives by Markov's inequality:

$$T_5 = \sigma_\varepsilon \|f_*\|_{L^2} \cdot o_{d,\mathbb{P}}(1) = (\sigma_\varepsilon^2 + \|f_*\|_{L^2}^2) \cdot o_{d,\mathbb{P}}(1). \tag{107}$$

Step 6. Finish the proof.

Combining Eqs. (101), (104), (105), (106) and (107), we have

$$\begin{aligned}R_{\mathrm{RF}}(f_*, \boldsymbol{X}, \boldsymbol{W}, \lambda) =& \|f_*\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5 \\ =& \|f_*\|_{L^2}^2 - 2\|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2) \\ =& \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2),\end{aligned}$$

which concludes the proof.

B.2. Proof of Proposition 6: structure of the feature matrix Z

Recall the definition $\boldsymbol{Z} = (\sigma_d(\boldsymbol{x}_i; \boldsymbol{\theta}_j))_{i\in[n], j\in[N]}$. Recall the decomposition $\boldsymbol{Z} = \boldsymbol{Z}_{\le \mathsf{m}} + \boldsymbol{Z}_{>\mathsf{m}}$ into a low- and high- degree parts, as per Eq. (88). For convenience, we will consider the normalized quantities

$$\begin{aligned}\tilde{\boldsymbol{Z}} = \boldsymbol{Z}/\sqrt{N}, \quad & \tilde{\boldsymbol{Z}}_{\le \mathsf{m}} = \boldsymbol{Z}_{\le \mathsf{m}}/\sqrt{N}, \quad & \tilde{\boldsymbol{Z}}_{>\mathsf{m}} = \boldsymbol{Z}_{>\mathsf{m}}/\sqrt{N}, \\ \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} = \boldsymbol{\phi}_{\le \mathsf{m}}/\sqrt{N}, \quad & \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}} = \boldsymbol{\psi}_{\le \mathsf{m}}/\sqrt{n}, \quad & \tilde{\boldsymbol{D}}_{\le \mathsf{m}} = \sqrt{n}\boldsymbol{D}_{\le \mathsf{m}}.\end{aligned}$$

In particular, notice that $\hat{\boldsymbol{U}}_\lambda = \tilde{\boldsymbol{Z}}^\mathsf{T}\tilde{\boldsymbol{Z}} + \lambda \mathbf{I}_N$ and $\tilde{\boldsymbol{Z}}_{\le \mathsf{m}} = \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^\mathsf{T}$.

By Proposition 3 applied to $\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}$ and $\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}$ (with assumptions satisfied by Assumption 1.(a) and Assumption 2.(a)), we get

$$\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^\mathsf{T}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_1, \qquad \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^\mathsf{T}\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}} = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_2, \tag{108}$$

with $\|\boldsymbol{\Delta}_i\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ for $i = 1, 2$. Furthermore, by Proposition 8 (stated in Section B.4), we have

$$\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}^\mathsf{T} = \kappa_{>\mathsf{m}} \cdot (\mathbf{I}_n + \boldsymbol{\Delta}_{\boldsymbol{Z}}), \qquad \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2} \cdot o_{d,\mathbb{P}}(1), \tag{109}$$

with $\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\rm op} = o_{d,\mathbb{P}}(1)$ and where we recall $\kappa_{>\mathsf{m}} = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$. Furthermore, Assumption 2.(a) implies that

$$\sigma_{\min}(\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}) = \min_{k\leq \mathsf{m}}\{\sqrt{n}|\lambda_{d,k}|\} = \omega_d(1)\cdot \kappa_{>\mathsf{m}}^{1/2}. \tag{110}$$

Hence, we expect $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Z}}_{\leq \mathsf{m}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}}$ to have m large singular values $\omega_d(1)\cdot\kappa_{>\mathsf{m}}^{1/2}$ associated to $\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}$ with left and right singular vectors spanned approximately by $\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}$ and $\tilde{\boldsymbol{\phi}}_{\leq \mathsf{m}}$, and $n-\mathsf{m}$ small singular values approximately equal to $\kappa_{>\mathsf{m}}^{1/2}$ associated to $\tilde{\boldsymbol{Z}}_{>\mathsf{m}}$.

Proof of Proposition 6. Claim (a)**. Bound on the singular values.**

Using Eqs. (108) and (110), we have

$$\begin{aligned}
\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}^{\mathsf{T}} &= \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\phi}}_{\leq \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}} \\
&= \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}(\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta})\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}} \\
&\succeq \Omega_{d,\mathbb{P}}(1)\cdot \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}^2\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}} \\
&\succeq \kappa_{>\mathsf{m}}\cdot\omega_d(1)\cdot \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}}.
\end{aligned}$$

Furthermore, by $\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}} = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_2$, we deduce that the singular values of $\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}$ are lower bounded as follows

$$\min_{i\in[\mathsf{m}]}\sigma_i(\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}) = \kappa_{>\mathsf{m}}\cdot\omega_{d,\mathbb{P}}(1). \tag{111}$$

By Lemma 8 stated below in Section B.2.1, we have for $i\in[n]$,

$$|\sigma_i(\tilde{\boldsymbol{Z}}) - \sigma_i(\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}})| \leq \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op}. \tag{112}$$

Recalling Eq. (109), $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op} = O_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{m}}^{1/2}$. Hence the first m singular values obey:

$$\sigma_i(\tilde{\boldsymbol{Z}}) \geq \sigma_i(\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}) - \kappa_{>\mathsf{m}}^{1/2}\cdot O_{d,\mathbb{P}}(1). \tag{113}$$

Using Eq. (111) implies $\sigma_{\min}(\boldsymbol{\Lambda}_1) = \min_{i\in[\mathsf{m}]}\sigma_i(\tilde{\boldsymbol{Z}}) = \kappa_{>\mathsf{m}}^{1/2}\cdot\omega_{d,\mathbb{P}}(1)$. This proves Eq. (90).

Using again Eq. (112), the $n-\mathsf{m}$ smallest singular values obey:

$$\max_{i=\mathsf{m}+1,\dots,n}\sigma_i(\tilde{\boldsymbol{Z}}) \leq \kappa_{>\mathsf{m}}^{1/2}\cdot(1+o_{d,\mathbb{P}}(1)). \tag{114}$$

In order to lower bound the $n-\mathsf{m}$ smallest singular values, we lower bound the eigenvalues of $\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathsf{T}}$. We decompose

$$\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathsf{T}} = \tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}}.$$

Recalling Eq. (108) and Eq. (109), we have

$$\begin{aligned}
\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}\tilde{\boldsymbol{Z}}_{\leq \mathsf{m}}^{\mathsf{T}} &= \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}(\tilde{\boldsymbol{\phi}}_{\leq \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_{\leq \mathsf{m}})\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}^{\mathsf{T}} = \tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}(\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_1)\tilde{\boldsymbol{D}}_{\leq \mathsf{m}}\tilde{\boldsymbol{\psi}}_{\leq \mathsf{m}}, \\
\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} &= \kappa_{>\mathsf{m}}\cdot(\mathbf{I}_n + \boldsymbol{\Delta}_{\boldsymbol{Z}}),
\end{aligned}$$

where $\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\rm op} = o_{d,\mathbb{P}}(1)$.

Denote $\boldsymbol{L} = \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_1)^{1/2}$ and $\boldsymbol{T} = \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_1)^{-1/2}$. By Eq. (109), we have $\|\boldsymbol{T}\boldsymbol{T}^{\mathsf{T}}\|_{\rm op} = \kappa_{>\mathsf{m}} \cdot o_{d,\mathbb{P}}(1)$. Combining these remarks leads to the lower bound

$$\begin{aligned}
\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathsf{T}} &= \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} + \boldsymbol{T}\boldsymbol{L}^{\mathsf{T}} + \boldsymbol{L}\boldsymbol{T}^{\mathsf{T}} + \boldsymbol{T}\boldsymbol{T}^{\mathsf{T}} - \boldsymbol{T}\boldsymbol{T}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} \\
&= (\boldsymbol{L} + \boldsymbol{T})(\boldsymbol{L} + \boldsymbol{T})^{\mathsf{T}} + \kappa_{>\mathsf{m}} \cdot (\mathbf{I}_n + \boldsymbol{\Delta}') \\
&\succeq \kappa_{>\mathsf{m}} \cdot (\mathbf{I}_n + \boldsymbol{\Delta}'),
\end{aligned}$$

where $\|\boldsymbol{\Delta}'\|_{\rm op} = o_{d,\mathbb{P}}(1)$. We deduce that

$$\sigma_{\min}(\tilde{\boldsymbol{Z}}) = \min_{i \in [n]} \sigma_i(\tilde{\boldsymbol{Z}}) \ge \kappa_{>\mathsf{m}}^{1/2} \cdot (1 + o_{d,\mathbb{P}}(1)),$$

which combined with Eq. (114) yields Eq. (91).

Part (b)**. Left and right singular vectors.**

Let us prove $\|\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Q}_2\|_{\rm op} = o_{d,\mathbb{P}}(1)$. The proof for $\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2$ follows from the same argument by replacing $\tilde{\boldsymbol{Z}}$ by $\tilde{\boldsymbol{Z}}^{\mathsf{T}}$ and using the bound $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\rm op} = \kappa_{>\mathsf{m}}^{1/2} \cdot o_{d,\mathbb{P}}(1)$, cf. Eq. (109).

Let us consider a sequence $\boldsymbol{u} \in \mathbb{R}^{n-\mathsf{m}}$ (where we keep the dependency on d implicit) such that $\|\boldsymbol{u}\|_2 = 1$ and $\|\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Q}_2 \boldsymbol{u}\|_2 = \|\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Q}_2\|_{\rm op}$. For convenience, denote $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{Q}_2 \boldsymbol{u}$. We have

$$\begin{aligned}
\boldsymbol{u}^{\mathsf{T}} \boldsymbol{\Lambda}_2^2 \boldsymbol{u} &= \boldsymbol{u}^{\mathsf{T}} \boldsymbol{Q}_2^{\mathsf{T}} \tilde{\boldsymbol{Z}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \boldsymbol{Q}_2 \boldsymbol{u} \\
&= \boldsymbol{u}^{\mathsf{T}} \boldsymbol{Q}_2^{\mathsf{T}} (\tilde{\boldsymbol{Z}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{\le \mathsf{m}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{\le \mathsf{m}} + \tilde{\boldsymbol{Z}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}) \boldsymbol{Q}_2 \boldsymbol{u} \\
&= \tilde{\boldsymbol{u}}^{\mathsf{T}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_2) \tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}} + 2 \tilde{\boldsymbol{u}}^{\mathsf{T}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \boldsymbol{u}) + \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}} \boldsymbol{Q}_2 \boldsymbol{u}\|_2^2.
\end{aligned} \tag{115}$$

From step 1, we know $\boldsymbol{u}^{\mathsf{T}} \boldsymbol{\Lambda}_2^2 \boldsymbol{u} = \kappa_{>\mathsf{m}} \cdot O_{d,\mathbb{P}}(1)$. Furthermore,

$$\begin{aligned}
\tilde{\boldsymbol{u}}^{\mathsf{T}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} (\mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}_2) \tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}} &= \Omega_{d,\mathbb{P}}(1) \cdot \|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2^2, \\
\tilde{\boldsymbol{u}}^{\mathsf{T}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \boldsymbol{u}) &\ge - \|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2 \|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op}, \\
\|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op} &\le \|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}\|_{\rm op} \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op} = \kappa_{>\mathsf{m}}^{1/2} \cdot O_{d,\mathbb{P}}(1).
\end{aligned} \tag{116}$$

Therefore, using the bounds (116) in Eq. (115), we get

$$\Omega_{d,\mathbb{P}}(1) \cdot \|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2^2 - 2 \|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2 \|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op} \le \kappa_{>\mathsf{m}} \cdot O_{d,\mathbb{P}}(1).$$

Hence,

$$\|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2 = O_{d,\mathbb{P}}\Big(\max\big(\kappa_{>\mathsf{m}}^{1/2}, \|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\rm op}\big)\Big) = \kappa_{>\mathsf{m}}^{1/2} \cdot O_{d,\mathbb{P}}(1). \tag{117}$$

Using the bound $\|\tilde{\boldsymbol{D}}_{\le \mathsf{m}} \tilde{\boldsymbol{u}}\|_2 = \kappa_{>\mathsf{m}}^{1/2} \cdot \omega_d(1) \cdot \|\tilde{\boldsymbol{u}}\|_2 = \kappa_{>\mathsf{m}}^{1/2} \cdot \omega_d(1) \cdot \|\boldsymbol{Q}_2^{\mathsf{T}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\rm op}$ in Eq. (117), we deduce that $\|\boldsymbol{Q}_2^{\mathsf{T}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\rm op} = o_{d,\mathbb{P}}(1)$. This concludes the proof of Proposition 6.(b).

Part (c)**. Cross-term bound.**

This is a direct application of Lemma 9 (stated below in Section B.2.1) with matrix $\kappa_{>\mathsf{m}}^{-1/2} \tilde{\boldsymbol{Z}} = \kappa_{>\mathsf{m}}^{-1/2} \tilde{\boldsymbol{Z}}_{\le \mathsf{m}} + \kappa_{>\mathsf{m}}^{-1/2} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}$. Indeed, Eq. (111) implies that $\sigma_{\min}(\kappa_{>\mathsf{m}}^{-1/2} \tilde{\boldsymbol{Z}}_{\le \mathsf{m}}) = \omega_{d,\mathbb{P}}(1)$ and Eq. (109) gives $\|\kappa_{>\mathsf{m}}^{-1} \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} - \mathbf{I}_n\|_{\rm op} = o_{d,\mathbb{P}}(1)$. Furthermore, the right singular vectors $\boldsymbol{V}_0$ of $\tilde{\boldsymbol{Z}}_{\le \mathsf{m}}$ are spanned by the left singular vectors of $\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}$. From Eq. (109), we have $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\rm op} = \kappa_{>\mathsf{m}}^{1/2} \cdot o_{d,\mathbb{P}}(1)$. Combined with $\|\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} - \mathbf{I}_{\mathsf{m}}\|_{\rm op} = o_{d,\mathbb{P}}(1)$, we get $\|\kappa_{>\mathsf{m}}^{-1/2} \tilde{\boldsymbol{Z}}_{>\mathsf{m}} \boldsymbol{V}_0\|_{\rm op} = o_{d,\mathbb{P}}(1)$. □

B.2.1. Auxiliary lemmas

We recall the following classical perturbation theory result, which we quote from [44, Page 102, Section 3].

Theorem 8 *(Sin(Θ) theorem for rectangular matrices [44]). Let $\boldsymbol{A}_0$ be a $n \times N$-matrix with singular value decomposition*

$$\boldsymbol{A}_0 = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^{\mathsf{T}},$$

where $\boldsymbol{U}_0 \in \mathbb{R}^{n\times \mathsf{m}}$, $\boldsymbol{V}_0 \in \mathbb{R}^{N\times \mathsf{m}}$ verify $\mathsf{m} \le \min(n, N)$ and $\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{U}_0 = \boldsymbol{V}_0^{\mathsf{T}}\boldsymbol{V}_0 = \mathbf{I}_{\mathsf{m}}$, and we denoted $\boldsymbol{\Sigma}_0 = \mathrm{diag}((\sigma_i(\boldsymbol{A}_0))_{i\in[\mathsf{m}]})$= the singular values. Let $\boldsymbol{M}$ be a perturbation $n \times N$-matrix and consider $\boldsymbol{B} = \boldsymbol{A}_0 + \boldsymbol{M}$ with singular value decomposition

$$\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q} = [\boldsymbol{P}_1, \boldsymbol{P}_2]\mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)[\boldsymbol{Q}_1, \boldsymbol{Q}_2]^{\mathsf{T}},$$

where $\boldsymbol{P}_1 \in \mathbb{R}^{n\times \mathsf{m}}, \boldsymbol{Q}_1 \in \mathbb{R}^{N\times \mathsf{m}}, \boldsymbol{P}_2 \in \mathbb{R}^{n\times(n-\mathsf{m})}, \boldsymbol{Q}_2 \in \mathbb{R}^{N\times(n-\mathsf{m})}$. Assume that $\sigma_{\min}(\boldsymbol{\Lambda}_1) > 0$. Then

$$\max(\|(\mathbf{I}_n - \boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}})\boldsymbol{P}_1\|_{\mathrm{op}}, \|(\mathbf{I}_N - \boldsymbol{V}_0\boldsymbol{V}_0^{\mathsf{T}})\boldsymbol{Q}_1\|_{\mathrm{op}}) \le \frac{\max(\|\boldsymbol{M}\boldsymbol{Q}_1\|_{\mathrm{op}}, \|\boldsymbol{M}^{\mathsf{T}}\boldsymbol{P}_1\|_{\mathrm{op}})}{\sigma_{\min}(\boldsymbol{\Lambda}_1)}. \tag{118}$$

Lemma 8 *(Weyl's inequality). Consider $\boldsymbol{A}_0, \boldsymbol{M} \in \mathbb{R}^{n\times N}$ and define $\boldsymbol{B} = \boldsymbol{A}_0 + \boldsymbol{M}$. Then for any $i \in [\min(n, N)]$, we have*

$$|\sigma_i(\boldsymbol{B}) - \sigma_i(\boldsymbol{A}_0)| \le \|\boldsymbol{M}\|_{\mathrm{op}}. \tag{119}$$

The next lemma implies that the projection of the noise matrix $\boldsymbol{M}$ on the top left singular vectors of the full matrix is approximately in the space orthogonal to the right singular vectors.

Lemma 9 *(Null space of right singular vectors). Let $\{N(d)\}_{d\ge1}$, $\{n(d)\}_{d\ge1}$ and $\{\mathsf{m}(d)\}_{d\ge1}$ be three sequences of integers. For convenience, we denote $N = N(d)$, $n = n(d)$ and $\mathsf{m} = \mathsf{m}(d)$. Assume that $N \ge n + \mathsf{m}$ and $n \ge \mathsf{m}$. Consider the following sequence of random spiked matrices:*

$$\boldsymbol{B} := \boldsymbol{B}(d) = \boldsymbol{A}_0 + \boldsymbol{M} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^{\mathsf{T}} + \boldsymbol{M} \in \mathbb{R}^{n\times N},$$

where $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^{\mathsf{T}}$ is the singular value decomposition of the rank m matrix $\boldsymbol{A}_0$ with $\boldsymbol{U}_0 \in \mathbb{R}^{n\times \mathsf{m}}$, $\boldsymbol{V}_0 \in \mathbb{R}^{N\times \mathsf{m}}$ and $\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{U}_0 = \boldsymbol{V}_0^{\mathsf{T}}\boldsymbol{V}_0 = \mathbf{I}_{\mathsf{m}}$, and $\boldsymbol{\Sigma}_0 = \mathrm{diag}((\sigma_{0,i}(\boldsymbol{A}_0))_{i\in[\mathsf{m}]}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}$ are the singular values. Further assume that

(a) $\sigma_{\min}(\boldsymbol{A}_0) = \min_{i\in[\mathsf{m}]} \sigma_{0,i}(\boldsymbol{A}_0) = \omega_{d,\mathbb{P}}(1)$,
(b) $\|\boldsymbol{M}\boldsymbol{V}_0\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$,
(c) $\|\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}} - \mathbf{I}_n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$.

Denote $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}} = [\boldsymbol{P}_1, \boldsymbol{P}_2]\mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)[\boldsymbol{Q}_1, \boldsymbol{Q}_2]^{\mathsf{T}}$ the singular value decomposition of $\boldsymbol{B}$ where $\boldsymbol{P}_1 \in \mathbb{R}^{n\times \mathsf{m}}$ and $\boldsymbol{Q}_1 \in \mathbb{R}^{N\times \mathsf{m}}$ correspond to the left and right singular vectors associated to the first m singular values $\boldsymbol{\Lambda}_1$, while $\boldsymbol{P}_2 \in \mathbb{R}^{n\times(n-\mathsf{m})}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{N\times(n-\mathsf{m})}$ correspond to the left and right singular vectors associated to the last $(n - \mathsf{m})$ singular values $\boldsymbol{\Lambda}_2$.

Then we have

$$\|\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{M}\boldsymbol{Q}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{120}$$

Proof of Lemma 9. Step 1. Simplification of the problem.

Without loss of generality, we can choose an orthonormal basis in $\mathbb{R}^N$ so that, in that basis

$$\boldsymbol{V}_0 = \begin{bmatrix} \mathbf{I}_{\mathsf{m}} \\ \mathbf{0}_{N-\mathsf{m},\mathsf{m}} \end{bmatrix}, \qquad \boldsymbol{M} = [\boldsymbol{M}_1 \quad \boldsymbol{M}_2 \quad \mathbf{0}_{n,N-(n+\mathsf{m})}], \tag{121}$$

where $\boldsymbol{M}_1 \in \mathbb{R}^{n\times\mathsf{m}}$ and $\boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$. Because the space corresponding to the last $N-(n+\mathsf{m})$ coordinates of the row is in the right null space of both $\boldsymbol{A}_0$ and $\boldsymbol{M}$, we can forget about them and consider –without loss of generality– $\boldsymbol{M} = [\boldsymbol{M}_1, \boldsymbol{M}_2] \in \mathbb{R}^{n\times(n+\mathsf{m})}$, $N = n+\mathsf{m}$.

From the assumption $\|\boldsymbol{M}\boldsymbol{V}_0\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, we have

$$\|\boldsymbol{M}_1\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{122}$$

Furthermore, from the assumption $\|\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}} - \mathbf{I}_n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$,

$$\|\boldsymbol{M}_2\boldsymbol{M}_2^{\mathsf{T}} - \mathbf{I}_n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{123}$$

Step 2. There exists an orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}$ such that $\|\boldsymbol{P}_1 - \boldsymbol{U}_0\boldsymbol{R}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$.

Recall that $\boldsymbol{\Lambda}_1 = \mathrm{diag}((\sigma_{1,i}(\boldsymbol{B}))_{i\in[\mathsf{m}]})$. By Lemma 8, we have for any $i \in [\mathsf{m}]$,

$$|\sigma_{1,i}(\boldsymbol{B}) - \sigma_{0,i}(\boldsymbol{A}_0)| \le \|\boldsymbol{M}\|_{\mathrm{op}}.$$

Using the assumption (a) that $\sigma_{\min}(\boldsymbol{A}_0) = \omega_{d,\mathbb{P}}(1)$ and assumption (c) $\|\boldsymbol{M}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$, we deduce that

$$\sigma_{\min}(\boldsymbol{\Lambda}_1) = \omega_{d,\mathbb{P}}(1). \tag{124}$$

Furthermore $\|\boldsymbol{M}\boldsymbol{Q}_1\|_{\mathrm{op}} \le \|\boldsymbol{M}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ and similarly $\|\boldsymbol{M}^{\mathsf{T}}\boldsymbol{P}_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. We can therefore apply Theorem 8 which gives

$$\|(\mathbf{I}_n - \boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}})\boldsymbol{P}_1\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Denote by $\boldsymbol{U}_{0,\perp} \in \mathbb{R}^{n\times(n-\mathsf{m})}$ a matrix such that $[\boldsymbol{U}_0, \boldsymbol{U}_{0,\perp}]$ is orthogonal. The last equation implies that $\|\boldsymbol{U}_{0,\perp}^{\mathsf{T}}\boldsymbol{P}_1\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Further,

$$\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{P}_1 = \mathbf{I}_{\mathsf{m}} - \boldsymbol{P}_1^{\mathsf{T}}(\mathbf{I}_n - \boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}})\boldsymbol{P}_1,$$

which shows that $\|\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{P}_1 - \mathbf{I}_{\mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. This implies $\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{P}_1$ is an approximately orthogonal matrix. Namely, let its singular value decomposition be $\boldsymbol{U}_0^{\mathsf{T}}\boldsymbol{P}_1 = \boldsymbol{R}_1\boldsymbol{S}\boldsymbol{R}_2^{\mathsf{T}}$. Then, by defining the orthogonal matrix $\boldsymbol{R} := \boldsymbol{R}_1\boldsymbol{R}_2^{\mathsf{T}} \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}$, we have $\|\boldsymbol{P}_1 - \boldsymbol{U}_0\boldsymbol{R}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$.

Step 3. The null space of the right eigenvectors $\boldsymbol{Q}$.

Let us explicitly describe the null space of $\boldsymbol{Q} \in \mathbb{R}^{(n+\mathsf{m})\times n}$ (recall that we removed the $N-(n+\mathsf{m})$ last coordinates of the columns). Consider $\boldsymbol{N}_1 \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}$ a rank m matrix and write $\boldsymbol{N}_2 \in \mathbb{R}^{n\times\mathsf{m}}$ as a function of $\boldsymbol{N}_1$ such that $\ker(\boldsymbol{Q})$ is spanned by the columns of the matrix $\boldsymbol{N} = \begin{bmatrix} \boldsymbol{N}_1 \\ \boldsymbol{N}_2 \end{bmatrix} \in \mathbb{R}^{(n+\mathsf{m})\times\mathsf{m}}$, i.e., $\boldsymbol{B}\boldsymbol{N} = \mathbf{0}$, that is

$$[\boldsymbol{U}_0\boldsymbol{\Sigma}_0 + \boldsymbol{M}_1 \quad \boldsymbol{M}_2]\begin{bmatrix} \boldsymbol{N}_1 \\ \boldsymbol{N}_2 \end{bmatrix} = (\boldsymbol{U}_0\boldsymbol{\Sigma}_0 + \boldsymbol{M}_1)\boldsymbol{N}_1 + \boldsymbol{M}_2\boldsymbol{N}_2 = \mathbf{0}.$$

Projecting on the two orthogonal subspaces $\boldsymbol{U}_0$ and $\boldsymbol{U}_{0,\perp}$, this is equivalent to

$$\boldsymbol{N}_1 = -(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1}\boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\boldsymbol{N}_2, \qquad \boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{M}_1\boldsymbol{N}_1 = -\boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{M}_2\boldsymbol{N}_2. \tag{125}$$

Let us do the following reparametrization $\boldsymbol{N}_2 = \boldsymbol{M}_2^{-1}\tilde{\boldsymbol{N}}_2$ and fix $\boldsymbol{N}_1 = -(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1}$. Then Eq. (125) gives

$$\boldsymbol{U}_0^\mathsf{T}\tilde{\boldsymbol{N}}_2 = \mathbf{I}_\mathsf{m}, \qquad \boldsymbol{U}_{0,\perp}^\mathsf{T}\tilde{\boldsymbol{N}}_2 = \boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{M}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1},$$

which gives $\tilde{\boldsymbol{N}}_2 = \boldsymbol{U}_0 + \boldsymbol{U}_{0,\perp}\boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{M}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1}$, and

$$\begin{aligned}\boldsymbol{N}_1 &= -(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1},\\ \boldsymbol{N}_2 &= \boldsymbol{M}_2^{-1}\boldsymbol{U}_0 + \boldsymbol{M}_2^{-1}\boldsymbol{U}_{0,\perp}\boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{M}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1}.\end{aligned}$$

By the assumption $\lambda_{\min}(\boldsymbol{\Sigma}_0) = \omega_{d,\mathbb{P}}(1)$ and Eq. (122), we have $\|(\boldsymbol{\Sigma}_0 + \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_1)^{-1}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Furthermore, from Eq. (123), we have $\|\boldsymbol{M}_2^{-1} - \boldsymbol{M}_2^\mathsf{T}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. We deduce that

$$\|\boldsymbol{N}^\mathsf{T} - \begin{bmatrix}\mathbf{0}_{n,\mathsf{m}} & \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\end{bmatrix}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{126}$$

Step 4. Concluding the proof.

By construction, $\boldsymbol{N}^\mathsf{T}\boldsymbol{Q} = \mathbf{0}$ and using Eq. (126), we get

$$\|\boldsymbol{N}^\mathsf{T}\boldsymbol{Q} - \begin{bmatrix}\mathbf{0}_{n,\mathsf{m}} & \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\end{bmatrix}\boldsymbol{Q}\|_{\mathrm{op}} = \|\begin{bmatrix}\mathbf{0}_{n,\mathsf{m}} & \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\end{bmatrix}\boldsymbol{Q}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{127}$$

Furthermore using step 2 and recalling that $\|\boldsymbol{M}_1\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$,

$$\|\boldsymbol{P}_1^\mathsf{T}\boldsymbol{M} - \begin{bmatrix}\mathbf{0}_{n,\mathsf{m}} & \boldsymbol{R}^\mathsf{T}\boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\end{bmatrix}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{128}$$

Combining Eqs. (127) and (128) yield

$$\|\boldsymbol{R}\boldsymbol{P}_1^\mathsf{T}\boldsymbol{M}\boldsymbol{Q} - \begin{bmatrix}\mathbf{0}_{n,\mathsf{m}} & \boldsymbol{U}_0^\mathsf{T}\boldsymbol{M}_2\end{bmatrix}\boldsymbol{Q}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

and $\|\boldsymbol{P}_1^\mathsf{T}\boldsymbol{M}\boldsymbol{Q}\|_{\mathrm{op}} = \|\boldsymbol{R}\boldsymbol{P}_1^\mathsf{T}\boldsymbol{M}\boldsymbol{Q}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, which concludes the proof. □

B.3. Proof of Proposition 7: technical bounds in the overparametrized regime

We prove the claims of this proposition in a different order than stated.

B.3.1. Proof of claim (c)

First, notice that $\mathbb{E}[\|\boldsymbol{f}\|_2^2] = n\|f_*\|_{L^2}^2$. Hence, by Markov's inequality, $\|\boldsymbol{f}\|_2^2 = n\|f_*\|_{L^2}^2 \cdot O_{d,\mathbb{P}}(1)$.

Let us now consider $\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{f}_{>\mathsf{m}}/n$. For any $\eta > 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{f}_{>\mathsf{m}}\|_2^2\right]/n^2 =& \mathbb{E}_{\boldsymbol{x}}\Big[\Big(\sum_{u\ge \mathsf{m}+1} \hat{f}_u \boldsymbol{\psi}_u^{\mathsf{T}}\Big)\boldsymbol{\psi}_{\le \mathsf{m}}\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}}\Big(\sum_{v\ge \mathsf{m}+1} \hat{f}_v \boldsymbol{\psi}_v\Big)\Big]/n^2 \\
=& \sum_{u,v\ge \mathsf{m}+1}\sum_{s=0}^{\mathsf{m}}\sum_{i,j\in[n]} \Big\{\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_j)\psi_v(\boldsymbol{x}_j)\Big]/n^2\Big\}\hat{f}_u\hat{f}_v \\
=& \sum_{u,v\ge \mathsf{m}+1}\sum_{s=0}^{\mathsf{m}}\sum_{i\in[n]} \Big\{\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_v(\boldsymbol{x}_i)\Big]/n^2\Big\}\hat{f}_u\hat{f}_v \\
=& \frac{1}{n}\sum_{s=0}^{\mathsf{m}} \mathbb{E}_{\boldsymbol{x}}\Big[\big(\mathsf{P}_{>\mathsf{m}} f_*(\boldsymbol{x})\big)^2 \psi_s(\boldsymbol{x})^2\Big] \le \frac{1}{n}\sum_{s=0}^{\mathsf{m}} \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2 \|\psi_s\|_{L^{(4+2\eta)/\eta}}^2 \\
\le& \tilde{C}(\eta)\frac{\mathsf{m}}{n}\|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2,
\end{aligned}$$

where the last inequality uses the hypercontractivity assumption of Assumption 1.(a):

$$\|\psi_s\|_{L^{(4+2\eta)/\eta}}^2 = \mathbb{E}_{\boldsymbol{x}}[\psi_s(\boldsymbol{x})^{2\cdot\frac{2+\eta}{\eta}}]^{\frac{2\eta}{4+2\eta}} \le C((2+\eta)/\eta)\mathbb{E}_{\boldsymbol{x}}[\psi_s(\boldsymbol{x})^2] = C((2+\eta)/\eta),$$

and $\tilde{C}(\eta) = C((2+\eta)/\eta)$. By Markov's inequality (using $\mathsf{m} \le n^{1-\delta_0}$ in Assumption 2.(a) for some fixed $\delta_0 > 0$), we get

$$\|\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{f}_{>\mathsf{m}}/n\|_2 = o_{d,\mathbb{P}}(1)\cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}.$$

B.3.2. Proof of Proposition 7.(a)

Throughout the proof, we will generically denote $\boldsymbol{\Delta}$ any matrix with $\|\boldsymbol{\Delta}\|_{\rm op} = o_{d,\mathbb{P}}(1)$. In particular, $\boldsymbol{\Delta}$ can change from line to line. For convenience, we will use the notations introduced in Section B.2.

Step 0. Bound $\|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\|_{\rm op} = \kappa_{>\mathsf{m}}^{-1/2}\cdot O_{d,\mathbb{P}}(1)$.

Recall the definition $\hat{\boldsymbol{U}}_\lambda = \tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{Z}} + \lambda \mathbf{I}_N$ and the singular value decomposition $\tilde{\boldsymbol{Z}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$. Hence, we can rewrite

$$\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1} = \boldsymbol{P}\frac{\boldsymbol{\Lambda}}{\boldsymbol{\Lambda}^2+\lambda}\boldsymbol{Q}^{\mathsf{T}},$$

where we denoted by a slight abuse of notation $\boldsymbol{\Lambda}/(\boldsymbol{\Lambda}^2+\lambda) := \mathrm{diag}((\Lambda_i/(\Lambda_i^2+\lambda))_{i\in[n]})$. From Proposition 6.$(a)$, $\sigma_{\min}(\boldsymbol{\Lambda}) = \kappa_{>\mathsf{m}}^{1/2}\cdot(1+o_{d,\mathbb{P}}(1))$. We deduce that

$$\|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\|_{\rm op} = \kappa_{>\mathsf{m}}^{-1/2}\cdot O_{d,\mathbb{P}}(1).$$

Step 1. Bound $\|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}} - \mathbf{I}_{\mathsf{m}}\|_{\rm op} = o_{d,\mathbb{P}}(1)$.

First notice that $\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}} = \tilde{\boldsymbol{Z}}_{\le \mathsf{m}}^{\mathsf{T}}(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger = (\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}_{>\mathsf{m}})^{\mathsf{T}}(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger$. Furthermore, by Eq. (108), we have $(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger = \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}} + \boldsymbol{\Delta}$. Hence,

$$\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}} = \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger - \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}}(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger. \tag{129}$$

Let us decompose the first term along the large singular values $\boldsymbol{\Lambda}_1$ and small singular values $\boldsymbol{\Lambda}_2$:

$$\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \hat{\boldsymbol{U}}_\lambda^{-1} \tilde{\boldsymbol{Z}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger = \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P} \frac{\boldsymbol{\Lambda}^2}{\boldsymbol{\Lambda}^2 + \lambda} \boldsymbol{P}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger$$
$$= \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_1 \frac{\boldsymbol{\Lambda}_1^2}{\boldsymbol{\Lambda}_1^2 + \lambda} \boldsymbol{P}_1^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger + \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2 \frac{\boldsymbol{\Lambda}_2^2}{\boldsymbol{\Lambda}_2^2 + \lambda} \boldsymbol{P}_2^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger.$$

From Eqs. (90) and (92) in Proposition 6 and the assumption in the theorem $\lambda = O_d(1) \cdot \kappa_{>\mathsf{m}}$, we have

$$\left\| \frac{\boldsymbol{\Lambda}_1^2}{\boldsymbol{\Lambda}_1^2 + \lambda} - \mathbf{I}_{\mathsf{m}} \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1), \qquad \|\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Hence,

$$\left\| \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2 \frac{\boldsymbol{\Lambda}_2^2}{\boldsymbol{\Lambda}_2^2 + \lambda} \boldsymbol{P}_2^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger \right\|_{\mathrm{op}} \le \|\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2\|_{\mathrm{op}} \|(\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

and

$$\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_1 \frac{\boldsymbol{\Lambda}_1^2}{\boldsymbol{\Lambda}_1^2 + \lambda} \boldsymbol{P}_1^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger = \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_1 \boldsymbol{P}_1^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger + \boldsymbol{\Delta} = \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{P}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger + \boldsymbol{\Delta}' = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta}',$$

where $\|\boldsymbol{\Delta}\|_{\mathrm{op}}, \|\boldsymbol{\Delta}'\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. We deduce

$$\left\| \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \hat{\boldsymbol{U}}_\lambda^{-1} \tilde{\boldsymbol{Z}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger - \mathbf{I}_{\mathsf{m}} \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{130}$$

Consider the second term in Eq. (129):

$$\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \hat{\boldsymbol{U}}_\lambda^{-1} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger = \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_1 \frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2 + \lambda} \boldsymbol{Q}_1^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger + \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2 \frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2 + \lambda} \boldsymbol{Q}_2^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger.$$

Using Eq. (90) in Proposition 6, we have $\sigma_{\min}(\boldsymbol{\Lambda}_1) = \kappa_{>\mathsf{m}}^{1/2} \cdot \omega_{d,\mathbb{P}}(1)$. Then, recalling that $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2} \cdot O_{d,\mathbb{P}}(1)$, we obtain

$$\left\| \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_1 \frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2 + \lambda} \boldsymbol{Q}_1^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger \right\|_{\mathrm{op}} \le \|\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}}\|_{\mathrm{op}} \|\boldsymbol{\Lambda}_1/(\boldsymbol{\Lambda}_1^2 + \lambda)\|_{\mathrm{op}} \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\mathrm{op}} \|\tilde{\psi}_{\le \mathsf{m}} + \boldsymbol{\Delta}\|_{\mathrm{op}}$$
$$= O_{d,\mathbb{P}}(1) \cdot o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2}) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2}) \cdot O_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1).$$

By Eqs. (91) and (92) in Proposition 6, we get

$$\left\| \tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2 \frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2 + \lambda} \boldsymbol{Q}_2^{\mathsf{T}} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger \right\|_{\mathrm{op}} \le \|\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{P}_2\|_{\mathrm{op}} \|\boldsymbol{\Lambda}_2/(\boldsymbol{\Lambda}_2^2 + \lambda)\|_{\mathrm{op}} \|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\mathrm{op}} \|\tilde{\psi}_{\le \mathsf{m}} + \boldsymbol{\Delta}\|_{\mathrm{op}}$$
$$= o_{d,\mathbb{P}}(1) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2}) \cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2}) \cdot O_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1).$$

We deduce that

$$\|\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \hat{\boldsymbol{U}}_\lambda^{-1} \tilde{\boldsymbol{Z}}_{>\mathsf{m}}^{\mathsf{T}} (\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{131}$$

Combining Eqs. (130) and (131) into Eq. (129) yields

$$\tilde{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \tilde{\boldsymbol{Z}} \hat{\boldsymbol{U}}_\lambda^{-1} \tilde{\phi}_{\le \mathsf{m}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} = \mathbf{I}_{\mathsf{m}} + \boldsymbol{\Delta},$$

where $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$.

Step 2. Bound $\|\tilde{\boldsymbol{D}}_{\le \mathsf{m}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{m}}/\sqrt{n}\|_2 = \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$.

Let us denote $\tilde{\boldsymbol{f}}_{>\mathsf{m}} = \boldsymbol{f}_{>\mathsf{m}}/\sqrt{n}$ for convenience. Let us use again that $\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}} = (\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}_{>\mathsf{m}})^{\mathsf{T}}(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger$:

$$\tilde{\boldsymbol{D}}_{\le \mathsf{m}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}} = (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}} - (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}. \tag{132}$$

First notice that because $\|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}} - \mathbf{I}_{\mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, we have $\|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\boldsymbol{P}_2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ in Proposition 6.(*b*) that implies $\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\boldsymbol{P}_2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ (for example by looking at the singular value decomposition of $\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}$). Similarly $\|\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_2 = \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$ (Proposition 7.(*c*)) implies $\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_2 = \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$. Using the same argument as in the proof of Eq. (130), we have

$$\begin{aligned}
&\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_2\\
&= \left\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\boldsymbol{P}\frac{\boldsymbol{\Lambda}^2}{\boldsymbol{\Lambda}^2+\lambda}\boldsymbol{P}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\right\|_2\\
&\le \|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_2 + o_{d,\mathbb{P}}(1)\cdot\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\|_{\mathrm{op}}\|\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_2 + \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}\cdot O_{d,\mathbb{P}}(1)\cdot\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\boldsymbol{P}_2\|_{\mathrm{op}}\\
&= \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1).
\end{aligned} \tag{133}$$

The second term (132) can be decomposed as

$$(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\hat{\boldsymbol{U}}_\lambda^{-1}\tilde{\boldsymbol{Z}}^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}} = (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_1\frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2+\lambda}\boldsymbol{P}_1^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}} + (\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_2\frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2+\lambda}\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}.$$

Using that $\sigma_{\min}(\boldsymbol{\Lambda}_1) = \kappa_{>\mathsf{m}}^{1/2}\cdot\omega_{d,\mathbb{P}}(1)$ and $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2}\cdot O_{d,\mathbb{P}}(1)$ yields

$$\begin{aligned}
\left\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_1\frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2+\lambda}\boldsymbol{P}_1^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\right\|_{\mathrm{op}} &\le \|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_1\|_{\mathrm{op}}\|\boldsymbol{\Lambda}_1/(\boldsymbol{\Lambda}_1^2+\lambda)\|_{\mathrm{op}}\|\boldsymbol{P}_1^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_{\mathrm{op}}\\
&= O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2})\cdot o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2})\cdot O_{d,\mathbb{P}}(\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2})\\
&= \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}\cdot o_{d,\mathbb{P}}(1).
\end{aligned} \tag{134}$$

For the second term, recall that $(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger = \tilde{\boldsymbol{\psi}}_{\le \mathsf{m}}^{\mathsf{T}} + \boldsymbol{\Delta}$ and introduce $\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{P}_1\boldsymbol{P}_1^{\mathsf{T}} + \boldsymbol{P}_2\boldsymbol{P}_2^{\mathsf{T}} = \mathbf{I}_n$:

$$\begin{aligned}
&\left\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_2\frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2+\lambda}\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\right\|_{\mathrm{op}}\\
&= \left\|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger[\boldsymbol{P}_1\boldsymbol{P}_1^{\mathsf{T}} + \boldsymbol{P}_2\boldsymbol{P}_2^{\mathsf{T}}]\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_2\frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2+\lambda}\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\right\|_{\mathrm{op}}\\
&\le \|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\boldsymbol{P}_1\|_{\mathrm{op}}\|\boldsymbol{P}_1^{\mathsf{T}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_2\|_{\mathrm{op}}\|\boldsymbol{\Lambda}_2/(\boldsymbol{\Lambda}_2^2+\lambda)\|_{\mathrm{op}}\|\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_{\mathrm{op}}\\
&\quad + \|(\tilde{\boldsymbol{\psi}}_{\le \mathsf{m}})^\dagger\boldsymbol{P}_2\|_{\mathrm{op}}\|\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\boldsymbol{Q}_2\|_{\mathrm{op}}\|\boldsymbol{\Lambda}_2/(\boldsymbol{\Lambda}_2^2+\lambda)\|_{\mathrm{op}}\|\boldsymbol{P}_2^{\mathsf{T}}\tilde{\boldsymbol{f}}_{>\mathsf{m}}\|_{\mathrm{op}}\\
&= o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2})\cdot O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2})\cdot\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}\\
&= \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}\cdot o_{d,\mathbb{P}}(1),
\end{aligned} \tag{135}$$

where we used Eq. (93) in Proposition 6, and $\sigma_{\min}(\boldsymbol{\Lambda}_2) = \kappa_{>\mathsf{m}}^{-1/2}\cdot\Omega_{d,\mathbb{P}}(1)$ to obtain the second to last line. Combining Eqs. (133), (134) and (135) yields the result.

Step 3. Bound $\sqrt{n}\|\boldsymbol{Z}\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{\phi}_{\le \mathsf{m}}\boldsymbol{D}_{\le \mathsf{m}}/N\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$.

First notice that $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1)$ implies $\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^\dagger\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1)$, where we used that $\|\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} - \mathbf{I}_{\mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$.

Using $\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}} \tilde{\boldsymbol{D}}_{\le \mathsf{m}} = (\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}_{>\mathsf{m}})(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^{\dagger}$, we have

$$\begin{aligned}
\|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_{\lambda}^{-1}\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}\tilde{\boldsymbol{D}}_{\le \mathsf{m}}\|_{\mathrm{op}} &\le \|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_{\lambda}^{-1}\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^{\dagger}\|_{\mathrm{op}} + \|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_{\lambda}^{-1}\tilde{\boldsymbol{Z}}_{>\mathsf{m}}(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^{\dagger}\|_{\mathrm{op}} \\
&\le \|\boldsymbol{\Lambda}^2/(\boldsymbol{\Lambda}^2+\lambda)\|_{\mathrm{op}}\|(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^{\dagger}\|_{\mathrm{op}} + \|\tilde{\boldsymbol{Z}}\hat{\boldsymbol{U}}_{\lambda}^{-1}\|_{\mathrm{op}}\|\tilde{\boldsymbol{Z}}_{>\mathsf{m}}(\tilde{\boldsymbol{\phi}}_{\le \mathsf{m}}^{\mathsf{T}})^{\dagger}\|_{\mathrm{op}} \\
&= O_{d,\mathbb{P}}(1) + O_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{-1/2}) \cdot o_{d,\mathbb{P}}(\kappa_{>\mathsf{m}}^{1/2}) \\
&= O_{d,\mathbb{P}}(1),
\end{aligned}$$

which concludes the proof of the claims in Proposition 7.(a).

B.3.3. Proof of Proposition 7.(b)

Denote

$$\begin{aligned}
\boldsymbol{D}_{\mathsf{m}:\mathsf{M}} &= \mathrm{diag}(\lambda_{d,\mathsf{m}+1}, \lambda_{d,\mathsf{m}+2}, \dots, \lambda_{d,\mathsf{M}}) \in \mathbb{R}^{(\mathsf{M}-\mathsf{m})\times(\mathsf{M}-\mathsf{m})}, \\
\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}} &= (\phi_k(\boldsymbol{\theta}_i))_{i\in[N],k=\mathsf{m}+1,\dots,\mathsf{M}} \in \mathbb{R}^{N\times(\mathsf{M}-\mathsf{m})}.
\end{aligned}$$

Applying Theorem 7 to $\boldsymbol{U}_{>\mathsf{m}}$ (where the assumptions are satisfied by Assumptions 1.(a) and (b) and Assumption 2.(a)), we get with Assumption 1.(d),

$$\boldsymbol{U}_{>\mathsf{m}} = \boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}\boldsymbol{D}_{\mathsf{m}:\mathsf{M}}^2\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}^{\mathsf{T}} + \kappa_{>\mathsf{M}}(\mathbf{I}_N + \boldsymbol{\Delta}),$$

where $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\kappa_{>\mathsf{M}} = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{M}})$. By assumption, we have $N \ge n^{1+\delta_0}$ for some fixed $\delta_0 > 0$ and therefore

$$\frac{n}{N}\|\kappa_{>\mathsf{M}}(\mathbf{I}_N + \boldsymbol{\Delta})\|_{\mathrm{op}} = \kappa_{>\mathsf{M}} \cdot o_{d,\mathbb{P}}(1). \tag{136}$$

By Proposition 3 (assumptions satisfied by Assumptions 1.(a) and 2.(a)), we get

$$\|\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}^{\mathsf{T}}\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}/N - \mathbf{I}_{\mathsf{M}-\mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Furthermore, by Assumption 2.(a), we have $n^{1+\delta_0} \cdot \|\mathbb{H}_{d,>\mathsf{m}}\|_{\mathrm{op}} = O_d(1) \cdot \kappa_{>\mathsf{m}}$ for a fixed $\delta_0 > 0$. Therefore $n\|\boldsymbol{D}_{\mathsf{m}:\mathsf{M}}^2\|_{\mathrm{op}} = \kappa_{>\mathsf{m}} \cdot o_d(1)$. Hence,

$$\frac{n}{N}\|\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}\boldsymbol{D}_{\mathsf{m}:\mathsf{M}}^2\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}^{\mathsf{T}}\|_{\mathrm{op}} \le \|\boldsymbol{\phi}_{\mathsf{m}:\mathsf{M}}/\sqrt{N}\|_{\mathrm{op}}^2\|n\boldsymbol{D}_{\mathsf{m}:\mathsf{M}}^2\|_{\mathrm{op}} = \kappa_{>\mathsf{m}} \cdot o_{d,\mathbb{P}}(1). \tag{137}$$

Combining Eqs. (136) and (137) yields

$$\frac{n}{N}\|\boldsymbol{U}_{>\mathsf{m}}\|_{\mathrm{op}} = \kappa_{>\mathsf{m}} \cdot o_{d,\mathbb{P}}(1).$$

B.3.4. Proof of Proposition 7.(d)

Recall

$$\boldsymbol{V}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k}\lambda_{d,k}\boldsymbol{\phi}_k.$$

Taking the expectation over $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$, we get

$$\frac{n}{N}\mathbb{E}[\|\boldsymbol{V}_{>\mathsf{m}}\|_2^2] = n\sum_{k\ge\mathsf{m}+1}\lambda_{d,k}^2\hat{f}_{d,k}^2 \le n \cdot \|\mathbb{H}_{d,>\mathsf{m}}\|_{\mathrm{op}} \cdot \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}^2.$$

From condition (19) in Assumption 2.(a), we have $n^{1+\delta_0}\|\mathbb{H}_{d,>\mathsf{m}}\|_{\rm op} = O_d(1)\cdot\kappa_{>\mathsf{m}}$, and we conclude, via Markov's inequality, that

$$\sqrt{\frac{n}{N}}\|\boldsymbol{V}_{>\mathsf{m}}\|_2 = \sqrt{\kappa_{>\mathsf{m}}}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}\cdot o_{d,\mathbb{P}}(1).$$

B.3.5. Bounds in the underparametrized regime

In the underparametrized case, we further prove the following lemma.

Lemma 10. *Follow the assumptions of Theorem 1 in the underparametrized case as well as the notations in Section B.1. Then, we have*

$$\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/n\|_2 = \kappa_{>\mathsf{M}}^{1/2}\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1), \tag{138}$$

$$\|\hat{U}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/n\|_{\rm op} = \kappa_{>\mathsf{M}}^{-1/2}\cdot o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}+\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}). \tag{139}$$

Proof of Lemma 10. Step 1. Bound $\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/n\|_2 = \kappa_{>\mathsf{M}}^{1/2}\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$.

Recall the decomposition of $\boldsymbol{Z}_{>\mathsf{M}}$ in the eigenbasis of functions:

$$\boldsymbol{Z}_{>\mathsf{M}} = \sum_{k=\mathsf{M}+1}^{\infty}\lambda_{d,k}\boldsymbol{\psi}_k\boldsymbol{\phi}_k^{\mathsf{T}}.$$

Consider the expected square norm (with respect to $\boldsymbol{\Theta} = (\boldsymbol{\theta}_j)_{j\in[N]}$)

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\big[\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}\|_2^2\big] &= \sum_{k,\ell=\mathsf{M}+1}^{\infty}\lambda_{d,k}\lambda_{d,\ell}\mathbb{E}_{\boldsymbol{\Theta}}[\boldsymbol{f}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{\psi}_{d,k}\boldsymbol{\phi}_{d,k}^{\mathsf{T}}\boldsymbol{\phi}_{d,\ell}\boldsymbol{\psi}_{d,\ell}\boldsymbol{f}_{>\mathsf{M}}]\\ &= N\sum_{k=\mathsf{M}+1}^{\infty}\lambda_{d,k}^2(\boldsymbol{f}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{\psi}_{d,k})^2,\end{aligned}$$

where we used that $\mathbb{E}_{\boldsymbol{\Theta}}[\boldsymbol{\phi}_{d,k}^{\mathsf{T}}\boldsymbol{\phi}_{d,\ell}] = N\delta_{k,\ell}$ by orthonormality of $\{\phi_{d,k}\}_{k\ge1}$. Expanding with respect to the $\boldsymbol{x}_i$'s, we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\big[\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}\|_2^2\big] &= N\sum_{i\in[n]}\Big\{H_{d,>\mathsf{M}:\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i)[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2 + H_{d,>\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i)[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2\Big\}\\ &\quad + N\sum_{i\neq j\in[n]}\sum_{k=\mathsf{M}+1}^{\infty}\lambda_{d,k}^2\psi_{d,k}(\boldsymbol{x}_i)\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)\cdot\psi_{d,k}(\boldsymbol{x}_j)\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_j),\end{aligned}$$

where we recall

$$H_{d,\mathsf{M}:u}(\boldsymbol{x}_i,\boldsymbol{x}_i) = \sum_{k=\mathsf{M}+1}^{u}\lambda_{d,k}^2\psi_{d,k}(\boldsymbol{x}_i)^2,$$

$$H_{d,>u}(\boldsymbol{x}_i,\boldsymbol{x}_i) = \sum_{k=u+1}^{\infty}\lambda_{d,k}^2\psi_{d,k}(\boldsymbol{x}_i)^2.$$

Consider the first term depending on $H_{d,\mathsf{M}:\mathsf{m}}$. Using the same computation as in the proof of Proposition 7.(c) and Lemma 6 (with the hypercontractivity assumption up to $u\ge\mathsf{m}$ of Assumption 1.(a)), by Hölder's inequality we have for the q

$$\mathbb{E}\big[H_{d,\mathsf{M}:\mathsf{m}}(\boldsymbol{x},\boldsymbol{x})[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x})]^2\big] \le \|H_{d,\mathsf{M}:\mathsf{m}}\|_{L^{1+2/\eta}}\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2 \le C(1+2/\eta)^2\cdot\mathbb{E}_{\boldsymbol{x}}[H_{d,\mathsf{M}:\mathsf{m}}(\boldsymbol{x},\boldsymbol{x})]\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2.$$

We deduce by Markov's inequality that the first term is bounded by

$$\sum_{i\in[n]} H_{d,\mathsf{M}:\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i)[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2 = O_{d,\mathbb{P}}(1)\cdot n\cdot\mathrm{Tr}(\mathbb{H}_{d,\mathsf{M}:\mathsf{m}})\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2. \tag{140}$$

For the second term, recall that by Assumption 1.(d), we have

$$\max_{\boldsymbol{x}_i\in[n]} H_{d,>\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i) = O_{d,\mathbb{P}}(1)\cdot\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}).$$

Hence

$$\sum_{i\in[n]} H_{d,>\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i)[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2 = O_{d,\mathbb{P}}(1)\cdot\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})\cdot\sum_{i\in[n]}[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2,$$

and by Markov's inequality

$$\sum_{i\in[n]} H_{d,>\mathsf{m}}(\boldsymbol{x}_i,\boldsymbol{x}_i)[\mathsf{P}_{>\mathsf{M}}f_*(\boldsymbol{x}_i)]^2 = O_{d,\mathbb{P}}(1)\cdot n\cdot\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^2}^2. \tag{141}$$

Taking the expectation of the third term gives

$$\begin{aligned} n(n-1)\sum_{k=\mathsf{M}+1}^{\infty}\lambda_{d,k}^2\mathbb{E}\big[\psi_{d,k}(\boldsymbol{x})[\mathsf{P}_{>\mathsf{M}}f_*(x)]\big]^2 &= n(n-1)\sum_{k=\mathsf{M}+1}^{\infty}\lambda_{d,k}^2\hat{f}_{d,k}^2 \\ &\le n(n-1)\|\mathbb{H}_{d,>u}\|_{\mathrm{op}}\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^2}^2. \end{aligned} \tag{142}$$

Merging Eqs. (140), (141) and (142), we get

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}\big[\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/n\|_2^2\big] &\le \frac{N}{n}\cdot O_{d,\mathbb{P}}(1)\cdot\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{M}})\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2 + N\|\mathbb{H}_{d,>u}\|_{\mathrm{op}}\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^2}^2 \\ &= o_d(1)\cdot\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{M}})\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2, \end{aligned}$$

where we used Assumption 2.(b) ($N\cdot\|\mathbb{H}_{d,>u}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(N^{-\delta_0})\mathrm{Tr}(\mathbb{H}_{d,>u})$ as well as $n\ge N^{1+\delta_0}$ for a fixed $\delta_0>0$). Using Markov's inequality proves Eq. (138).

Step 2. Bound on $\|\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/n\|_2$.

By Proposition 6.(a) in the underparametrized case, we have

$$\hat{\boldsymbol{U}}_\lambda^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}/n = \boldsymbol{Q}_1\frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2+\lambda}\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{f}/\sqrt{n} + \boldsymbol{Q}_2\frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2+\lambda}\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{f}/\sqrt{n}, \tag{143}$$

where $\sigma_{\min}(\boldsymbol{\Lambda}_1) = \omega_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{M}}^{1/2}$ and $\sigma_{\min}(\boldsymbol{\Lambda}_2) = \kappa_{>\mathsf{M}}^{1/2}\cdot(1+o_{d,\mathbb{P}}(1))$. In particular, this shows that

$$\Big\|\boldsymbol{Q}_1\frac{\boldsymbol{\Lambda}_1}{\boldsymbol{\Lambda}_1^2+\lambda}\boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{f}/\sqrt{n}\Big\|_2 \le \sigma_{\min}(\boldsymbol{\Lambda}_1)^{-1}\|\boldsymbol{f}/\sqrt{n}\|_2 \le o_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{M}}^{-1/2}\cdot\|f_*\|_{L^2}. \tag{144}$$

For the second term (143), decompose $\boldsymbol{f} = \boldsymbol{f}_{\le\mathsf{M}} + \boldsymbol{f}_{>\mathsf{M}}$. Recall $\boldsymbol{f}_{\le\mathsf{M}} = \boldsymbol{\psi}_{\le\mathsf{M}}\hat{\boldsymbol{f}}_{\le\mathsf{M}}$. By Proposition 6.($b$), we have $\|\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{\psi}_{\le\mathsf{m}}/\sqrt{n}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Furthermore, using Eq. (138), namely $\|\boldsymbol{Z}_{>\mathsf{M}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/n\|_2 = \kappa_{>\mathsf{M}}^{1/2}\cdot\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$, we get $\|\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/\sqrt{n}\|_{\mathrm{op}} = \|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}\cdot o_{d,\mathbb{P}}(1)$. We deduce

$$\begin{aligned}\left\|\boldsymbol{Q}_2\frac{\boldsymbol{\Lambda}_2}{\boldsymbol{\Lambda}_2^2+\lambda}\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{f}/\sqrt{n}\right\|_2 &\le \sigma_{\min}(\boldsymbol{\Lambda}_2)^{-1}(\|\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{f}_{\le \mathsf{M}}/\sqrt{n}\|_2+\|\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{M}}/\sqrt{n}\|_2)\\ &= O_{d,\mathbb{P}}(\kappa_{>\mathsf{M}}^{-1/2})\cdot o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}+\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}})\\ &= \kappa_{>\mathsf{M}}^{-1/2}(\|f_*\|_{L^2}+\|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}})\cdot o_{d,\mathbb{P}}(1).\end{aligned} \tag{145}$$

Combining Eqs. (144) and (145) yields Eq. (139). □

B.4. Concentration of the random feature kernel matrix $Z^{\mathsf{T}}Z$

We recall the following standard result on concentration of random matrices with independent rows:

Lemma 11 *([42] Theorem 5.45). Let $\boldsymbol{A}$ be a $p\times q$ matrix whose rows $\boldsymbol{a}_i$ are independent random vectors in $\mathbb{R}^q$ with common second moment matrix $\boldsymbol{\Sigma}=\mathbb{E}[\boldsymbol{a}_i\otimes\boldsymbol{a}_i]$. Let $\Gamma:=\mathbb{E}[\max_{i\in[p]}\|\boldsymbol{a}_i\|_2^2]$. Then*

$$\mathbb{E}\left[\|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}/p-\boldsymbol{\Sigma}\|_{\rm op}\right]\le\max(\|\boldsymbol{\Sigma}\|_{\rm op}^{1/2}\eta,\eta^2),$$

where $\eta=C\sqrt{\frac{\Gamma\log(\min(p,q))}{p}}$ and C is an absolute constant.

We will also use the following corollary for asymmetric matrices:

Corollary 1. *Let $\boldsymbol{A}$ be a $n\times N$ matrix whose rows $\boldsymbol{a}_i$ are independent random vectors in $\mathbb{R}^N$ with common second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}}=\mathbb{E}[\boldsymbol{a}_i\otimes\boldsymbol{a}_i]$. Let $\boldsymbol{B}$ be a $n\times\mathsf{m}$ matrix whose rows $\boldsymbol{b}_i$ are independent random vectors in $\mathbb{R}^{\mathsf{m}}$ with common second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}}=\mathbb{E}[\boldsymbol{b}_i\otimes\boldsymbol{b}_i]$. Let $\Gamma_{\boldsymbol{a}}:=\mathbb{E}[\max_{i\in[n]}\|\boldsymbol{a}_i\|_2^2]$ and $\Gamma_{\boldsymbol{b}}:=\mathbb{E}[\max_{i\in[n]}\|\boldsymbol{b}_i\|_2^2]$. Denote $\boldsymbol{\Sigma}_{\boldsymbol{ab}}=\mathbb{E}[\boldsymbol{a}_i\otimes\boldsymbol{b}_i]$. Then,*

$$\mathbb{E}\left[\|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}/n-\boldsymbol{\Sigma}_{\boldsymbol{ab}}\|_{\rm op}\right]\le\max\left((\|\boldsymbol{\Sigma}_{\boldsymbol{a}}\|_{\rm op}+\|\boldsymbol{\Sigma}_{\boldsymbol{b}}\|_{\rm op})^{1/2}\eta,\eta^2\right), \tag{146}$$

where $\eta=C\sqrt{\frac{(\Gamma_{\boldsymbol{a}}+\Gamma_{\boldsymbol{b}})\log(\min(n,N,\mathsf{m}))}{n}}$ and C is an absolute constant.

Proof of Corollary 1. Define $\boldsymbol{C}=[\boldsymbol{A},\boldsymbol{B}]\in\mathbb{R}^{n\times(N+\mathsf{m})}$ whose rows $\boldsymbol{c}_i=[\boldsymbol{a}_i,\boldsymbol{b}_i]$ are independent random vectors in $\mathbb{R}^{N+\mathsf{m}}$ with common second matrix $\boldsymbol{\Sigma}_{\boldsymbol{c}}=\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{a}} & \boldsymbol{\Sigma}_{\boldsymbol{ab}}\\ \boldsymbol{\Sigma}_{\boldsymbol{ba}} & \boldsymbol{\Sigma}_{\boldsymbol{b}}\end{bmatrix}$. By Lemma 11, we have

$$\mathbb{E}\left[\|\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}/n-\boldsymbol{\Sigma}_{\boldsymbol{c}}\|_{\rm op}\right]\le\max(\|\boldsymbol{\Sigma}_{\boldsymbol{c}}\|_{\rm op}^{1/2}\eta,\eta^2),$$

where $\eta=C\sqrt{\frac{\Gamma\log(\min(n,N+\mathsf{m}))}{n}}$ with

$$\Gamma=\mathbb{E}[\max_{i\in[n]}\|\boldsymbol{c}_i\|_2^2]\le\mathbb{E}[\max_{i\in[n]}\|\boldsymbol{a}_i\|_2^2]+\mathbb{E}[\max_{i\in[n]}\|\boldsymbol{b}_i\|_2^2]\le\Gamma_{\boldsymbol{a}}+\Gamma_{\boldsymbol{b}}.$$

Notice that $\|\boldsymbol{\Sigma}_{\boldsymbol{c}}\|_{\rm op}\le C(\|\boldsymbol{\Sigma}_{\boldsymbol{a}}\|_{\rm op}+\|\boldsymbol{\Sigma}_{\boldsymbol{b}}\|_{\rm op})$, and

$$\|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}/n-\boldsymbol{\Sigma}_{\boldsymbol{ab}}\|_{\rm op}\le\|\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}/n-\boldsymbol{\Sigma}_{\boldsymbol{c}}\|_{\rm op}.$$

Combining these bounds yields Eq. (146). □

Consider the feature matrix $\boldsymbol{Z}=(\sigma_d(\boldsymbol{x}_i;\boldsymbol{\theta}_j))_{i\in[n],j\in[N]}$. We recall the decomposition $\boldsymbol{Z}=\boldsymbol{Z}_{\le\mathsf{m}}+\boldsymbol{Z}_{>\mathsf{m}}$ into a low and high degree parts:

$$\boldsymbol{Z}_{\le \mathsf{m}} = \boldsymbol{\psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}^{\mathsf{T}}, \qquad \boldsymbol{Z}_{>\mathsf{m}} = \sum_{k \ge \mathsf{m}+1} \lambda_{d,k} \boldsymbol{\psi}_k \boldsymbol{\phi}_k^{\mathsf{T}}.$$

We prove the following concentration result on $\boldsymbol{Z}_{>\mathsf{m}}$.

Proposition 8 *(Concentration of $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}$ matrix). Consider the overparametrized case $N(d) \ge n(d)^{1+\delta_0}$ for some fixed $\delta_0 > 0$. Let $\{\sigma_d\}_{d\ge 1}$ be a sequence of activation functions satisfying the feature map concentration (Assumption 1) and the spectral gap (Assumption 2) at level $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d))\}_{d\ge 1}$. Then, we have*

$$\frac{\boldsymbol{Z}_{>\mathsf{m}} \boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}}{N} = \kappa_{>\mathsf{m}} \cdot (\mathbf{I}_n + \boldsymbol{\Delta}_{\boldsymbol{Z}}), \tag{147}$$

where $\kappa_{>\mathsf{m}} = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$ and $\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Furthermore,

$$\Big\| \frac{\boldsymbol{Z}_{>\mathsf{m}} \boldsymbol{\phi}_{\le \mathsf{m}}}{N} \Big\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}^{1/2} \cdot o_{d,\mathbb{P}}(1). \tag{148}$$

Proof of Proposition 8. For convenience, we will drop the subscript d.

Step 1. Bound on $\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}/N - \kappa_{>\mathsf{m}}\mathbf{I}_n\|_{\mathrm{op}}$.

Denote $\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{Z}_{>\mathsf{m}} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_N] \in \mathbb{R}^{n\times N}$ with $\boldsymbol{a}_i = (\sigma_{>\mathsf{m}}(\boldsymbol{x}_1;\boldsymbol{\theta}_i), \dots, \sigma_{>\mathsf{m}}(\boldsymbol{x}_n;\boldsymbol{\theta}_i)) \in \mathbb{R}^n$. Conditioned on $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, the rows $\boldsymbol{a}_i$ are independent with common second moment matrix

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{a}_i \otimes \boldsymbol{a}_i] = \boldsymbol{H}_{>\mathsf{m}},$$

where $\boldsymbol{H}_{>\mathsf{m}} = (H_{>\mathsf{m},ij})_{1\le i,j\le N}$ with $H_{>\mathsf{m},ij} = \mathbb{E}_{\boldsymbol{\theta}}[\sigma_{>\mathsf{m}}(\boldsymbol{x}_i;\boldsymbol{\theta})\sigma_{>\mathsf{m}}(\boldsymbol{x}_j;\boldsymbol{\theta})]$. By applying Theorem 7 to the kernel matrix $\boldsymbol{H}_{>\mathsf{m}}$ (assumptions satisfied by Assumptions 1 and 2), we have $\boldsymbol{H}_{>\mathsf{m}} = \kappa_{>\mathsf{m}} \cdot (\mathbf{I}_n + \boldsymbol{\Delta}_H)$ where $\|\boldsymbol{\Delta}_H\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Therefore it is sufficient to show that

$$\Big\| \frac{\boldsymbol{Z}_{>\mathsf{m}} \boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}}{N} - \boldsymbol{H}_{>\mathsf{m}} \Big\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Let us decompose $\sigma_{>\mathsf{m}}$ into a low- and high- degree parts $\sigma_{>\mathsf{m}} = \sigma_{\mathsf{m}:u} + \sigma_{>u}$ (recall that $u(d) > \mathsf{m}(d)$):

$$\sigma_{\mathsf{m}:u}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=\mathsf{m}+1}^{u} \lambda_{d,k}\psi_k(\boldsymbol{x})\phi_k(\boldsymbol{\theta}),$$

$$\sigma_{>u}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=u+1}^{\infty} \lambda_{d,k}\psi_k(\boldsymbol{x})\phi_k(\boldsymbol{\theta}).$$

Let $\underline{\boldsymbol{a}}_i = (\sigma_{\mathsf{m}:u}(\boldsymbol{x}_1;\boldsymbol{\theta}_i), \dots, \sigma_{\mathsf{m}:u}(\boldsymbol{x}_n;\boldsymbol{\theta}_i)) \in \mathbb{R}^n$ and $\overline{\boldsymbol{a}}_i = (\sigma_{>u}(\boldsymbol{x}_1;\boldsymbol{\theta}_i), \dots, \sigma_{>u}(\boldsymbol{x}_n;\boldsymbol{\theta}_i)) \in \mathbb{R}^n$, $\boldsymbol{a}_i = \underline{\boldsymbol{a}}_i + \overline{\boldsymbol{a}}_i$. Then

$$\Gamma = \mathbb{E}_{\boldsymbol{\theta}}[\max_{i\in[N]} \|\boldsymbol{a}_i\|_2^2] \le 2\mathbb{E}_{\boldsymbol{\theta}}[\max_{i\in[N]} \|\overline{\boldsymbol{a}}_i\|_2^2] + 2\mathbb{E}_{\boldsymbol{\theta}}[\max_{i\in[N]} \|\underline{\boldsymbol{a}}_i\|_2^2].$$

Let $q > 0$ be an integer as in Assumption 1.(c). We have

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\max_{i\in[N]} \|\overline{\boldsymbol{a}}_i\|_2^2\Big] \le \mathbb{E}_{\boldsymbol{\theta}}\Big[\max_{i\in[N]} \|\overline{\boldsymbol{a}}_i\|_2^q\Big]^{1/q} \le N^{1/q}\mathbb{E}_{\boldsymbol{\theta}}[\|\underline{\boldsymbol{a}}_i\|_2^{2q}]^{1/q}.$$

By Jensen's inequality and Assumption 1.(c), there exists a fixed $\delta_0 > 0$ such that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\Big[\|\underline{\boldsymbol{a}}_i\|_2^{2q}\Big] &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\Big[\Big(\sum_{j\in[n]}\sigma_{>u}(\boldsymbol{x}_j;\boldsymbol{\theta})^2\Big)^q\Big]\\ &\le n^{q-1}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\Big[\sum_{j\in[n]}\sigma_{>u}(\boldsymbol{x}_j;\boldsymbol{\theta})^{2q}\Big]\\ &\le n^q\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}[\sigma_{>u}(\boldsymbol{x}_j;\boldsymbol{\theta})^{2q}] = O_d(1)\cdot n^{q(1+2\delta_0)}\cdot\kappa_{>u}^q,\end{aligned}$$

where $\kappa_{>u} = \mathrm{Tr}(\mathbb{H}_{>u}) = \sum_{k=u+1}^{\infty}\lambda_k^2$. Hence, by Markov's inequality, we get

$$\mathbb{E}_{\boldsymbol{\theta}}[\max_{i\in[N]}\|\overline{\boldsymbol{a}}_i\|_2^2] = O_{d,\mathbb{P}}(1)\cdot N^{1/q}n^{1+2\delta_0}\cdot\kappa_{>u}. \tag{149}$$

Similarly, by the hypercontractivity assumption (Assumption 1.(a)), we obtain

$$\mathbb{E}_{\boldsymbol{x}}\Big[\mathbb{E}_{\boldsymbol{\theta}}\Big[\max_{i\in[N]}\|\underline{\boldsymbol{a}}_i\|_2^2\Big]\Big] \le C_qN^{1/q}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[\|\underline{\boldsymbol{a}}_i\|_2^2\big] = C_qN^{1/q}n\cdot\kappa_{\mathsf{m}:u},$$

where $\kappa_{\mathsf{m}:u} = \sum_{k=\mathsf{m}+1}^{u}\lambda_k^2$. Hence, by Markov's inequality,

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\max_{i\in[N]}\|\underline{\boldsymbol{a}}_i\|_2^2\Big] = O_{d,\mathbb{P}}(1)\cdot N^{1/q}n\cdot\kappa_{\mathsf{m}:u}. \tag{150}$$

Combining Eqs. (149) and (150) yields

$$\Gamma_{\boldsymbol{a}} = O_{d,\mathbb{P}}(1)\cdot N^{1/q}n^{1+2\delta_0}\kappa_{>\mathsf{m}}.$$

We can therefore apply Lemma 11. Recalling $\|\boldsymbol{H}_{>\mathsf{m}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{m}}$, we have

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}/N - \boldsymbol{H}_{>\mathsf{m}}\big\|_{\mathrm{op}}\Big] \le O_{d,\mathbb{P}}(1)\cdot\max(\kappa_{>\mathsf{m}}^{1/2}\eta,\eta^2),$$

with $\eta = \big(\kappa_{>\mathsf{m}}N^{1/q-1}n^{1+2\delta_0}\log(N)\big)^{1/2} = \kappa_{>\mathsf{m}}^{1/2}\cdot o_{d,\mathbb{P}}(1)$ by the choice of q in Assumption 1.(c). We conclude

$$\big\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{Z}_{>\mathsf{m}}^{\mathsf{T}}/N - \boldsymbol{H}_{>\mathsf{m}}\big\|_{\mathrm{op}} = \kappa_{>\mathsf{m}}\cdot o_{d,\mathbb{P}}(1).$$

Step 2. Bound on $\|\boldsymbol{Z}_{>\mathsf{m}}\boldsymbol{\phi}_{\le\mathsf{m}}/N\|_{\mathrm{op}}$.

Consider $\boldsymbol{B} = \kappa_{>\mathsf{m}}^{1/2}\boldsymbol{\phi}_{\le\mathsf{m}} = [\boldsymbol{b}_1,\dots,\boldsymbol{b}_N]^{\mathsf{T}}\mathbb{R}^{N\times\mathsf{m}}$ where $\boldsymbol{b}_i = \kappa_{>\mathsf{m}}^{1/2}[\boldsymbol{\phi}_1(\boldsymbol{\theta}_i),\dots,\boldsymbol{\phi}_{\mathsf{m}}(\boldsymbol{\theta}_i)] \in \mathbb{R}^{\mathsf{m}}$ are independent rows with second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}} = \mathbb{E}[\boldsymbol{b}_i\otimes\boldsymbol{b}_i] = \kappa_{>\mathsf{m}}\mathbf{I}_{\mathsf{m}}$. Furthermore, by the hypercontractivity assumption (Assumption 1.(a)), we have

$$\Gamma_{\boldsymbol{b}} = \mathbb{E}_{\boldsymbol{\theta}}\Big[\max_{i\in[N]}\|\boldsymbol{b}_i\|_2^2\Big] \le C_qN^{1/q}\mathbb{E}_{\boldsymbol{\theta}}\Big[\|\boldsymbol{b}_i\|_2^2\Big] = C_qN^{1/q}\mathsf{m}\cdot\kappa_{>\mathsf{m}}.$$

Notice that $\mathbb{E}[\boldsymbol{a}_i\otimes\boldsymbol{b}_i] = 0$. Furthermore, recalling the previous step, we have $\|\boldsymbol{\Sigma}_{\boldsymbol{a}}\|_{\mathrm{op}} = \|\boldsymbol{H}_{>\mathsf{m}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{m}}$ and

$$\begin{aligned}\|\boldsymbol{\Sigma}_{\boldsymbol{a}}\|_{\mathrm{op}} + \|\boldsymbol{\Sigma}_{\boldsymbol{b}}\|_{\mathrm{op}} &= O_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{m}},\\ \Gamma_{\boldsymbol{a}} + \Gamma_{\boldsymbol{b}} &= O_{d,\mathbb{P}}(1)\cdot N^{1/q}n^{1+2\delta_0}\cdot\kappa_{>\mathsf{m}}.\end{aligned}$$

Then by Corollary 1 applied to $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}/N$ and recalling the assumption on q in Assumption 1.(c),

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{Z}_{\mathsf{m}}\boldsymbol{\phi}_{\le\mathsf{m}}/N\|_{\mathrm{op}}] = o_{d,\mathbb{P}}(1)\cdot\kappa_{>\mathsf{m}}^{1/2},$$

which concludes the proof by Markov's inequality. □

Appendix C. Generalization error of kernel ridge regression: Proof of Theorem 5

In this section, we prove Theorem 5. We will then prove a different version of the same theorem in Section C.2, under somewhat different assumptions. Namely, we will relax Assumption 4.(*c*) and instead impose a gap condition on the eigenvalues of the kernel.

C.1. Proof of Theorem 5

In this section, we prove Theorem 9. Throughout the proof, we will denote $\boldsymbol{\Delta}$ any matrix with $\|\boldsymbol{\Delta}\|_{\rm op} = o_{d,\mathbb{P}}(1)$. In particular, $\boldsymbol{\Delta}$ can change from one line to line. We defer the proofs of some more technical results to Section C.1.1.

Step 1. Expressing the risk in terms of empirical kernel matrix.

Recall that the KRR estimator is given by

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

where $\boldsymbol{y} = (y_1, \dots, y_n)$ and $\boldsymbol{H} = (H(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\in[n]}$, $\boldsymbol{h}(\boldsymbol{x}) = (H_d(\boldsymbol{x}, \boldsymbol{x}_1), \dots, H_d(\boldsymbol{x}, \boldsymbol{x}_n)) \in \mathbb{R}^n$. The resulting test error is

$$\begin{aligned} R_{\rm KR}(f_*, \boldsymbol{X}, \lambda) &\equiv \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_*(\boldsymbol{x}) - \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{h}(\boldsymbol{x})\Big)^2\Big] \\ &= \mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})^2] - 2\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{E} + \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{E} = (E_1, \dots, E_n)^{\mathsf{T}}$, $\boldsymbol{M} = (M_{ij})_{ij\in[n]}$ and $\boldsymbol{H} = (H_{ij})_{ij\in[n]}$ are defined by

$$\begin{aligned} E_i &= \mathbb{E}_{\boldsymbol{x}}[f_*(\boldsymbol{x})H_d(\boldsymbol{x}, \boldsymbol{x}_i)], \\ M_{ij} &= \mathbb{E}_{\boldsymbol{x}}[H_d(\boldsymbol{x}_i, \boldsymbol{x})H_d(\boldsymbol{x}_j, \boldsymbol{x})], \\ H_{ij} &= H_d(\boldsymbol{x}_i, \boldsymbol{x}_j). \end{aligned}$$

We recall that the eigendecomposition of H_d is given by

$$H_d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{\infty} \lambda_{d,k}^2 \psi_k(\boldsymbol{x})\psi_k(\boldsymbol{y}).$$

We write the orthogonal decomposition of f_* in the basis $\{\psi_k\}_{k\ge 1}$ as

$$f_*(\boldsymbol{x}) = \sum_{k=1}^{\infty} \hat{f}_{d,k}\psi_k(\boldsymbol{x}).$$

Define

$$\begin{aligned} \boldsymbol{\psi}_k &= (\psi_k(\boldsymbol{x}_1), \dots, \psi_k(\boldsymbol{x}_n))^{\mathsf{T}} \in \mathbb{R}^n, \\ \boldsymbol{D}_{\le \mathsf{m}} &= \mathrm{diag}(\lambda_{d,1}, \lambda_{d,2}, \dots, \lambda_{d,\mathsf{m}}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}, \\ \boldsymbol{\Psi}_{\le \mathsf{m}} &= (\psi_k(\boldsymbol{x}_i))_{i\in[n],k\in[\mathsf{m}]} \in \mathbb{R}^{n\times\mathsf{m}}, \\ \hat{\boldsymbol{f}}_{\le \mathsf{m}} &= (\hat{f}_{d,1}, \hat{f}_{d,2}, \dots, \hat{f}_{d,\mathsf{m}})^{\mathsf{T}} \in \mathbb{R}^{\mathsf{m}}. \end{aligned}$$

We decompose the vectors and matrices $\boldsymbol{f}$, $\boldsymbol{E}$, $\boldsymbol{H}$, and $\boldsymbol{M}$ in terms of orthogonal basis

$$\begin{aligned}
&\boldsymbol{f} = \boldsymbol{f}_{\le \mathsf{m}} + \boldsymbol{f}_{>m}, && \boldsymbol{f}_{\le \mathsf{m}} = \boldsymbol{\Psi}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}, && \boldsymbol{f}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k} \boldsymbol{\psi}_k,\\
&\boldsymbol{E} = \boldsymbol{E}_{\le \mathsf{m}} + \boldsymbol{E}_{>m}, && \boldsymbol{E}_{\le \mathsf{m}} = \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}^2_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}}, && \boldsymbol{E}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda^2_{d,k} \hat{f}_{d,k} \boldsymbol{\psi}_k,\\
&\boldsymbol{H} = \boldsymbol{H}_{\le \mathsf{m}} + \boldsymbol{H}_{>m}, && \boldsymbol{H}_{\le \mathsf{m}} = \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}^2_{\le \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\le \mathsf{m}}, && \boldsymbol{H}_{>\mathsf{m}} \sum_{k=\mathsf{m}+1}^{\infty} \lambda^2_{d,k} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^{\mathsf{T}},\\
&\boldsymbol{M} = \boldsymbol{M}_{\le \mathsf{m}} + \boldsymbol{M}_{>m}, && \boldsymbol{M}_{\le \mathsf{m}} = \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}^4_{\le \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\le \mathsf{m}}, && \boldsymbol{M}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda^4_{d,k} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^{\mathsf{T}}.
\end{aligned} \tag{151}$$

Applying Theorem 7 with respect to the operator $\mathbb{H}_d$ and $\mathbb{H}_d^2$ where the assumptions are satisfied by Assumptions 4.(a), 4.(b), cf. Eqs. (43) and (45), and 5.(a), cf. Eq. (49), and using Assumption 4.(c), the kernel matrices $\boldsymbol{H}$ and $\boldsymbol{M}$ can be rewritten as

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}^2_{\le \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\le \mathsf{m}} + \kappa_H (\mathbf{I} + \boldsymbol{\Delta}_H),\\
\boldsymbol{M} &= \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}^4_{\le \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\le \mathsf{m}} + \kappa_M (\mathbf{I} + \boldsymbol{\Delta}_M),
\end{aligned} \tag{152}$$

where

$$\begin{aligned}
\kappa_H &= \operatorname{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \sum_{k \ge \mathsf{m}+1} \lambda^2_{d,k},\\
\kappa_M &= \operatorname{Tr}(\mathbb{H}^2_{d,>\mathsf{m}}) = \sum_{k \ge \mathsf{m}+1} \lambda^4_{d,k},
\end{aligned}$$

and

$$\max\{\|\boldsymbol{\Delta}_M\|_{\mathrm{op}}, \|\boldsymbol{\Delta}_H\|_{\mathrm{op}}\} = o_{d,\mathbb{P}}(1). \tag{153}$$

Let us introduce the shrinkage matrix

$$\boldsymbol{S}_{\le \mathsf{m}} = \left(\mathbf{I}_{\mathsf{m}} + \frac{\kappa_H + \lambda}{n} \boldsymbol{D}^{-2}_{\le \mathsf{m}}\right)^{-1} = \operatorname{diag}((s_j)_{j \in [\mathsf{m}]}) \qquad \text{where } s_j = \frac{\lambda^2_{d,j}}{\lambda^2_{d,j} + \frac{\kappa_H + \lambda}{n}}. \tag{154}$$

Step 2. Decompose the risk

Recalling $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, we decompose the risk as follows

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

where

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

Step 3. Term T_2

Note we have

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

where

$$\begin{aligned} T_{21} &= \boldsymbol{f}_{\le \mathsf{m}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{f}_{\le \mathsf{m}}, \\ T_{22} &= 2 \boldsymbol{f}_{\le \mathsf{m}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{f}_{> \mathsf{m}}, \\ T_{23} &= \boldsymbol{f}_{> \mathsf{m}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{f}_{> \mathsf{m}}. \end{aligned} \tag{155}$$

By Lemma 12 which is stated in Section C.1.1 below, we have

$$\| n (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} - \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{S}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} / n \|_{\mathrm{op}} = o_{d,\mathbb{P}}(1), \tag{156}$$

hence

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

By Assumption 4.(a), the conditions of Theorem 7.(b) are satisfied, and we have (with $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$)

$$\hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{S}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} / n^2 = \hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} (\mathbf{I} + \boldsymbol{\Delta}) \boldsymbol{S}_{\le \mathsf{m}}^2 (\mathbf{I} + \boldsymbol{\Delta}) \hat{\boldsymbol{f}}_{\le \mathsf{m}} = \| \boldsymbol{S}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2.$$

Moreover,

$$\| \boldsymbol{\Psi}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 / n = \hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} (\mathbf{I} + \boldsymbol{\Delta}) \hat{\boldsymbol{f}}_{\le \mathsf{m}} = \| \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 (1 + o_{d,\mathbb{P}}(1)).$$

As a result, we obtain

$$T_{21} = \| \boldsymbol{S}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 = \| \boldsymbol{S}_{\le \mathsf{m}} \hat{\boldsymbol{f}}_{\le \mathsf{m}} \|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \mathsf{m}} f_* \|_{L^2}^2. \tag{157}$$

By Eq. (156) again, we simplify

$$\begin{aligned} T_{23} &= \Big(\sum_{k \ge \mathsf{m}+1} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} \Big) (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \Big(\sum_{k \ge \mathsf{m}+1} \boldsymbol{\psi}_k \hat{f}_k \Big) \\ &= \Big(\sum_{k \ge \mathsf{m}+1} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} \Big) \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{S}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} \Big(\sum_{k \ge \mathsf{m}+1} \boldsymbol{\psi}_k \hat{f}_k \Big) / n^2 + \Big[\Big\| \sum_{k \ge \mathsf{m}+1} \boldsymbol{\psi}_k \hat{f}_k \Big\|_2^2 / n \Big] \cdot o_{d,\mathbb{P}}(1). \end{aligned}$$

Note that $\boldsymbol{S}_{\le \mathsf{m}} \preceq \mathbf{I}_{\mathsf{m}}$ and we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{k\geq \mathsf{m}+1}\hat{f}_k\boldsymbol{\psi}_k^{\mathsf{T}}\Big)\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\Big(\sum_{k\geq \mathsf{m}+1}\boldsymbol{\psi}_k\hat{f}_k\Big)\Big]/n^2\\
&\leq \mathbb{E}\Big[\Big(\sum_{k\geq \mathsf{m}+1}\hat{f}_k\boldsymbol{\psi}_k^{\mathsf{T}}\Big)\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\Big(\sum_{k\geq \mathsf{m}+1}\boldsymbol{\psi}_k\hat{f}_k\Big)\Big]/n^2\\
&= \sum_{u,v\geq \mathsf{m}+1}\sum_{s=1}^{\mathsf{m}}\sum_{i,j\in[n]}\Big\{\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_j)\psi_v(\boldsymbol{x}_j)\Big]/n^2\Big\}\hat{f}_v\hat{f}_u\\
&= \sum_{u,v\geq \mathsf{m}+1}\sum_{s=1}^{\mathsf{m}}\sum_{i\in[n]}\Big\{\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_v(\boldsymbol{x}_i)\Big]/n^2\Big\}\hat{f}_v\hat{f}_u\\
&= \frac{1}{n}\sum_{s=1}^{\mathsf{m}}\mathbb{E}_{\boldsymbol{x}}\Big[\big(\mathsf{P}_{>\mathsf{m}}f_*(\boldsymbol{x})\big)^2\psi_s(\boldsymbol{x})^2\Big] \leq \frac{1}{n}\sum_{s=1}^{\mathsf{m}}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2\|\psi_s\|_{L^{(4+2\eta)/\eta}}^2\\
&\leq C(\eta)\frac{\mathsf{m}}{n}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2,
\end{aligned}
$$

where the last inequality used the hypercontractivity assumption as in Assumption 4.(a). Moreover

$$
\mathbb{E}\Big[\frac{1}{n}\Big\|\sum_{k\geq \mathsf{m}+1}\boldsymbol{\psi}_k\hat{f}_k\Big\|_2^2\Big] = \sum_{k=\mathsf{m}+1}^{\infty}\hat{f}_k^2 = \|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^2}^2.
$$

Using the last two displays, and the fact that $\mathsf{m}(d)\leq n(d)^{1-\delta}$ by Assumption 5.(b),

$$
T_{23} = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2. \tag{158}
$$

Using Cauchy-Schwarz inequality with term T_{22}, we get

$$
T_{22}\leq 2(T_{21}T_{23})^{1/2} = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\leq \mathsf{m}}f_*\|_{L^2}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}. \tag{159}
$$

As a result, combining Eqs. (157), (158) and (159) leads to

$$
T_2 = \|\boldsymbol{S}_{\leq \mathsf{m}}\hat{\boldsymbol{f}}_{\leq \mathsf{m}}\|_2^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}}f_*\|_{L^{2+\eta}}^2). \tag{160}
$$

Step 4. Term T_1.

We decompose

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

where

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

By Lemma 13 stated in Section C.1.1 below, we have

$$
\|\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2 - \boldsymbol{S}_{\leq \mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),
$$

so that

$$T_{11} = \hat{\boldsymbol{f}}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \hat{\boldsymbol{f}}_{\le \mathsf{m}}$$
$$= \|\boldsymbol{S}_{\le \mathsf{m}}^{1/2} \hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 = \|\boldsymbol{S}_{\le \mathsf{m}}^{1/2} \hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\le \mathsf{m}} f_*\|_2^2. \tag{161}$$

Applying Cauchy-Schwarz inequality to T_{12}, and by the expression $\boldsymbol{M} = \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^4 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} + \kappa_M (\mathbf{I}_M + \boldsymbol{\Delta}_M)$, cf. Eq. (152), we get with high probability

$$\begin{aligned}
|T_{12}| &= \Big| \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \hat{\boldsymbol{f}}_{\le \mathsf{m}} \Big| \\
&\overset{(a)}{\le} \Big\| \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \Big\|_2 \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2 \\
&\overset{(b)}{=} \Big[\Big(\sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} \Big) (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^4 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \Big(\sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k \Big) \Big]^{1/2} \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2 \\
&\overset{(c)}{\le} \Big[\Big(\sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k^{\mathsf{T}} \Big) (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \Big(\sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k \boldsymbol{\psi}_k \Big) \Big]^{1/2} \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2 \\
&\overset{(d)}{=} T_{23}^{1/2} \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2 \overset{(e)}{=} o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2} \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}.
\end{aligned} \tag{162}$$

Here (a) follows by Cauchy-Schwarz; (b) by the definition of norm; (c) because $\boldsymbol{M} \succeq \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^4 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} + \kappa_M (\mathbf{I} + \boldsymbol{\Delta}_M) \succeq \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^4 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}$; (d) by the definition of T_{23} as in Eq. (155); (e) by Eq. (158).

For term T_{13}, we have

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

Note that we have $\mathbb{E}[\|\boldsymbol{f}\|_2^2] = n \|f_*\|_{L^2}^2$. Further by Eq. (152), we have $\|(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\|_{\mathrm{op}} \le 2/(\kappa_H + \lambda)$ with high probability. Finally,

$$\mathbb{E}[\|\boldsymbol{E}_{>\mathsf{m}}\|_2^2] = n \sum_{k=\mathsf{m}+1}^{\infty} \lambda_{d,k}^4 \hat{f}_k^2 \le n \Big[\max_{k \ge \mathsf{m}+1} \lambda_{d,k}^4 \Big] \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2.$$

As a result, we obtain

$$\begin{aligned}
|T_{13}| &\le O_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} \|f_*\|_{L^2} \Big[n^2 \max_{k \ge \mathsf{m}+1} \lambda_{d,k}^4 \Big]^{1/2} / (\kappa_H + \lambda) \\
&= O_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} \|f_*\|_{L^2} \Big[n \max_{k \ge \mathsf{m}+1} \lambda_{d,k}^2 \Big] / \Big(\sum_{k \ge \mathsf{m}+1} \lambda_{d,k}^2 + \lambda \Big) \\
&= o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2} \|f_*\|_{L^2},
\end{aligned} \tag{163}$$

where the last equality used Eq. (49) in Assumption 5.(a) and the fact that $\lambda \in [0, \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})]$. Combining Eqs. (161), (162) and (163) yields

$$T_1 = \|\boldsymbol{S}_{\le \mathsf{m}}^{1/2} \hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^{2+\eta}}^2). \tag{164}$$

Step 5. Terms T_3.

Again, by Lemma 12,

$$\frac{1}{\sigma_\varepsilon^2} \mathbb{E}_{\boldsymbol{\varepsilon}}[T_3] = \mathrm{Tr}((\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}) = \mathrm{Tr}(\boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{S}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} / n^2) + o_{d,\mathbb{P}}(1).$$

Hence, using Proposition 3 and noting that $\boldsymbol{S}_{\le \mathsf{m}} \preceq \mathbf{I}_{\mathsf{m}}$, we obtain

$$\frac{1}{n^2}\mathrm{Tr}(\boldsymbol{\Psi}_{\le \mathsf{m}}\boldsymbol{S}_{\le \mathsf{m}}^2\boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}) \le \frac{1}{n^2}\mathrm{Tr}(\boldsymbol{\Psi}_{\le \mathsf{m}}\boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}) = \frac{1}{n^2}\mathrm{Tr}(\boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}\boldsymbol{\Psi}_{\le \mathsf{m}}) = \frac{1}{n^2}n\mathsf{m}\big(1+o_{d,\mathbb{P}}(1)\big) = o_{d,\mathbb{P}}(1).$$

We conclude

$$T_3 = o_{d,\mathbb{P}}(1)\cdot\sigma_\varepsilon^2. \tag{165}$$

Step 6. Terms T_4.

Note that

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

Notice that $\boldsymbol{M} \succeq \boldsymbol{\Psi}_{\le L}\boldsymbol{D}_{\le L}^4\boldsymbol{\Psi}_{\le L}^{\mathsf{T}}$ for any $L\in\mathbb{N}$, by the decomposition of Eq. (151). Therefore:

$$\begin{aligned}\|\boldsymbol{D}_{\le L}^2\boldsymbol{\Psi}_{\le L}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\boldsymbol{\Psi}_{\le L}\boldsymbol{D}_{\le L}^2\|_{\mathrm{op}} &= \|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{\Psi}_{\le L}\boldsymbol{D}_{\le L}^4\boldsymbol{\Psi}_{\le L}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}}\\ &\le \|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}}.\end{aligned} \tag{166}$$

Further notice that, using Lemma 12 (stated below) followed by Proposition 3, we get

$$\begin{aligned}\|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}} &= \|\boldsymbol{\Psi}_{\le \mathsf{m}}\boldsymbol{S}_{\le \mathsf{m}}^2\boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}/n\|_{\mathrm{op}}/n + o_{d,\mathbb{P}}(1/n)\\ &\le \|\boldsymbol{\Psi}_{\le \mathsf{m}}\boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}}/n\|_{\mathrm{op}}/n + o_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1)\,.\end{aligned} \tag{167}$$

Hence,

$$\begin{aligned}\boldsymbol{E}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\boldsymbol{E} &\overset{(a)}{=} \lim_{L\to\infty}\boldsymbol{E}_{\le L}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\boldsymbol{E}_{\le L}\\ &\overset{(b)}{=} \lim_{L\to\infty}\hat{\boldsymbol{f}}_{\le L}^{\mathsf{T}}[\boldsymbol{D}_{\le L}^2\boldsymbol{\Psi}_{\le L}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\boldsymbol{\Psi}_{\le L}\boldsymbol{D}_{\le L}^2]\hat{\boldsymbol{f}}_{\le L}\\ &\overset{(c)}{\le} \limsup_{L\to\infty}\|\boldsymbol{D}_{\le L}^2\boldsymbol{\Psi}_{\le L}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\boldsymbol{\Psi}_{\le L}\boldsymbol{D}_{\le L}^2\|_{\mathrm{op}}\cdot\lim_{L\to\infty}\|\hat{\boldsymbol{f}}_{\le L}\|_2^2\\ &\overset{(d)}{\le} \|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\mathrm{op}}\cdot\|f_*\|_{L^2}^2\\ &\overset{(e)}{\le} o_{d,\mathbb{P}}(1)\cdot\|f_*\|_{L^2}^2,\end{aligned}$$

where the limits for $L\to\infty$ exist with high probability. In particular, (a) holds with high probability since $\|\boldsymbol{E}_{\le L}^{\mathsf{T}}-\boldsymbol{E}\|_2\to 0$ as $L\to\infty$, and $\|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-2}\|_{\mathrm{op}} \le 1/\lambda_{\min}(\boldsymbol{H})^2 \le (2/\kappa_H)^2$, by the decomposition (152), together with the fact that $\|\boldsymbol{\Delta}_H\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, cf. Eq. (153). Further, (b) is by definition of $\boldsymbol{E}_{\le L}$; (c) by definition of operator norm; (d) by Eq. (166); (e) by Eq. (167).

We thus obtain

$$T_4 = o_{d,\mathbb{P}}(1)\cdot\sigma_\varepsilon\cdot\|f_*\|_{L^2} = o_{d,\mathbb{P}}(1)\cdot(\sigma_\varepsilon^2+\|f_*\|_{L^2}^2). \tag{168}$$

Step 7. Terms T_5.

We decompose T_5 using $\boldsymbol{f} = \boldsymbol{f}_{\le \mathsf{m}} + \boldsymbol{f}_{>\mathsf{m}}$,

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

where

$$T_{51} = \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\le \mathsf{m}},$$
$$T_{52} = \boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{>\mathsf{m}}.$$

First notice that, by Eq. (167),

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

Then by Lemma 12,

$$\begin{aligned}
\frac{1}{\sigma_\varepsilon^2}\mathbb{E}_{\boldsymbol{\varepsilon}}[T_{51}^2] &= \frac{1}{\sigma_\varepsilon^2}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\le \mathsf{m}}\boldsymbol{f}_{\le \mathsf{m}}^{\mathsf{T}}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{\varepsilon}] \\
&= \boldsymbol{f}_{\le \mathsf{m}}^{\mathsf{T}}[(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}]^2\boldsymbol{f}_{\le \mathsf{m}} \\
&\le \|\boldsymbol{M}^{1/2}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-2}\boldsymbol{M}^{1/2}\|_{\rm op}\|\boldsymbol{M}^{1/2}(\boldsymbol{H}+\lambda \mathbf{I}_n)^{-1}\boldsymbol{f}_{\le \mathsf{m}}\|_2^2 \\
&= o_{d,\mathbb{P}}(1)\cdot T_{21} \\
&= o_{d,\mathbb{P}}(1)\cdot \|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2}^2,
\end{aligned}$$

where the last equality follows by Eq. (157). Similarly, we get

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[T_{52}^2]/\sigma_\varepsilon^2 = o_{d,\mathbb{P}}(1)\cdot T_{23} = o_{d,\mathbb{P}}(1)\cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2.$$

By Markov's inequality, we deduce that

$$T_5 = o_{d,\mathbb{P}}(1)\cdot \sigma_\varepsilon(\|\mathsf{P}_{\le \mathsf{m}} f_*\|_{L^2} + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}) = o_{d,\mathbb{P}}(1)\cdot(\sigma_\varepsilon^2 + \|f_*\|_{L^2}^2). \tag{169}$$

Step 8. Finish the proof.

Combining Eqs. (160), (164), (165), (168) and (169), we have

$$\begin{aligned}
R_{\rm KR}(f_*, \boldsymbol{X}, \lambda) &= \|f_*\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5 \\
&= \|\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 - 2\|\boldsymbol{S}_{\le \mathsf{m}}^{1/2}\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + \|\boldsymbol{S}_{\le \mathsf{m}}\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2 \\
&\quad + o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2) \\
&= \|(\mathbf{I}-\boldsymbol{S}_{\le \mathsf{m}})\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).
\end{aligned}$$

Recall the expression (55) of $\hat{f}_{\gamma^{\rm eff}}^{\rm eff}$:

$$\hat{f}_{\gamma^{\rm eff}}^{\rm eff} = \sum_{j=1}^{\infty} \frac{\lambda_{d,j}^2}{\lambda_{d,j}^2 + \frac{\gamma^{\rm eff}}{n}}\,\hat{f}_{d,k}\psi_{d,j}\,,$$

with $\gamma^{\rm eff} = \lambda + \kappa_H$. From Assumption 5.(*a*), we have $\max_{j>\mathsf{m}} \lambda_{d,j}^2 = o_d(1)\cdot \kappa_H/n$ and we deduce

$$\|(\mathbf{I}-\boldsymbol{S}_{\le \mathsf{m}})\hat{\boldsymbol{f}}_{\le \mathsf{m}}\|_2^2 + \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^2}^2 = \|f_* - \hat{f}_{\gamma^{\rm eff}}^{\rm eff}\|_{L^2}^2 + \|f_*\|_{L^2}^2\cdot o_{d,\mathbb{P}}(1).$$

We conclude

$$R_{\rm KR}(f_*, \boldsymbol{X}, \lambda) = \|f_* - \hat{f}_{\gamma^{\rm eff}}^{\rm eff}\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

Proceeding analogously (with $\hat{f}^{\rm eff}_{\gamma^{\rm eff}}$ replacing f_*) we obtain

$$\|\hat{f}_\lambda - \hat{f}^{\rm eff}_{\gamma^{\rm eff}}\|_{L^2}^2 = o_{d,\mathbb{P}}(1)\cdot(\|f_*\|_{L^2}^2 + \|\mathsf{P}_{>\mathsf{M}} f_*\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

C.1.1. Auxiliary lemmas

Lemma 12. *Follow the assumptions and notations in the proof of Theorem 5. We have*

$$\|n(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1} - \boldsymbol{\Psi}_{\le\mathsf{m}}\boldsymbol{S}^2_{\le\mathsf{m}}\boldsymbol{\Psi}^{\mathsf{T}}_{\le\mathsf{m}}/n\|_{\rm op} = o_{d,\mathbb{P}}(1),$$

where $\boldsymbol{S}_{\le\mathsf{m}}$ is the shrinkage matrix defined in Equation (154).

Proof of Lemma 12. We simplify the notations by defining $\boldsymbol{\psi}_k = (\psi_k(\boldsymbol{x}_i))_{i\in[n]} \in \mathbb{R}^n$ $\boldsymbol{\Psi} = \boldsymbol{\psi}_{\le\mathsf{m}} \in \mathbb{R}^{n\times\mathsf{m}}$, $\boldsymbol{D} = \boldsymbol{D}_{\le\mathsf{m}} = \mathrm{diag}(\lambda_{d,1},\dots,\lambda_{d,\mathsf{m}}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}$.

Then recalling Eq. (152), we have

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^{\mathsf{T}} + \kappa_H(\mathbf{I}+\boldsymbol{\Delta}_H),\\ \boldsymbol{M} &= \boldsymbol{\Psi}\boldsymbol{D}^4\boldsymbol{\Psi}^{\mathsf{T}} + \kappa_M(\mathbf{I}+\boldsymbol{\Delta}_M),\end{aligned} \tag{170}$$

where $\kappa_H = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})$ and $\kappa_M = \mathrm{Tr}(\mathbb{H}^2_{d,>\mathsf{m}})$, and

$$\max\{\|\boldsymbol{\Delta}_M\|_{\rm op}, \|\boldsymbol{\Delta}_H\|_{\rm op}\} = o_{d,\mathbb{P}}(1). \tag{171}$$

As a result, we obtain the decomposition

$$n(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1} = T_1 + T_2,$$

where

$$\begin{aligned}T_1 &= n\kappa_M(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}(\mathbf{I}_M+\boldsymbol{\Delta}_M)(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1},\\ T_2 &= n(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{\Psi}\boldsymbol{D}^4\boldsymbol{\Psi}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}.\end{aligned}$$

Step 1. Bound term T_1.

For T_1, by Eqs. (170) and (171),

$$\|T_1\|_{\rm op} \le n\kappa_M\|(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\|_{\rm op}^2\|\mathbf{I}+\boldsymbol{\Delta}_M\|_{\rm op} = O_{d,\mathbb{P}}(1)\cdot n[\kappa_M/\kappa_H^2]. \tag{172}$$

By Eq. (49) in Assumption 5.(a), we have

$$\frac{\kappa_M}{\kappa_H^2} = \frac{\mathrm{Tr}(\mathbb{H}^2_{d,>\mathsf{m}})}{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})^2} \le \frac{\|\mathbb{H}_{d,>\mathsf{m}}\|_{\rm op}}{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})} = O_d(n^{-1-\delta_0}).$$

We conclude that

$$\|T_1\|_{\rm op} = o_{d,\mathbb{P}}(1).$$

Step 2. Bound term T_2.

For T_2, setting $\boldsymbol{\Delta}'_H = \kappa_H\boldsymbol{\Delta}_H/(\lambda+\kappa_H)$, the Sherman-Morrison-Woodbury formula produces the identity

$$\begin{aligned}T_2 =& n(\mathbf{I}+\boldsymbol{\Delta}'_H)^{-1}\boldsymbol{\Psi}((\kappa_H+\lambda)\boldsymbol{D}^{-2} + \boldsymbol{\Psi}^{\mathsf{T}}(\mathbf{I}+\boldsymbol{\Delta}'_H)^{-1}\boldsymbol{\Psi})^{-2}\boldsymbol{\Psi}^{\mathsf{T}}(\mathbf{I}+\boldsymbol{\Delta}'_H)^{-1}\\ =& \boldsymbol{E}\boldsymbol{\Psi}\boldsymbol{R}^2\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{E}/n,\end{aligned}$$

where

$$\boldsymbol{E} = (\mathbf{I} + \boldsymbol{\Delta}'_H)^{-1},$$
$$\boldsymbol{R} = [(\kappa_H + \lambda)(n\boldsymbol{D}^2)^{-1} + \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}\boldsymbol{\Psi}/n]^{-1}.$$

Denote $\boldsymbol{S} := \boldsymbol{S}_{\leq \mathsf{m}} = [\mathbf{I}_\mathsf{m} + (\kappa_H + \lambda)(n\boldsymbol{D}^2)^{-1}]^{-1}$. We get

$$\begin{aligned}\|T_2 - \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{S}^2\boldsymbol{\Psi}/n\|_{\mathrm{op}} &\leq (1 + \|\boldsymbol{E}\|_{\mathrm{op}})\|\boldsymbol{E} - \mathbf{I}\|_{\mathrm{op}}\|\boldsymbol{\Psi}\boldsymbol{R}^2\boldsymbol{\Psi}^\mathsf{T}/n\| + \|\boldsymbol{\Psi}\boldsymbol{R}^{-2}\boldsymbol{\Psi}^\mathsf{T}/n - \boldsymbol{\Psi}\boldsymbol{S}^2\boldsymbol{\Psi}^\mathsf{T}/n\|_{\mathrm{op}} \\ &\leq (1 + \|\boldsymbol{E}\|_{\mathrm{op}})\|\boldsymbol{E} - \mathbf{I}\|_{\mathrm{op}}\|\boldsymbol{\Psi}\boldsymbol{R}^2\boldsymbol{\Psi}^\mathsf{T}/n\|_{\mathrm{op}} + \|\boldsymbol{\Psi}\boldsymbol{\Psi}^\mathsf{T}/n\|_{\mathrm{op}}\|\boldsymbol{R}^2 - \boldsymbol{S}^2\|_{\mathrm{op}}.\end{aligned}$$

Recalling Eq. (171), we have $\|\boldsymbol{E} - \mathbf{I}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and by Theorem 7.(b), we have $\|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}/n - \mathbf{I}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Furthermore

$$\|\boldsymbol{R}^2 - \boldsymbol{S}^2\|_{\mathrm{op}} = \|\boldsymbol{R} - \boldsymbol{S}\|_{\mathrm{op}}(\|\boldsymbol{R}\|_{\mathrm{op}} + \|\boldsymbol{S}\|_{\mathrm{op}}0 \leq \|\boldsymbol{R}\|_{\mathrm{op}}\|\boldsymbol{S}\|_{\mathrm{op}}(\|\boldsymbol{R}\|_{\mathrm{op}} + \|\boldsymbol{S}\|_{\mathrm{op}})\|\boldsymbol{R}^{-1} - \boldsymbol{S}^{-1}\|_{\mathrm{op}}$$

and

$$\begin{aligned}\|\boldsymbol{R}^{-1} - \boldsymbol{S}^{-1}\|_{\mathrm{op}} &\leq \|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}\boldsymbol{\Psi}/n - \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}/n\|_{\mathrm{op}} + \|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}/n - \mathbf{I}\|_{\mathrm{op}} \\ &\leq \|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}/n\|_{\mathrm{op}}\|\boldsymbol{E} - \mathbf{I}\|_{\mathrm{op}} + \|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}/n - \mathbf{I}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).\end{aligned}$$

Using that $\|\boldsymbol{S}\|_{\mathrm{op}} \leq 1$, we obtain by the above computation that $\|\boldsymbol{R}\|_{\mathrm{op}} \leq 1 + o_{d,\mathbb{P}}(1)$. Combining the above inequalities yields

$$\|T_2 - \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{S}^2\boldsymbol{\Psi}/n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

This gives

$$\|n(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{M}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1} - \boldsymbol{\Psi}\boldsymbol{S}^2\boldsymbol{\Psi}^\mathsf{T}/n\|_{\mathrm{op}} \leq \|T_1\|_{\mathrm{op}} + \|T_2 - \boldsymbol{\Psi}\boldsymbol{S}^2\boldsymbol{\Psi}^\mathsf{T}/n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)\,,$$

which completes the proof. □

Lemma 13. *Follow the assumptions and notations in the proof of Theorem 5. We have*

$$\|\boldsymbol{S}_{\leq \mathsf{m}} - \boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

where $\boldsymbol{S}_{\leq \mathsf{m}}$ is the shrinkage matrix defined in Equation (154).

Proof of Lemma 13. We will follow the notations in the proof of Proposition 12. Applying Theorem 7 with respect to operator $\mathbb{H}_d$ and by Eq. (174) in Assumption 4.(c1),

$$\boldsymbol{H} + \lambda\mathbf{I}_n = \boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^\mathsf{T} + \kappa_H(\mathbf{I}_n + \boldsymbol{\Delta}_H) + \lambda\mathbf{I}_n = \boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^\mathsf{T} + (\kappa_H + \lambda)(\mathbf{I}_n + \boldsymbol{\Delta}'_H),$$

where $\|\boldsymbol{\Delta}_H\|_{\mathrm{op}}, \|\boldsymbol{\Delta}'_H\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. The Sherman-Morrison-Woodbury formula gives the identity

$$\boldsymbol{\Psi}^\mathsf{T}[\boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^\mathsf{T} + (\kappa_H + \lambda)(\mathbf{I}_n + \boldsymbol{\Delta}'_H)]^{-1}\boldsymbol{\Psi}\boldsymbol{D}^2 = \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{\Psi}\boldsymbol{R}/n,$$

where $\boldsymbol{E} = \mathbf{I}_n + \boldsymbol{\Delta}_H$ and $\boldsymbol{R} = [(\kappa_H + \lambda)(n\boldsymbol{D}^2)^{-1} + \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{\Psi}/n]^{-1}$. We obtain the inequality

$$\|\boldsymbol{S} - \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{\Psi}\boldsymbol{R}/n\|_{\mathrm{op}} \leq \|\boldsymbol{R}\|_{\mathrm{op}}\|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{\Psi}/n - \mathbf{I}\|_{\mathrm{op}} + \|\boldsymbol{R} - \boldsymbol{S}\|_{\mathrm{op}}.$$

In the proof of Lemma 12, we already showed that $\|\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{\Psi}/n - \mathbf{I}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, $\|\boldsymbol{R} - \boldsymbol{S}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\|\boldsymbol{R}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. This concludes the proof. □

C.2. Kernel ridge regression under relaxed assumptions on the diagonal

In this section, we state and prove a version of Theorem 5 that holds under weaker assumptions. Namely, instead of the concentration bound in Assumption 4.(c) we only require that the diagonal terms are upper bounded by a sub-polynomial factor times their expectation. Instead, we assume a spectral gap condition that was not required in the previous section.

We will first describe the modified assumption, then state the new version of the theorem. The proof is very similar to the one in the previous section. We will therefore use the same notations and only sketch the differences.

Assumption 8 *(Relaxed kernel concentration at level $\{(n(d), \mathsf{m}(d))\}_{d\ge 1}$). We assume the kernel concentration property at level $\{(n(d), \mathsf{m}(d))\}_{d\ge 1}$, as stated in Assumption 4, with condition (c) replaced by the following*

(c') (Upper bound on the diagonal elements of the kernel) For $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$ and any $\delta > 0$, we have

$$\max_{i\in[n(d)]} \mathbb{E}_{\boldsymbol{x}\sim u(d)}\big[H_{d,>u(d)}(\boldsymbol{x}_i,\boldsymbol{x})^2\big] = O_{d,\mathbb{P}}(n(d)^{\delta})\cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}\big[H_{d,>u(d)}(\boldsymbol{x},\boldsymbol{x}')^2\big], \tag{173}$$

$$\max_{i\in[n(d)]} H_{d,>u(d)}(\boldsymbol{x}_i,\boldsymbol{x}_i) = O_{d,\mathbb{P}}(n(d)^{\delta})\cdot \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,>u(d)}(\boldsymbol{x},\boldsymbol{x})]. \tag{174}$$

Assumption 9 *(Eigenvalue condition at level $\{(n(d), \mathsf{m}(d))\}_{d\ge 1}$). We assume the eigenvalue condition Assumption 5 and, in addition, the following to hold*

(c) There exists a fixed $\delta_0 > 0$, such that

$$n^{1-\delta_0} \ge \frac{1}{\lambda^2_{d,\mathsf{m}(d)}} \sum_{k=\mathsf{m}(d)+1} \lambda^2_{d,k}\,.$$

Assumption 10 *(Regularization and lower bound on diagonal elements). Consider the regularization parameter $\lambda \in \mathbb{R}_{\ge 0}$. We assume that one of the following holds:*

(i) For $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$ and any $\delta > 0$, we have

$$\min_{i\in[n(d)]} \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x})^2] = \Omega_{d,\mathbb{P}}(n(d)^{-\delta})\cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x}')^2], \tag{175}$$

$$\min_{i\in[n(d)]} H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x}_i) = \Omega_{d,\mathbb{P}}(n(d)^{-\delta})\cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x})], \tag{176}$$

and $\lambda = O_d(1)\cdot \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}(d)})$ (in particular, taking $\lambda = 0$ is fine).

(ii) We have $\lambda = \Theta_d(1)\cdot \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}(d)})$.

Theorem 9. *Let $\{f_* \in \mathcal{D}_d\}_{d\ge 1}$ be a sequence of functions, $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim \nu_d$ independently, $\{\mathbb{H}_d\}_{d\ge 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, n(d), \mathsf{m}(d))\}_{d\ge 1}$ and the regularization parameter λ satisfy Assumptions 8, 9, and 10. Then for any fixed $\eta > 0$, we have*

$$|R_{\mathrm{KR}}(f_*, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \|\mathsf{P}_{>\mathsf{m}} f_*\|^2_{L^2}| = o_{d,\mathbb{P}}(1)\cdot(\|f_*\|^2_{L^2} + \|\mathsf{P}_{>\mathsf{m}} f_*\|^2_{L^{2+\eta}} + \sigma_\varepsilon^2). \tag{177}$$

C.2.1. Proof outline for Theorem 9

Throughout this section, we will denote $\delta_0 > 0$ a fixed constant and $\delta > 0$ a constant that can be arbitrarily small. The value of δ_0 is allowed to change from line to line.

By the spectral gap condition (Assumption 9), the population estimator $\hat{f}^{\rm eff}_{\gamma^{\rm eff}}$ defined in Eq. (55) is approximately given by

$$\|\hat{f}^{\rm eff}_{\gamma^{\rm eff}} - \mathsf{P}_{\leq \mathsf{m}} f_*\|_{L^2}^2 = o_{d,\mathbb{P}}(1) \cdot \|f_*\|_{L^2}^2.$$

Similarly, the shrinkage matrix defined in Eq. (154) verifies

$$\|\boldsymbol{S}_{\leq \mathsf{m}} - \mathbf{I}_{\mathsf{m}}\|_{\rm op} = O_{d,\mathbb{P}}(n^{-\delta_0}).$$

From Theorem 7 applied to the operator $\mathbb{H}_d$ and $\mathbb{H}_d^2$, the kernel matrices $\boldsymbol{H}$ and $\boldsymbol{M}$ can be rewritten as

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{\Psi}_{\leq \mathsf{m}} \boldsymbol{D}^2_{\leq \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\leq \mathsf{m}} + \kappa_H(\boldsymbol{\Lambda}_H + \boldsymbol{\Delta}_H), \\ \boldsymbol{M} &= \boldsymbol{\Psi}_{\leq \mathsf{m}} \boldsymbol{D}^4_{\leq \mathsf{m}} \boldsymbol{\Psi}^{\mathsf{T}}_{\leq \mathsf{m}} + \kappa_M(\boldsymbol{\Lambda}_M + \boldsymbol{\Delta}_M), \end{aligned} \tag{178}$$

where

$$\begin{aligned} \kappa_H &= \operatorname{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \sum_{k \geq \mathsf{m}+1} \lambda_{d,k}^2, \\ \kappa_M &= \operatorname{Tr}(\mathbb{H}^2_{d,>\mathsf{m}}) = \sum_{k \geq \mathsf{m}+1} \lambda_{d,k}^4, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\Lambda}_H =& \operatorname{diag}\left(\left\{H_{d,>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x}_i)/\kappa_H\right\}_{i\in[n]}\right), \\ \boldsymbol{\Lambda}_M =& \operatorname{diag}\left(\left\{\mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x})^2]/\kappa_M\right\}_{i\in[n]}\right); \end{aligned}$$

and there exists a fixed $\delta_0 > 0$ such that

$$\max\{\|\boldsymbol{\Delta}_M\|_{\rm op}, \|\boldsymbol{\Delta}_H\|_{\rm op}\} = O_{d,\mathbb{P}}(n^{-\delta_0}). \tag{179}$$

From Lemma 7 applied to $\boldsymbol{\Lambda}_H$ and $\boldsymbol{\Lambda}_M$ with Assumptions 4.(a) and 8.(c'), we have

$$\begin{aligned} \boldsymbol{\Lambda}_H &\preceq O_{d,\mathbb{P}}(n^{\delta}) \cdot \mathbf{I}_n, \\ \boldsymbol{\Lambda}_M &\preceq O_{d,\mathbb{P}}(n^{\delta}) \cdot \mathbf{I}_n. \end{aligned} \tag{180}$$

Furthermore from Assumption 10 and Eq. (179), we have for any $\delta < \delta'$,

$$\boldsymbol{H} + \lambda \mathbf{I}_n \succeq \kappa_H(\boldsymbol{\Lambda}_H + \boldsymbol{\Delta}_H) + \lambda \mathbf{I}_n \succeq \Omega_{d,\mathbb{P}}(n^{-\delta}) \cdot \kappa_H \cdot \mathbf{I}_n. \tag{181}$$

The handling of the bounds on T_1, T_2, T_3, T_4 and T_5 follows from the same computation as Section C.1, where every $o_{d,\mathbb{P}}(1)$ is replaced by $O_{d,\mathbb{P}}(n^{-\delta_0})$ for some fixed $\delta_0 > 0$ while every $O_{d,\mathbb{P}}(1)$ is replaced by $O_{d,\mathbb{P}}(n^{\delta})$ with $\delta > 0$ arbitrary small. In particular, bounds of the form $O_{d,\mathbb{P}}(1) \cdot o_{d,\mathbb{P}}(1)$ should be replaced by $O_{d,\mathbb{P}}(n^{\delta}) \cdot O_{d,\mathbb{P}}(n^{-\delta_0})$, and taking $\delta > 0$ sufficiently small yields a bound $o_{d,\mathbb{P}}(1)$ (see the proofs bellow for some examples).

Below we detail the proof of the updated auxiliary lemmas from Section C.1.1. Eq. (182) is used to bound the term T_{21}, Eq. (183) is used to bound the term T_{23}, while Eq. (184) is used to bound the term T_3, T_4 and T_5.

Lemma 14. *Follow the assumptions of Theorem 9 and the same notations as in Section C.1. Define*

$$\boldsymbol{G} = n(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{M} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}.$$

Then, there exists a fixed $\delta_0 > 0$ such that for any $\delta > 0$,

$$\|\boldsymbol{\psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{\psi}_{\le \mathsf{m}}/n - \mathbf{I}_{\mathsf{m}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n^{-\delta_0}), \tag{182}$$

$$\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{f}_{>\mathsf{m}}/n = O_{d,\mathbb{P}}(n^{-\delta_0}) \cdot \|\mathsf{P}_{>\mathsf{m}} f_*\|_{L^{2+\eta}}^2, \tag{183}$$

$$\|\boldsymbol{G}\|_{\mathrm{op}} \le O_{d,\mathbb{P}}(n^{\delta}). \tag{184}$$

Proof of Lemma 14. Recall that we denote $\delta_0 > 0$ a fixed constant and $\delta > 0$ a constant that can be arbitrarily small. The value of δ_0 is allowed to change from line to line.

Following the notations as in the proof of Lemma 12, we have

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{\Psi} \boldsymbol{D}^2 \boldsymbol{\Psi}^{\mathsf{T}} + \kappa_H (\boldsymbol{\Lambda}_H + \boldsymbol{\Delta}_H), \\
\boldsymbol{M} &= \boldsymbol{\Psi} \boldsymbol{D}^4 \boldsymbol{\Psi}^{\mathsf{T}} + \kappa_M (\boldsymbol{\Lambda}_M + \boldsymbol{\Delta}_M).
\end{aligned}$$

Consider the same decomposition $\boldsymbol{G} = \boldsymbol{T}_1 + \boldsymbol{T}_2$ as in the proof of Lemma 12, where

$$\begin{aligned}
\boldsymbol{T}_1 &= n\kappa_M (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} (\boldsymbol{\Lambda}_M + \boldsymbol{\Delta}_M) (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}, \\
\boldsymbol{T}_2 &= n(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Psi} \boldsymbol{D}^4 \boldsymbol{\Psi}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}.
\end{aligned}$$

Step 1. Bound term $\boldsymbol{T}_1$.

For $\boldsymbol{T}_1$, by Eqs. (180) and (181), we have for any $\delta > 0$,

$$\begin{aligned}
\|\boldsymbol{T}_1\|_{\mathrm{op}} &\le n\kappa_M \|(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1}\|_{\mathrm{op}}^2 \|(\boldsymbol{\Lambda}_M + \boldsymbol{\Delta}_M)\|_{\mathrm{op}} \\
&\le O_{d,\mathbb{P}}(1) \cdot n\kappa_M \cdot n^{2\delta} \kappa_H^{-2} \cdot n^{2\delta} \\
&\le O_{d,\mathbb{P}}(1) \cdot n^{1+4\delta} \frac{\kappa_M}{\kappa_H^2}.
\end{aligned} \tag{185}$$

By Eq. (49) in Assumption 5.(a),

$$\frac{\kappa_M}{\kappa_H^2} = \frac{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}^2)}{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})^2} \le \frac{\|\mathbb{H}_{d,>\mathsf{m}}\|_{\mathrm{op}}}{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})} = O_d(n^{-1-\delta_0}).$$

Hence, taking δ sufficiently small in Eq. (172) yields

$$\|\boldsymbol{T}_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n^{-\delta_0}).$$

Step 2. Simplifying the term $\boldsymbol{T}_2$.

Introduce $\boldsymbol{A} = \boldsymbol{\Lambda}_H + \boldsymbol{\Delta} + (\lambda/\kappa_H) \cdot \mathbf{I}_n$ so that

$$\boldsymbol{H} + \lambda \mathbf{I}_n = \boldsymbol{\Psi} \boldsymbol{D}^2 \boldsymbol{\Psi}^{\mathsf{T}} + \kappa_H \boldsymbol{A}.$$

The Sherman-Morrison-Woodbury formula gives the identity

$$\boldsymbol{T}_2 = \boldsymbol{A}^{-1} \boldsymbol{\psi} \big(\kappa_H (n\boldsymbol{D}^2)^{-1} + \boldsymbol{\psi}^{\mathsf{T}} \boldsymbol{A}^{-1} \boldsymbol{\psi}/n \big)^{-2} \boldsymbol{\psi}^{\mathsf{T}} \boldsymbol{A}^{-1}/n.$$

From Assumption 9.(*b*), we have $\kappa_H(n\boldsymbol{D}^2)^{-1} \preceq \Omega_{d,\mathbb{P}}(n^{-\delta_0})\cdot \mathbf{I}_n$. Furthermore, recalling that $\|\boldsymbol{\psi}^\mathsf{T}\boldsymbol{\psi}-\mathbf{I}_\mathsf{m}\|_{\rm op} = o_{d,\mathbb{P}}(1)$ and $\boldsymbol{A}^{-1} \succeq O_{d,\mathbb{P}}(n^{-\delta})\mathbf{I}_n$ for any $\delta > 0$, we deduce (for example from Lemma 8)

$$\big\|\boldsymbol{T}_2 - \boldsymbol{A}^{-1}\boldsymbol{\psi}\big(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{\psi}/n\big)^{-2}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{A}^{-1}/n\big\|_{\rm op} = O_{d,\mathbb{P}}(n^{-\delta_0}).$$

Denote $\boldsymbol{S} = \boldsymbol{\Lambda}_H + (\lambda/\kappa_H)\cdot \mathbf{I}_n$ the diagonal matrix such that we have $\|\boldsymbol{A}-\boldsymbol{S}\|_{\rm op} = O_{d,\mathbb{P}}(n^{-\delta_0})$. We have $\Omega_d(n^{-\delta})\cdot \mathbf{I}_n \preceq \boldsymbol{S} \preceq O_d(n^{\delta})\cdot \mathbf{I}_n$. Similarly to the previous line, we get

$$\big\|\boldsymbol{A}^{-1}\boldsymbol{\psi}\big(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{\psi}/n\big)^{-2}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{A}^{-1}/n - \boldsymbol{S}^{-1}\boldsymbol{\psi}\big(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{\psi}/n\big)^{-2}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}/n\big\|_{\rm op} = O_{d,\mathbb{P}}(n^{-\delta_0}).$$

Denote $\boldsymbol{R} = \boldsymbol{S}^{-1}\boldsymbol{\psi}\big(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{\psi}/n\big)^{-2}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}/n$.

Step 3. Proving the bounds.

First notice that because $\Omega_d(n^{-\delta})\cdot \mathbf{I}_n \preceq \boldsymbol{S} \preceq O_d(n^{\delta})\cdot \mathbf{I}_n$ and $\|\boldsymbol{\psi}^\mathsf{T}\boldsymbol{\psi} - \mathbf{I}_\mathsf{m}\|_{\rm op} = o_{d,\mathbb{P}}(1)$,

$$\|\boldsymbol{G}\|_{\rm op} \le \|\boldsymbol{T}_1\|_{\rm op} + \|\boldsymbol{T}_2 - \boldsymbol{R}\|_{\rm op} + \|\boldsymbol{R}\|_{\rm op} = O_{d,\mathbb{P}}(n^{\delta}),$$

for any $\delta > 0$, which proves Eq. (184). Similarly,

$$\|\boldsymbol{\psi}^\mathsf{T}\boldsymbol{G}\boldsymbol{\psi}/n - \mathbf{I}_\mathsf{m}\|_{\rm op} \le (\|\boldsymbol{T}_1\|_{\rm op} + \|\boldsymbol{T}_2 - \boldsymbol{R}\|_{\rm op})\|\boldsymbol{\psi}/\sqrt{n}\|_{\rm op}^2 + \|\boldsymbol{\psi}^\mathsf{T}\boldsymbol{R}\boldsymbol{\psi}/n - \mathbf{I}_\mathsf{m}\|_{\rm op} = O_{d,\mathbb{P}}(n^{-\delta_0}),$$

which proves Eq. (182).

Notice that

$$\boldsymbol{R} = \boldsymbol{S}^{-1}\boldsymbol{\psi}\big(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{\psi}/n\big)^{-2}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}/n \preceq O_{d,\mathbb{P}}(n^{\delta})\cdot .\boldsymbol{S}^{-1}\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}/n$$

Denote $\boldsymbol{S} = \mathrm{diag}((s_i)_{i\in[n]})$ and recall the decomposition

$$\boldsymbol{f}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_k\boldsymbol{\psi}_k.$$

We have

$$\begin{aligned}
\mathbb{E}\Big[\boldsymbol{f}_{>\mathsf{m}}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{f}_{>\mathsf{m}}\Big]/n^2 &= \mathbb{E}\Big[\Big(\sum_{k\ge\mathsf{m}+1}\hat{f}_k\boldsymbol{\psi}_k^\mathsf{T}\Big)\boldsymbol{S}^{-1}\boldsymbol{\Psi}_{\le\mathsf{m}}\boldsymbol{\Psi}_{\le\mathsf{m}}^\mathsf{T}\boldsymbol{S}^{-1}\Big(\sum_{k\ge\mathsf{m}+1}\boldsymbol{\psi}_k\hat{f}_k\Big)\Big]/n^2\\
&= \sum_{u,v\ge\mathsf{m}+1}\sum_{t=1}^{\mathsf{m}}\sum_{i,j\in[n]}\Big\{s_i^{-1}s_j^{-1}\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_t(\boldsymbol{x}_i)\psi_t(\boldsymbol{x}_j)\psi_v(\boldsymbol{x}_j)\Big]/n^2\Big\}\hat{f}_v\hat{f}_u\\
&= \sum_{u,v\ge\mathsf{m}+1}\sum_{t=1}^{\mathsf{m}}\sum_{i\in[n]}s_i^{-2}\Big\{\mathbb{E}\Big[\psi_u(\boldsymbol{x}_i)\psi_t(\boldsymbol{x}_i)\psi_t(\boldsymbol{x}_i)\psi_v(\boldsymbol{x}_i)\Big]/n^2\Big\}\hat{f}_v\hat{f}_u\\
&= O_d(n^{\delta})\cdot\frac{1}{n}\sum_{s=1}^{\mathsf{m}}\mathbb{E}_{\boldsymbol{x}}\Big[\big(\mathsf{P}_{>\mathsf{m}}f_*(\boldsymbol{x})\big)^2\psi_s(\boldsymbol{x})^2\Big]\\
&\le O_d(n^{\delta})\cdot\frac{1}{n}\sum_{t=1}^{\mathsf{m}}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2\|\psi_t\|_{L^{(4+2\eta)/\eta}}^2\\
&\le O_d(n^{\delta})\cdot\frac{\mathsf{m}}{n}\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+eta}}^2 = O_d(n^{-\delta_0})\cdot\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2.
\end{aligned}$$

By Markov's inequality, we get

$$|\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{R}\boldsymbol{f}_{>\mathsf{m}}/n| \le |\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{\psi}\boldsymbol{\psi}^{\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{f}_{>\mathsf{m}}/n| = O_d(n^{-\delta_0})\cdot\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2.$$

We deduce that

$$|\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{>\mathsf{m}}| \le (\|\boldsymbol{T}_1\|_{\mathrm{op}} + \|\boldsymbol{T}_2-\boldsymbol{R}\|_{\mathrm{op}})\|\boldsymbol{f}_{>\mathsf{m}}/\sqrt{n}\|_2^2 + |\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{R}\boldsymbol{f}_{>\mathsf{m}}/n| = O_d(n^{-\delta_0})\cdot\|\mathsf{P}_{>\mathsf{m}}f_*\|_{L^{2+\eta}}^2,$$

which concludes the proof. □

Lemma 15. *Follow the assumptions of Theorem 9 and the same notations as in Section C.1. There exists a fixed $\delta_0 > 0$ such that*

$$\|\mathbf{I}_{\le\mathsf{m}} - \boldsymbol{\Psi}_{\le\mathsf{m}}^{\mathsf{T}}(\boldsymbol{H}+\lambda\mathbf{I}_n)^{-1}\boldsymbol{\Psi}_{\le\mathsf{m}}\boldsymbol{D}_{\le\mathsf{m}}^2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n^{-\delta_0}).$$

Proof of Lemma 15. We follow the same argument as in Lemma 13. We decompose

$$\boldsymbol{H}+\lambda\mathbf{I}_n = \boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^{\mathsf{T}} + \kappa_H\cdot\boldsymbol{A},$$

where we denoted $\boldsymbol{A} = \boldsymbol{\Lambda}_H + \boldsymbol{\Delta} + (\lambda/\kappa_H)\cdot\mathbf{I}$. By the Sherman-Morrison-Woodbury formula, we have the identity

$$\boldsymbol{\Psi}^{\mathsf{T}}[\boldsymbol{\Psi}\boldsymbol{D}^2\boldsymbol{\Psi}^{\mathsf{T}}+\boldsymbol{A}]^{-1}\boldsymbol{\Psi}\boldsymbol{D}^2 = \boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}[\kappa_H(n\boldsymbol{D}^2)^{-1}+\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}/n]^{-1}/n.$$

Hence

$$\|\mathbf{I}_{\mathsf{m}} - \boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}[\kappa_H(n\boldsymbol{D}^2)^{-1}+\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}/n]^{-1}/n\|_{\mathrm{op}} = \|\kappa_H(n\boldsymbol{D}^2)^{-1}(\kappa_H(n\boldsymbol{D}^2)^{-1}+\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}/n)^{-1}\|_{\mathrm{op}}.$$

We know that by Assumption 9, $n\boldsymbol{D}^2 \succeq \Omega_d(n^{\delta_0})\cdot\kappa_H\cdot\mathbf{I}_{\mathsf{m}}$. Furthermore, by Eq. (181), we have $\boldsymbol{A}^{-1} \succeq \Omega_d(n^{-\delta})\cdot\kappa_H\cdot\mathbf{I}_n$. Using that $\|\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}-\mathbf{I}_{\mathsf{m}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, we deduce that for any $\delta>0$,

$$\|(\kappa_H(n\boldsymbol{D}^2)^{-1}+\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}/n)^{-1}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n^{\delta}).$$

We obtain the bound

$$\|\mathbf{I}_{\mathsf{m}} - \boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}[\kappa_H(n\boldsymbol{D}^2)^{-1}+\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\Psi}/n]^{-1}/n\|_{\mathrm{op}} = O_{d,\mathbb{P}}(n^{-\delta_0})\cdot O_{d,\mathbb{P}}(n^{\delta}).$$

Taking δ sufficiently small concludes the proof. □

Appendix D. Proof of Theorem 2: generalization error of RFRR on the sphere and hypercube

We check that Assumption 3 implies the assumptions of Theorem 1 on the sphere (Section D.1) and on the hypercube (Section D.2).

D.1. On the sphere

Proof of Theorem 2 on the sphere. Consider the spherical case $\boldsymbol{\theta}, \boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $d^{\mathsf{s}+\delta_0} \le n \le d^{\mathsf{s}+1-\delta_0}$ and $d^{\mathsf{S}+\delta_0} \le N \le d^{\mathsf{S}+1-\delta_0}$. Take $\sigma_d(\boldsymbol{x};\boldsymbol{\theta}) = \bar{\sigma}_d(\langle\boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})$ for some activation function $\bar{\sigma}_d : \mathbb{R}\to\mathbb{R}$ satisfying Assumption 3 at level (s,S) (see Section 2.4 in the main text).

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

By rotational invariance, we can decompose $\bar{\sigma}$ in the basis of spherical harmonics (see Section E.1)

$$\sigma_d(\boldsymbol{x};\boldsymbol{\theta}) = \bar{\sigma}_d(\langle \boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d}) = \sum_{k=0}^{\infty} \xi_{d,k} B(\mathbb{S}^{d-1};k) Q_k^{(d)}(\langle \boldsymbol{x},\boldsymbol{\theta}\rangle) = \sum_{k=0}^{\infty} \xi_{d,k} \sum_{s\in[B(d,k)]} Y_{ks}(\boldsymbol{x}) Y_{ks}(\boldsymbol{\theta}),$$

where the distinct eigenvalues are $\xi_{d,k}$ with degeneracy

$$B(\mathbb{S}^{d-1};k) = \frac{d-2+2k}{d-2}\binom{d-3+k}{k}.$$

We have for fixed k, $B(\mathbb{S}^{d-1};k) = \Theta_d(d^k)$. Furthermore, we have uniformly $\sup_{k\geq \ell} B(\mathbb{S}^{d-1};k)^{-1} = O_d(d^{-\ell})$ (see Lemma 1 in [20]). Notice that by Assumption 3 (see for example Lemma 5 in [20]), there exists a constant $C>0$ such that

$$\|\bar{\sigma}_d\|_{L^2}^2 = \sum_{k=0}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) \leq C, \tag{186}$$

which implies that $\xi_{d,k}^2 = O_d(B(\mathbb{S}^{d-1};k)^{-1})$. In particular,

$$\sup_{k>\mathsf{s}} \xi_{d,k}^2 = O_d(d^{-\mathsf{s}-1}), \tag{187}$$

$$\sup_{k>\mathsf{S}} \xi_{d,k}^2 = O_d(d^{-\mathsf{S}-1}). \tag{188}$$

Furthermore, by noting that $\xi_{d,k}^2 = B(\mathbb{S}^{d-1};k)^{-1}\|\overline{\mathsf{P}}_k \bar{\sigma}_d(\langle \boldsymbol{e},\cdot\rangle)\|_{L^2}^2$, conditions (26), (27) and (28) can be rewritten as follows in terms of the coefficients $(\xi_{d,k})_{k\geq 0}$:

$$\min_{k\leq \mathsf{s}} \xi_{d,k}^2 = \Omega_d(d^{-\mathsf{s}}), \tag{189}$$

$$\min_{k\leq \mathsf{S}} \xi_{d,k}^2 = \Omega_d(d^{-\mathsf{S}}), \tag{190}$$

$$\sum_{k=2\max(\mathsf{s},\mathsf{S})+2}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) = \Omega_d(1). \tag{191}$$

Denote $\{\lambda_{d,j}\}_{j\geq 1}$ the eigenvalues $\{\xi_{d,k}\}_{k\geq 0}$ with their degeneracy in non increasing order of their absolute value. Set M and m to be the number of eigenvalues associated to spherical harmonics of degree less or equal to S and s respectively, i.e.,

$$\mathsf{M} = \sum_{k=0}^{\mathsf{S}} B(\mathbb{S}^{d-1};k) = \Theta_d(d^{\mathsf{S}}), \qquad \mathsf{m} = \sum_{k=0}^{\mathsf{s}} B(\mathbb{S}^{d-1};k) = \Theta_d(d^{\mathsf{s}}). \tag{192}$$

Notice that Eqs. (187) and (189) imply that $(\lambda_{d,j})_{j\leq \mathsf{m}}$ corresponds exactly to all the eigenvalues associated to spherical harmonics of degree less or equal to s. Similarly Eqs. (188) and (190) imply that $(\lambda_{d,j})_{j\leq \mathsf{M}}$ corresponds exactly to all the eigenvalues associated to spherical harmonics of degree less or equal to S.

Notice that the diagonal elements of the truncated kernels are given by (for any $\boldsymbol{x}, \boldsymbol{\theta} \in \mathbb{S}^{d-1}(\sqrt{d})$)

$$H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{x}) = \sum_{k=\mathsf{s}+1}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) Q_k^{(d)}(\langle \boldsymbol{x},\boldsymbol{x}\rangle) = \sum_{k=\mathsf{s}+1}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) = \|\overline{\mathsf{P}}_{>\mathsf{s}}\bar{\sigma}_d\|_{L^2}^2 = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}),$$
$$U_{d,>\mathsf{M}}(\boldsymbol{\theta},\boldsymbol{\theta}) = \sum_{k=\mathsf{S}+1}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) Q_k^{(d)}(\langle \boldsymbol{\theta},\boldsymbol{\theta}\rangle) = \sum_{k=\mathsf{S}+1}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) = \|\overline{\mathsf{P}}_{>\mathsf{S}}\bar{\sigma}_d\|_{L^2}^2 = \mathrm{Tr}(\mathbb{U}_{d,>\mathsf{M}}), \tag{193}$$

where we used that $Q_k^{(d)}(d) = 1$.

Step 2. Checking the assumptions at level $\{(N(d), \mathsf{M}(d), n(d), \mathsf{m}(d))\}_{d\geq 1}$.

We are now in position to verify the assumptions of Theorem 1. Choose $u := u(d)$ to be the number of eigenvalues with absolute value $\Omega_d(d^{-2\max(\mathsf{s},\mathsf{S})-2+\delta})$ for some $\delta > 0$ that will be chosen small enough, see Eq. (194). In particular, $(\lambda_{d,j})_{j\in[u]}$ contains all the eigenvalues associated to the spherical harmonics of degree less or equal to $\max(\mathsf{S},\mathsf{s})$, and none of the eigenvalues associated to spherical harmonics of degree $2\max(\mathsf{S},\mathsf{s})+2$ and bigger. We therefore must have $u \geq \max(\mathsf{M}(d), \mathsf{m}(d))$.

Let us verify the conditions of $(N, \mathsf{M}, n, \mathsf{m})$-FMCP in Assumption 1 with the sequence of integers $u(d)$:

(a) The hypercontractivity of the space of polynomials of degree less or equal $2\max(\mathsf{S},\mathsf{s})+1$ is a consequence of a classical result due to Beckner [9] (see Section E.3).

(b) Let us lower bound the right-hand side of Eq. (18). We have

$$\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2 \geq \sum_{k=2\max(\mathsf{s},\mathsf{S})+2}^{\infty} \xi_{d,k}^2 B(\mathbb{S}^{d-1};k) = \Omega_d(1),$$
$$\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4 \leq \Big\{\sup_{j>u} \lambda_{d,j}^2\Big\} \cdot \sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2 = O_d(d^{-2\max(\mathsf{s},\mathsf{S})-2+\delta}) \cdot \sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2.$$

Hence,

$$\frac{\big(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2\big)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4} = \Omega_d(d^{2\max(\mathsf{s},\mathsf{S})+2-\delta}) \geq \max(n,N)^{2+\delta}, \tag{194}$$

for $\delta > 0$ small enough, where we recall that $n \leq d^{\mathsf{s}+1-\delta_0}$ and $N \leq d^{\mathsf{S}+1-\delta_0}$ for some fixed $\delta_0 > 0$.

(c) From Eq. (28) in Assumption 3, we only need to check that for q such that

$$\min(n,N)\max(N,n)^{1/q-1}\log(\max(N,n)) = o_d(1),$$

we have

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[[\mathsf{P}_{>u}\bar{\sigma}_d](\langle \boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})^{2q}\big]^{1/(2q)} = O_d(1).$$

Denote S the set of eigenvalues $\lambda_{d,j}$, with $j > u$, associated to spherical harmonics of degree less of equal to $2\max(\mathsf{s},\mathsf{S})+1$. By triangular inequality, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[[\mathsf{P}_{>u}\bar{\sigma}_d](\langle \boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})^{2q}\big]^{1/(2q)}\\ &\leq \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[[\mathsf{P}_{S}\bar{\sigma}_d](\langle \boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})^{2q}\big]^{1/(2q)} + \mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[[\overline{\mathsf{P}}_{>2\max(\mathsf{s},\mathsf{S})+1}\bar{\sigma}_d](\langle \boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})^{2q}\big]^{1/(2q)} = O_d(1),\end{aligned}$$

where we used that $\mathsf{P}_S\bar{\sigma}_d$ is a polynomial of degree less or equal to $2\max(\mathsf{S},\mathsf{s})+1$ in each variable $\boldsymbol{x}$ and $\boldsymbol{\theta}$ and satisfies the hypercontractivity property (see Lemma 6), i.e.,

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}}\big[[\mathsf{P}_S\bar{\sigma}_d](\langle\boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})^{2q}\big] = O_d(1)\cdot\mathbb{E}_{\boldsymbol{x}}\big[H_{d,S}(\boldsymbol{x},\boldsymbol{x})^q\big] = O_d(1)\cdot\mathrm{Tr}(\mathbb{H}_{d,S})^q = O_d(1),$$

while the bound on $\overline{\mathsf{P}}_{>2\max(\mathsf{s},\mathsf{S})+1}\bar{\sigma}_d$ follows from Assumption 3.(*a*) and Lemma 16 stated below.

(*d*) This is automatically verified because the diagonal elements are constant in this case (Eq. (193)).

Next, we check Assumption 2 at level $(N,\mathsf{M},n,\mathsf{m})$. Consider the overparametrized case $N(d)\geq n(d)$, and therefore $\mathsf{M}\geq\mathsf{m}$. The underparametrized case $N(d)\leq n(d)$ is treated analogously.

(*i*) The eigenvalue sums in Eq. (19) can be estimated as follows

$$\frac{1}{\lambda^2_{d,\mathsf{m}(d)}}\sum_{k=\mathsf{m}(d)+1}^{\infty}\lambda^2_{d,k} = \frac{1}{\xi^2_{d,\mathsf{s}}}\sum_{k=\mathsf{s}+1}^{\infty}\xi^2_{d,k}B(\mathbb{S}^{d-1};k) = O_d(d^{\mathsf{s}}), \tag{195}$$

$$\frac{1}{\lambda^2_{d,\mathsf{m}(d)+1}}\sum_{k=\mathsf{m}(d)+1}^{\infty}\lambda^2_{d,k} = \frac{1}{\xi^2_{d,\mathsf{s}+1}}\sum_{k=\mathsf{s}+1}^{\infty}\xi^2_{d,k}B(\mathbb{S}^{d-1};k) \geq B(\mathbb{S}^{d-1};\mathsf{s}+1) = \Omega_d(d^{\mathsf{s}+1}). \tag{196}$$

The last equality in (195) follows from Eq. (186) and the assumption (189). Hence condition (19) in Assumption 2 is satisfied since, by the statement of Theorem 2, we assume $d^{\mathsf{s}+\delta}\leq n\leq d^{\mathsf{s}+1-\delta}$. Furthermore, by Eq. (192), we have $\mathsf{m}\leq n^{1-\delta}$ for some $\delta>0$ chosen small enough.

(*ii*) The eigenvalue sum in Eq. (20) is

$$\frac{1}{\lambda^2_{d,\mathsf{M}(d)+1}}\sum_{k=\mathsf{M}(d)+1}^{\infty}\lambda^2_{d,k} = \frac{1}{\xi^2_{d,\mathsf{S}+1}}\sum_{k=\mathsf{S}+1}^{\infty}\xi^2_{d,k}B(\mathbb{S}^{d-1};k) \geq B(\mathbb{S}^{d-1};\mathsf{S}+1) = \Omega_d(d^{\mathsf{S}+1}). \tag{197}$$

Hence condition (20) in Assumption 2 is satisfied since, by the statement of Theorem 2, $N\leq d^{\mathsf{S}+1-\delta_0}$. By Eq. (192), we have $\mathsf{M}\leq N^{1-\delta}$ for some $\delta>0$ chosen small enough. □

Lemma 16. *Consider m,ℓ two fixed integers. Assume $|\bar{\sigma}_d(x)|\leq c_0\exp(c_1x^2/(4m))$ with $c_0>0$ and $c_1<1$. Then*

$$\mathbb{E}_{x_1\sim\tau_d^1}\big[\bar{\sigma}_{d,>\ell}(x_1)^{2m}\big] = O_d(1),$$

where τ_d^1 is the marginal distribution of $\langle\boldsymbol{e},\boldsymbol{x}\rangle$ with $\|\boldsymbol{e}\|_2=1$ and $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and we denoted $\bar{\sigma}_{d,>\ell}=\overline{\mathsf{P}}_{>\ell}\bar{\sigma}_d$.

Proof of Lemma 16. Recall that

$$\bar{\sigma}_{d,>\ell}(x) = \bar{\sigma}_d(x) - \sum_{k=0}^{\ell}\xi_{d,k}(\sigma)B(\mathbb{S}^{d-1};k)Q_k^{(d)}(\sqrt{d}x), \tag{198}$$

where $\xi_{d,k}(\sigma)^2B(d,k)\leq\|\bar{\sigma}_d\|^2_{L^2}\leq C$ for some constant $C>0$ (using $|\bar{\sigma}_d(x)|\leq c_0\exp(c_1x^2/(4m))$) and $\sqrt{B(\mathbb{S}^{d-1};k)}Q_k^{(d)}$ is a degree-k polynomial that converges to the Hermite polynomial $\mathrm{He}_k/\sqrt{k!}$ (see Section E.1.3). Therefore, $\bar{\sigma}_{d,>\ell}$ is equal to $\bar{\sigma}_d$ plus a polynomial of degree ℓ with bounded coefficients. In particular, from the assumption $|\bar{\sigma}_d(x)|\leq c_0\exp(c_1x^2/(4m))$, we deduce there exists $c_0'>0$ such that $|\bar{\sigma}_{d,>\ell}(x)|\leq c_0'\exp(c_1x^2/(4m))$, whence:

$$\left|\mathbb{E}_{x_1}[\bar{\sigma}_{d,>\ell}(x_1)^{2m}]\right| \le (c_0')^{2m}\mathbb{E}_{x_1}[\exp(c_1 x_1^2/2)]. \tag{199}$$

Furthermore, recall that $\tau_d^1(\mathrm{d}x) = C_d(1-x^2/d)^{(d-3)/2}\mathbf{1}_{x\in[-\sqrt{d},\sqrt{d}]}\mathrm{d}x \le C\exp(-x^2/2)\mathrm{d}x$. We can therefore upper bound the right hand side of Eq. (199) and use dominated convergence, which concludes the proof. □

D.2. On the hypercube

The proof for the hypercube $\mathscr{Q}^d$ follows from the same proof as for the sphere. We refer to [35] for background on Fourier analysis on $\mathscr{Q}^d$, and Section E.2 for notations that make the analogy with the sphere transparent. In particular, an analog of Lemma 16 follows by noticing that the law $\langle \mathbf{1}, \boldsymbol{x}\rangle/\sqrt{d}$ is a standardized binomial, which can be in terms of the standard normal distribution, times polynomial factors. The only difference comes from the degeneracy

$$B(\mathscr{Q}^d; k) = \binom{d}{\ell}.$$

Hence Assumption 3.(a) only implies $\xi_{d,d-\ell}^2 = O_d(d^{-\ell})$ for the last coefficients, which is the reason for the further requirement Assumption 3.(c).

Let us check that Assumption 3.(c) holds for a class of smooth activation functions. We believe that indeed this assumption holds much more generally, but we leave such generalizations to future work.

Lemma 17. *Consider ℓ a fixed integer. Assume there exist constants $c_0 > 0$ and $c_1 < 1$ such that $|\bar{\sigma}^{(\ell)}(x)| \le c_0\exp(c_1 x^2/4)$ for all $x \in \mathbb{R}$. Then, we have*

$$\max_{k\le\ell} \xi_{d,d-k}(\bar{\sigma})^2 = O_d(d^{-\ell}),$$

where $\xi_{d,d-k}(\bar{\sigma}) = \langle \bar{\sigma}(\langle \boldsymbol{e}, \cdot\rangle), Q_{d-k}(\sqrt{d}\langle \boldsymbol{e}, \cdot\rangle)\rangle_{L^2(\mathscr{Q}^d)}$, and Q_k is the k-th hypercubic Gegenbauer polynomial (see Appendix E.2).

Proof of Lemma 17. By the mean value theorem, we have for any $k \le \ell$,

$$\begin{aligned}
\xi_{d-k,d}(\bar{\sigma}) =& \mathbb{E}_{\boldsymbol{x}}[\bar{\sigma}(\langle \mathbf{1}, \boldsymbol{x}\rangle/\sqrt{d})Q_{d-k}^{(d)}(\langle \mathbf{1}, \boldsymbol{x}\rangle)] \\
=& \mathbb{E}_{\boldsymbol{x}}\Big[x_1\cdots x_{d-k}\bar{\sigma}\Big(\frac{x_1+\ldots+x_d}{\sqrt{d}}\Big)\Big] \\
=& \frac{1}{2}\mathbb{E}_{x_2,\ldots,x_3}\Big[x_2\ldots x_{d-k}\Big(\bar{\sigma}\Big(\frac{1+x_2+\ldots+x_d}{\sqrt{d}}\Big) - \bar{\sigma}\Big(\frac{-1+x_2+\ldots+x_d}{\sqrt{d}}\Big)\Big)\Big] \\
=& \frac{1}{\sqrt{d}}\mathbb{E}_{x_2,\ldots,x_d}\Big[x_2\ldots x_{d-k}\bar{\sigma}^{(1)}(\zeta^1(x_2,\ldots,x_d))\Big],
\end{aligned}$$

where on the third line we integrated over the first coordinate x_1 and on the last line $|\zeta^1(x_2,\ldots,x_d)) - (x_2+\ldots+x_d)/\sqrt{d}| \le 1/\sqrt{d}$. By iterating this computation ℓ times, we get

$$\xi_{d-k,d}(\bar{\sigma}) = \frac{1}{d^{\ell/2}}\mathbb{E}_{x_{\ell+1},\ldots,x_d}\Big[x_{\ell+1}\ldots x_{d-k}\bar{\sigma}^{(\ell)}(\zeta^\ell(x_{\ell+1},\ldots,x_d))\Big],$$

where $|\zeta^\ell(x_{\ell+1},\ldots,x_d) - (x_{\ell+1}+\ldots+x_d)/\sqrt{d}| \le \ell/\sqrt{d}$. Hence,

$$
\begin{aligned}
|\xi_{d-k,d}(\bar{\sigma})| \le & \frac{1}{d^{\ell/2}} \mathbb{E}_{x_{\ell+1},\dots,x_d} \left[|\bar{\sigma}^{(\ell)}(\zeta^{\ell}(x_{\ell+1},\dots,x_d))| \right] \\
\le & \frac{1}{d^{\ell/2}} \mathbb{E}_{X=(x_{\ell+1}+\dots+x_d)/\sqrt{d}} [c_0 \exp(c_1 X^2/2 + c_1 \ell^2/d)] \\
= & O_d(d^{-\ell/2}),
\end{aligned}
$$

where we used that X converges weakly to the standard normal distribution and dominated convergence. □

Appendix E. Technical background

E.1. Functions on the sphere

E.1.1. Functional spaces over the sphere

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

$$
\langle f, g \rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\, g(\boldsymbol{x})\, \tau_d(\mathrm{d}\boldsymbol{x})\,. \tag{200}
$$

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

$$
L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{201}
$$

The dimension of each subspace is given by

$$
\dim(V_{d,\ell}) = B(\mathbb{S}^{d-1}; \ell) = \frac{2\ell + d - 2}{d-2} \binom{\ell + d - 3}{\ell}. \tag{202}
$$

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

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

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

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

$$
\overline{\mathsf{P}}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(\mathbb{S}^{d-1};k)} \langle f, Y_{kl} \rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{203}
$$

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

E.1.2. Gegenbauer polynomials

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell \geq 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_d^1)$, where $\tilde{\tau}_d^1$ is the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle$ when $\boldsymbol{x} \sim \tau_d$, satisfying the normalization condition:

$$\langle Q_k^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle), Q_j^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = \frac{1}{B(\mathbb{S}^{d-1}; k)}\, \delta_{jk}\,. \tag{204}$$

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

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

We will use the following properties of Gegenbauer polynomials

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

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(\mathbb{S}^{d-1}; k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{205}$$

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

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(\mathbb{S}^{d-1}; k)} \sum_{i=1}^{B(\mathbb{S}^{d-1};k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{206}$$

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

$$(\overline{\mathsf{P}}_k f)(\boldsymbol{x}) = B(\mathbb{S}^{d-1}; k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \tau_d(\mathrm{d}\boldsymbol{y})\,. \tag{207}$$

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

$$\xi_{d,k}(\bar{\sigma}) = \int_{[-\sqrt{d}, \sqrt{d}]} \bar{\sigma}(x) Q_k^{(d)}(\sqrt{d}x) \tau_d^1(\mathrm{d}x), \tag{208}$$

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

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

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

$$\mathscr{H}_d f(\boldsymbol{x}) \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} h_d(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d)\, f(\boldsymbol{x}_1)\, \tau_d(\mathrm{d}\boldsymbol{x}_1)\,. \tag{209}$$

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

$$\xi_{d,k}(h_d) = \int_{[-\sqrt{d},\sqrt{d}]} h_d\big(x/\sqrt{d}\big) Q_k^{(d)}(\sqrt{d}x)\tau_{d-1}^1(\mathrm{d}x)\,. \tag{210}$$

E.1.3. Hermite polynomials

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

$$\mathbb{E}\big\{\mathrm{He}_j(G)\,\mathrm{He}_k(G)\big\} = k!\,\delta_{jk}\,. \tag{211}$$

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

$$g(x) = \sum_{k=0}^{\infty} \frac{\mu_k(g)}{k!}\,\mathrm{He}_k(x)\,, \qquad \mu_k(g) \equiv \mathbb{E}\big\{g(G)\,\mathrm{He}_k(G)\big\}\,. \tag{212}$$

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

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(\mathbb{S}^{d-1};k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}\,. \tag{213}$$

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

$$\mu_k(\bar{\sigma}) = \lim_{d\to\infty} \xi_{d,k}(\bar{\sigma})(B(\mathbb{S}^{d-1};k)k!)^{1/2}, \tag{214}$$

where $\mu_k(\bar{\sigma})$ and $\xi_{d,k}(\bar{\sigma})$ are given in Eq. (212) and (208).

E.2. Functions on the hypercube

Fourier analysis on the hypercube is a well studied subject [35]. The purpose of this section is to introduce some notations that make the correspondence with proofs on the sphere straightforward. For convenience, we will adopt the same notations as for their spherical case.

E.2.1. Fourier basis

Denote $\mathscr{Q}^d = \{-1,+1\}^d$ the hypercube in d dimension. Let us denote τ_d to be the uniform probability measure on $\mathscr{Q}^d$. All the functions will be assumed to be elements of $L^2(\mathscr{Q}^d,\tau_d)$ (which contains all the bounded functions $f : \mathscr{Q}^d \to \mathbb{R}$), with scalar product and norm denoted as $\langle\cdot,\cdot\rangle_{L^2}$ and $\|\cdot\|_{L^2}$:

$$\langle f, g\rangle_{L^2} \equiv \int_{\mathscr{Q}^d} f(\boldsymbol{x})g(\boldsymbol{x})\tau_d(\mathrm{d}\boldsymbol{x}) = \frac{1}{2^n}\sum_{\boldsymbol{x}\in\mathscr{Q}^d} f(\boldsymbol{x})g(\boldsymbol{x}).$$

Notice that $L^2(\mathscr{Q}^d, \tau_d)$ is a 2^n dimensional linear space. By analogy with the spherical case we decompose $L^2(\mathscr{Q}^d, \tau_d)$ as a direct sum of $d+1$ linear spaces obtained from polynomials of degree $\ell = 0, \ldots, d$

$$L^2(\mathscr{Q}^d, \tau_d) = \bigoplus_{\ell=0}^{d} V_{d,\ell}.$$

For each $\ell \in \{0, \ldots, d\}$, consider the Fourier basis $\{Y_{\ell,S}^{(d)}\}_{S\subseteq[d],|S|=\ell}$ of degree ℓ, where for a set $S \subseteq [d]$, the basis is given by

$$Y_{\ell,S}^{(d)}(\boldsymbol{x}) \equiv x^S \equiv \prod_{i\in S} x_i.$$

It is easy to verify that (notice that $x_i^k = x_i$ if k is odd and $x_i^k = 1$ if k is even)

$$\langle Y_{\ell,S}^{(d)}, Y_{k,S'}^{(d)}\rangle_{L^2} = \mathbb{E}[x^S \times x^{S'}] = \delta_{\ell,k}\delta_{S,S'}.$$

Hence $\{Y_{\ell,S}^{(d)}\}_{S\subseteq[d],|S|=\ell}$ form an orthonormal basis of $V_{d,\ell}$ and

$$\dim(V_{d,\ell}) = B(\mathscr{Q}^d; \ell) = \binom{d}{\ell}.$$

As above, we will omit the superscript (d) in $Y_{\ell,S}^{(d)}$ when clear from the context.

E.2.2. Hypercubic Gegenbauer

We consider the following family of polynomials $\{Q_\ell^{(d)}\}_{\ell=0,\ldots,d}$ that we will call hypercubic Gegenbauer, defined as

$$Q_\ell^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(\mathscr{Q}^d; \ell)}\sum_{S\subseteq[d],|S|=\ell} Y_{\ell,S}^{(d)}(\boldsymbol{x})Y_{\ell,S}^{(d)}(\boldsymbol{y}).$$

Notice that the right hand side only depends on $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ and therefore these polynomials are uniquely defined. In particular,

$$\langle Q_\ell^{(d)}(\langle \mathbf{1}, \cdot\rangle), Q_k^{(d)}(\langle \mathbf{1}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)}\delta_{\ell k}.$$

Hence $\{Q_\ell^{(d)}\}_{\ell=0,\ldots,d}$ form an orthogonal basis of $L^2(\{-d, -d+2, \ldots, d-2, d\}, \tilde{\tau}_d^1)$ where $\tilde{\tau}_d^1$ is the distribution of $\langle \mathbf{1}, \boldsymbol{x}\rangle$ when $\boldsymbol{x} \sim \tau_d$, i.e., $\tilde{\tau}_d^1 \sim 2\mathrm{Bin}(d, 1/2) - d/2$.

We have

$$\langle Q_\ell^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)}Q_k(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\delta_{\ell k}.$$

For a function $\bar{\sigma}(\cdot/\sqrt{d}) \in L^2(\{-d, -d+2, \ldots, d-2, d\}, \tilde{\tau}_d^1)$, denote its hypercubic Gegenbauer coefficients $\xi_{d,k}(\bar{\sigma})$ to be

$$\xi_{d,k}(\bar{\sigma}) = \int_{\{-d,-d+2,\ldots,d-2,d\}} \bar{\sigma}(x/\sqrt{d})Q_k^{(d)}(x)\tilde{\tau}_d^1(\mathrm{d}x).$$

Notice that by weak convergence of $\langle \mathbf{1}, \boldsymbol{x}\rangle/\sqrt{d}$ to the normal distribution, we have also convergence of the (rescaled) hypercubic Gegenbauer polynomials to the Hermite polynomials, i.e., for any fixed k, we have

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(\mathscr{Q}^d; k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}. \tag{215}$$

E.3. Hypercontractivity of Gaussian measure and uniform distributions on the sphere and the hypercube

By Holder's inequality, we have $\|f\|_{L^p} \le \|f\|_{L^q}$ for any f and any $p \le q$. The reverse inequality does not hold in general, even up to a constant. However, for some measures, the reverse inequality will hold for some sufficiently nice functions. These measures satisfy the celebrated hypercontractivity properties [22,13,8,9].

Lemma 18 *(Hypercube hypercontractivity [8]). For any $\ell = \{0, \dots, d\}$ and $f_* \in L^2(\mathscr{Q}^d)$ to be a degree ℓ polynomial, then for any integer $q \ge 2$, we have*

$$\|f_*\|^2_{L^q(\mathscr{Q}^d)} \le (q-1)^\ell \cdot \|f_*\|^2_{L^2(\mathscr{Q}^d)}.$$

Lemma 19 *(Spherical hypercontractivity [9]). For any $\ell \in \mathbb{N}$ and $f_* \in L^2(\mathbb{S}^{d-1})$ to be a degree ℓ polynomial, for any $q \ge 2$, we have*

$$\|f_*\|^2_{L^q(\mathbb{S}^{d-1})} \le (q-1)^\ell \cdot \|f_*\|^2_{L^2(\mathbb{S}^{d-1})}.$$

Lemma 20 *(Gaussian hypercontractivity). For any $\ell \in \mathbb{N}$ and $f \in L^2(\mathbb{R}, \gamma)$ to be a degree ℓ polynomial on $\mathbb{R}$, where γ is the standard Gaussian distribution. Then for any $q \ge 2$, we have*

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

The Gaussian hypercontractivity is a direct consequence of hypercube hypercontractivity.

References

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruosong Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, in: International Conference on Machine Learning, in: PMLR, 2019, pp. 322–332.

[2] Ahmed El Alaoui, Michael W. Mahoney, Fast randomized kernel ridge regression with statistical guarantees, Adv. Neural Inf. Process. Syst. (2015) 775–783.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, Yingyu Liang, Learning and generalization in overparameterized neural networks, going beyond two layers, Adv. Neural Inf. Process. Syst. (2019) 6158–6169.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song, A convergence theory for deep learning via over-parameterization, in: International Conference on Machine Learning, in: PMLR, vol. 97, 2019, pp. 242–252.

[5] Francis Bach, Sharp analysis of low-rank kernel matrix approximations, in: Conference on Learning Theory, 2013, pp. 185–209.

[6] Francis Bach, On the equivalence between kernel quadrature rules and random feature expansions, J. Mach. Learn. Res. 18 (2017) 714–751.

[7] Maria-Florina Balcan, Avrim Blum, Santosh Vempala, Kernels as features: on kernels, margins, and low-dimensional mappings, Mach. Learn. 65 (1) (2006) 79–94.

[8] William Beckner, Inequalities in Fourier analysis, Ann. Math. (1975) 159–182.

[9] William Beckner, Sobolev inequalities, the Poisson semigroup, and analysis on the sphere S^n, Proc. Natl. Acad. Sci. 89 (11) (1992) 4816–4819.

[10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.

[11] Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proc. Natl. Acad. Sci. 116 (32) (2019) 15849–15854.

[12] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, Alexander Tsigler, Benign overfitting in linear regression, Proc. Natl. Acad. Sci. (2020).

[13] Aline Bonami, Etude des coefficients de Fourier des fonctions de $L^p(G)$, Ann. Inst. Fourier 20 (1970) 335–402.

[14] Mikhail Belkin, Alexander Rakhlin, Alexandre B. Tsybakov, Does data interpolation contradict statistical optimality?, in: The 22nd International Conference on Artificial Intelligence and Statistics, in: PMLR, vol. 89, 2019, pp. 1611–1619.
[15] Alain Berlinet, Christine Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Springer Science & Business Media, 2011.
[16] Andrea Caponnetto, Ernesto De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (3) (2007) 331–368.
[17] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, Xiyu Zhai, Gradient descent finds global minima of deep neural networks, in: International Conference on Machine Learning, in: PMLR, 2019, pp. 1675–1685.
[18] Richard M. Dudley, Real Analysis and Probability, CRC Press, 2018.
[19] Simon S. Du, Xiyu Zhai, Barnabas Poczos, Aarti Singh, Gradient descent provably optimizes over-parameterized neural networks, in: International Conference on Learning Representations, 2018.
[20] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Andrea Montanari, Linearized two-layers neural networks in high dimension, Ann. Stat. 49 (2) (2021) 1029–1054.
[21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Andrea Montanari, When do neural networks outperform kernel methods?, Adv. Neural Inf. Process. Syst. 33 (2020).
[22] Leonard Gross, Logarithmic sobolev inequalities, Am. J. Math. 97 (4) (1975) 1061–1083.
[23] Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, Ann. Stat. (2021), in press, arXiv:1903.08560.
[24] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer, 2009.
[25] Arthur Jacot, Franck Gabriel, Clément Hongler, Neural tangent kernel: convergence and generalization in neural networks, Adv. Neural Inf. Process. Syst. (2018) 8571–8580.
[26] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, Franck Gabriel, Kernel alignment risk estimator: risk prediction from training data, Adv. Neural Inf. Process. Syst. 33 (2020).
[27] Yuanzhi Li, Yingyu Liang, Learning overparameterized neural networks via stochastic gradient descent on structured data, Adv. Neural Inf. Process. Syst. (2018) 8157–8166.
[28] David G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[29] Tengyuan Liang, Alexander Rakhlin, Just interpolate: kernel "ridgeless" regression can generalize, Ann. Stat. 48 (3) (2020) 1329–1347.
[30] Tengyuan Liang, Alexander Rakhlin, Xiyu Zhai, On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels, in: Conference on Learning Theory, in: PMLR, 2020, pp. 2683–2711.
[31] Song Mei, Andrea Montanari, The generalization error of random features regression: precise asymptotics and double descent curve, Commun. Pure Appl. Math. (2019).
[32] Song Mei, Theodor Misiakiewicz, Andrea Montanari, Learning with invariances in random features and kernel models, in: Conference on Learning Theory, 2021.
[33] Chao Ma, Stephan Wojtowytsch, Lei Wu, Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't, CSIAM Trans. Appl. Math. 1 (2020) 561–615.
[34] Andrea Montanari, Yiqiao Zhong, The interpolation phase transition in neural networks: memorization and generalization under lazy training, arXiv:2007.12826, 2020.
[35] Ryan O'Donnell, Analysis of Boolean Functions, Cambridge University Press, 2014.
[36] Samet Oymak, Mahdi Soltanolkotabi, Towards moderate overparameterization: global convergence guarantees for training shallow neural networks, IEEE J. Select. Areas Inform. Theory 1 (1) (2020) 84–105.
[37] Alessandro Rudi, Raffaello Camoriano, Lorenzo Rosasco, Less is more: Nyström computational regularization, Adv. Neural Inf. Process. Syst. (2015) 1657–1665.
[38] Ali Rahimi, Benjamin Recht, Random features for large-scale kernel machines, Adv. Neural Inf. Process. Syst. (2008) 1177–1184.
[39] Ali Rahimi, Benjamin Recht, Weighted sums of random kitchen sinks: replacing minimization with randomization in learning, Adv. Neural Inf. Process. Syst. (2009) 1313–1320.
[40] Alessandro Rudi, Lorenzo Rosasco, Generalization properties of learning with random features, Adv. Neural Inf. Process. Syst. (2017) 3215–3225.
[41] Alexander Tsigler, Peter L. Bartlett, Benign overfitting in ridge regression, arXiv:2009.14286, 2020.
[42] Roman Vershynin, Introduction to the non-asymptotic analysis of random matrices, arXiv:1011.3027, 2010.
[43] Martin J. Wainwright, High-Dimensional Statistics: A Non-asymptotic Viewpoint, vol. 48, Cambridge University Press, 2019.
[44] Per-Åke Wedin, Perturbation bounds in connection with singular value decomposition, BIT Numer. Math. 12 (1) (1972) 99–111.
[45] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, Zhi-Hua Zhou, Nyström method vs random Fourier features: a theoretical and empirical comparison, Adv. Neural Inf. Process. Syst. (2012) 476–484.
[46] Difan Zou, Yuan Cao, Dongruo Zhou, Quanquan Gu, Stochastic gradient descent optimizes over-parameterized deep relu networks, Mach. Learn. 109 (2020) 467–492.