Iteratively reweighted kernel machines efficiently learn sparse functions

Libin Zhu[*]   Damek Davis[†]   Dmitriy Drusvyatskiy[‡]   Maryam Fazel[§]

May 2025

Abstract

The impressive practical performance of neural networks is often attributed to their ability to learn low-dimensional data representations and hierarchical structure directly from data. In this work, we argue that these two phenomena are not unique to neural networks, and can be elicited from classical kernel methods. Namely, we show that the derivative of the kernel predictor can detect the influential coordinates with low sample complexity. Moreover, by iteratively using the derivatives to reweight the data and retrain kernel machines, one is able to efficiently learn hierarchical polynomials with finite leap complexity. Numerical experiments illustrate the developed theory.

1 Introduction

Over the past decade, deep learning has revolutionized the field of artificial intelligence, driving remarkable advances in areas such as computer vision, natural language processing, reinforcement learning and robotics. While impressive empirical achievements are apparent, understanding *why* deep learning methods are so successful remains an active area of research. In particular, it is now believed that two closely related phenomena—called *feature learning* and *hierarchical learning*—largely underpin deep learning's impressive performance.

Heuristically, feature learning refers to the fact that gradient methods for training neural networks automatically filter out much of the irrelevant information in the data and provide a concise summary of only the most salient aspects for the learning task at hand. As a rudimentary example, detecting whether a person in an image is wearing a necktie only requires inspecting the neck area. Gradient methods for this classification task are able to automatically filter out the irrelevant pixels in the image by learning appropriate weights in the network. Importantly, feature learning implies that the trained neural network automatically performs a kind of nonlinear dimension reduction, which then leads to better generalization. Feature learning is a surprising phenomenon because it suggests that learning a data-dependent compression mechanism is an easier problem than making predictions directly. See, for example, papers [9, 27] which explore feature learning in neural networks.

The second phenomenon is rooted in the observation that typical learning tasks are hierarchical, with features appearing in a natural-order of significance. For example, an airplane in an image may be detected by first learning simpler concepts like 'doors', 'wheels', 'wings', and so forth. There is now ample evidence that gradient methods for training neural networks are able to gradually learn these intermediate concepts, which significantly improves the overall sample efficiency. We refer the reader to [3, 14, 40, 74] for details.

Although recent work in AI has focused almost exclusively on neural networks, these models do have downsides. Neural networks are trained by running gradient-based methods on highly nonlinear and nonconvex functions. As a result, the training process can be brittle, requiring significant computational resources for tuning algorithmic parameters such as step-size, batch-size, initialization scale, momentum terms, etc. Consequently, it is worthwhile to ask whether simpler and more interpretable models can be as effective at learning features and hierarchical structure. One apparent candidate is kernel ridge regression—the preeminent model in machine learning prior to deep learning's ascension.

> In this work, we argue that both phenomena of feature learning and hierarchical learning are not unique to neural networks, and can be elicited by iteratively retraining classical kernel machines.

Consequently, our paper suggests that kernel methods, appropriately instantiated, are much more powerful than previously thought and are worth revisiting in contemporary settings. The algorithms we consider are largely inspired by the classical iteratively reweighted least squares (IRLS) for sparse recovery [29] and the recently introduced recursive feature machines (RFM) algorithm [60] for learning multi-index models. Indeed, the retraining process we introduce is closely related to the "diagonal" variant of RFM for learning sparse non-linear models.

Importantly, the kernel-based algorithms we consider make no distributional assumptions on the data and can be applied on any supervised learning task. That being said, the theoretical guarantees we develop rely on strong distributional assumptions such as Gaussianity or uniform sampling on the hypercube. The reader should regard these assumptions as a "simplification of reality" that allows us to theoretically justify the use of an algorithm in more practical settings. This is in contrast for example to Fourier-based techniques for learning on the hypercube [52, 56], which are tailor-made for this specific task. In this sense, our work is in line with the vast literature on compressed sensing [17, 19, 31, 32], structured signal recovery [5, 22, 62, 71], and gradient dynamics for neural networks [27, 73], where generative data assumptions are imposed in order to support the use of a general purpose algorithm in practice.

1.1 Summary of results

Setting the stage, consider the problem of learning a degree ℓ polynomial f on $\mathbb{R}^d$ from n samples $(x^{(i)}, y_i)$ with noisy labels,

$$y_i = f(x^{(i)}) + \varepsilon_i \qquad \text{for} \quad i = 1, \ldots, n,$$

where the feature vectors $x^{(i)}$ are drawn independently from some distribution μ on $\mathbb{R}^d$ and the noise terms are independent Gaussian $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Since we aim to learn a degree ℓ polynomial, we focus on the high-dimensional regime where n scales polynomially in d, that is $n = d^{p+\delta}$ for some small constant $\delta \in (0, 1)$ and an integer p. Note that the interesting regime is when the power p is smaller, and in some cases *much smaller*, than the degree ℓ of the target polynomial.

Now fix a positive semidefinite inner product kernel and let $\hat{f}$ be the Kernel Ridge Regression (KRR) estimator fit to the data set. A central question is to understand the generalization error $\|\hat{f} - f\|_{L_2(\mu)}^2$ of the prediction function. Since p can be smaller than the degree of the target function, the generalization error will typically not tend to zero as d increases. Nonetheless, in a number of interesting cases, the error can be controlled based on an expansion of f in a distinguished orthonormal polynomial basis of $L_2(\mu)$. For example, for Gaussian, spherical, and hypercube data, the relevant polynomial bases are comprised of Hermite, Gaugenbauer, and Fourier-Walsh polynomials, respectively. In particular, the seminal work [39] showed that for the spherical data, the generalization error of $\hat{f}$ satisfies

$$\|\hat{f} - f\|_{L_2(\mu)}^2 = \|f_{>p}\|_{L_2(\mu)}^2 + o_{d,\mathbb{P}}(1), \tag{1}$$

where $f_{>p}$ is the function obtained by keeping the higher-order terms (the polynomial bases of degree greater than p) in the expansion of f. Further extensions of the estimate (1) appear in [53].

Looking at equation (1), we see that $\hat{f}$ asymptotically recovers f only in the regime $p \geq \ell$. In particular, it rules out favorable performance of kernel methods even if f has "low intrinsic complexity",

e.g., dependence on only a few coordinates. In this work, we show how to overcome this limitation in the case of such "sparse" functions. The method we propose alternates between two stages: (1) "active" coordinate detection and (2) kernel ridge regression on reweighted data.

1.1.1 Sparse feature learning.

As a motivating example, Figure 1a (blue curve) depicts the generalization error of the KRR predictor $\hat{f}$ when learning the degree three polynomial with a Laplace kernel and uniform sampling on the hypercube:

$$f(x) = x_1 + x_2 + x_3 + x_1 x_2 x_3 \qquad \text{for } x \in \mathbb{R}^d. \tag{2}$$

Even though this polynomial depends only on the first three coordinates and therefore has low intrinsic complexity, the generalization error remains large even in the regime $n = d^2$ as implied by (1). Thus, kernel methods are not adaptive to the low complexity of the function that is being learned. This observation often serves as the reason to discard kernel methods and instead focus on neural networks. Interestingly, when learning the cubic polynomial (2), the KRR predictor $\hat{f}$ performs comparably to neural networks (orange curve) trained by Adam in the low sample regime $d^{0.3} \le n \le d^{1.0}$ and significantly worse in the high sample regime $d^{1.2} \le n \le d^{1.3}$ (Figure 1a).
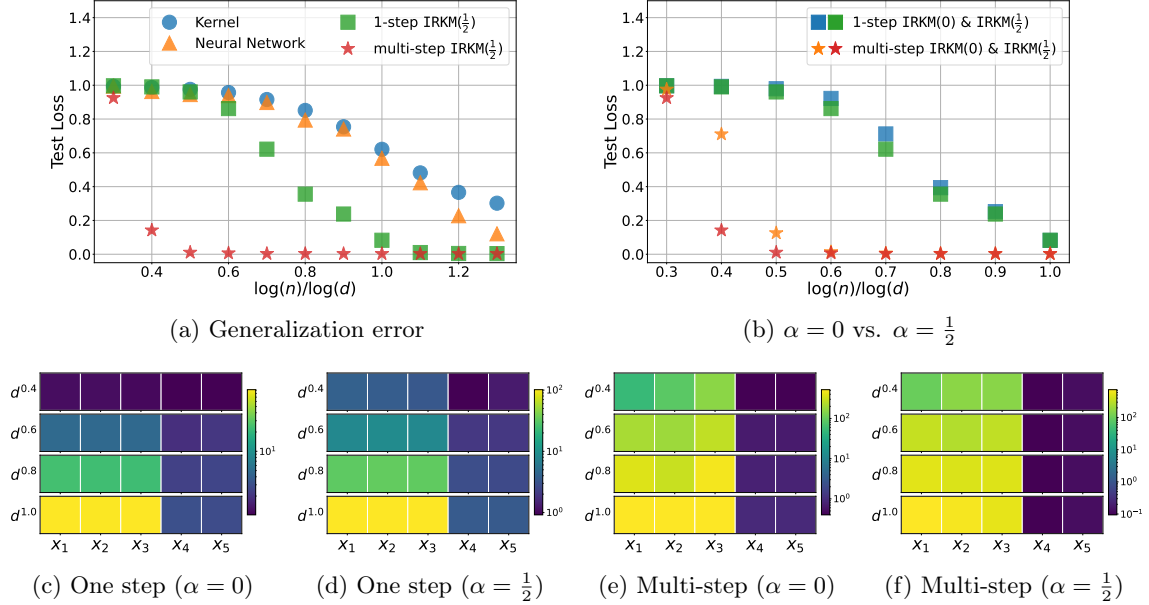

(a) Generalization error

(b) $\alpha = 0$ vs. $\alpha = \frac{1}{2}$

(c) One step ($\alpha = 0$)   (d) One step ($\alpha = \frac{1}{2}$)   (e) Multi-step ($\alpha = 0$)   (f) Multi-step ($\alpha = \frac{1}{2}$)

Figure 1: **(a) Test loss for kernel machines, neural networks and** IRKM$(\frac{1}{2})$**. (b) Comparison of** IRKM(α) **performance for** $\alpha = 0$ **versus** $\alpha = \frac{1}{2}$**. (c,d) The empirical coordinate weights at the first step of** IRKM(α) **with** $\alpha = 0$ **and** $\frac{1}{2}$ **respectively. (e,f) The empirical coordinate weights after** T **steps of** IRKM(α) **with** $\alpha = 0$ **and** $\frac{1}{2}$ **respectively.** The samples are i.i.d. uniformly drawn from the hypercube $\{\pm 1\}^d$ with $d = 500$ and the label is $y = x_1 + x_2 + x_3 + x_1 x_2 x_3 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. We use the Laplacian kernel for both kernel machine and IRKM(α). The IRKM(α) algorithm is run for at most $T = 20$ steps. We train a fully-connected network by Adam, varying the widths and batch sizes as $\{128, 256, 512\}$, depth as $\{2, 3, 4, 5, 6\}$ and learning rate as $\{10^{-3}, 10^{-4}\}$ and report the smallest test loss. For all experiments, we report an average of 10 independent runs.

Detecting coordinates with empirical weights. On the other hand, wishful thinking suggests that it may be possible to extract the relevant coordinates (features) from $\hat{f}$ directly, even when $\hat{f}$ does not generalize well. One simple approach is to use an empirical plug-in estimator $\mathbb{E}_n(\partial_j \hat{f})^2 :=$

$\frac{1}{n} \sum_{i=1}^{n} (\partial_j \hat{f}(x^{(i)}))^2$ to measure the sensitivity of the prediction function $\hat{f}$ to the j'th coordinate. Returning to our running example (2), Figure 1c depicts the values $\mathbb{E}_n(\partial_j \hat{f})^2$ for different sample sizes n. The first three coordinates are clearly identified as being the most significant. Our first contribution is to show that this is no accident: the empirical coordinate weights $\mathbb{E}_n(\partial_j \hat{f})^2$ allow us to asymptotically estimate the population coordinate weights for the truncation of f. In all theorem statements appearing in the introduction, we will assume that μ is the uniform measure on the hypercube $\{\pm 1\}^d$.

Theorem 1.1 (Informal). *Consider the regime $n = d^{p+\delta}$ for any $\delta \in (0, 1)$ and $p \in \mathbb{N}$. Then there exists a constant $c > 0$ such that the following estimate holds uniformly over all coordinates:*

$$\frac{\mathbb{E}_n(\partial_j \hat{f})^2}{1 + o_{d,\mathbb{P}}(1)} = \underbrace{\mathbb{E}(\partial_j f_{\leq p})^2}_{order\ p} + c \cdot d^{2\delta-2} \cdot \underbrace{\mathbb{E}(\partial_j f_{p+1})^2}_{order\ p+1} + \underbrace{O_{d,\mathbb{P}}(d^{-1})}_{noise}. \tag{3}$$

Theorem 1.1 is conceptually interesting because it shows that kernel ridge regression can detect relevant coordinates despite exhibiting a high generalization error. One unsatisfying aspect of (3) is that the term $d^{2\delta-2}\mathbb{E}(\partial_j f_{p+1})^2$ is only meaningful in the high-sampling regime $\delta \in (\frac{1}{2}, 1)$, since in the complementary regime $\delta \in (0, \frac{1}{2})$ this term is dominated by the noise $O_{d,\mathbb{P}}(d^{-1})$ and therefore the "signal" $\mathbb{E}(\partial_j f_{p+1})^2$ is drowned out by the noise. With this in mind, we will now show that the coefficient $d^{2\delta-2}$ in (3) can be reduced by a square root factor to $d^{\delta-1}$ by carefully modifying the estimator $\mathbb{E}_n(\partial_j \hat{f})^2$. This improved dependence on dimension will ultimately improve the sample efficiency of learning by a significant factor of $\sqrt{d}$. In particular, returning to our running example (2), the modified estimator will be able to learn the first three coordinates with $n = d^\delta$ samples *for any $\delta \in (0, 1)$.* This is in contrast to the pure empirical weights $\mathbb{E}_n(\partial_j \hat{f})^2$ that require at least $n = \sqrt{d}$ samples to learn the first three coordinates; see Figure 1c.

Improved coordinate detection with the DN-estimator. The key idea underlying the modified estimator is to introduce a weighted kernel function

$$K_w(x, z) = K(\sqrt{w} \odot x, \sqrt{w} \odot z),$$

parametrized by a weight vector $w \in \mathbb{R}^d$.[1] Intuitively, the weight vector w encodes the importance of each coordinate for prediction. Now let $\hat{f}_w$ be the KRR estimator with respect to the kernel K_w and let $\|\cdot\|_w$ denote the RKHS norm induced by K_w. Intuitively, seeking to find the relevant coordinates of f, we should focus on those weights w for which the norm of the KRR predictor $\|\hat{f}_w\|_w^2$ is as small as possible. Since the function $w \mapsto \|\hat{f}_w\|_w^2$ is difficult to optimize directly, we may instead follow its gradient. A quick computation shows that up to a sign flip, the partial derivatives of this function can be approximated[2] by

$$\mathcal{D}_j(w) := \beta_w^\top \frac{\partial K_w(X,X)}{\partial w_j} \beta_w,$$

where $\beta_w \in \mathbb{R}^n$ are the dual coefficients of f_w.[3] We call $\mathcal{D}_j(w)$ the *derivative norm (DN) estimator*.[4] The following theorem shows that the average of the two estimators $\mathbb{E}_n(\partial_j \hat{f})^2$ and $\mathcal{D}_j(w)$ with $w = \mathbf{1}_d$ enjoys an improved guarantee over (3), wherein the coefficient $d^{2\delta-2}$ is replaced by $d^{\delta-1}$.

Theorem 1.2 (Informal). *Consider the regime $n = d^{p+\delta}$ for any $\delta \in (0, 1)$ and $p \in \mathbb{N}$. Then there exists a constant $c > 0$ such that the following estimate holds uniformly over all coordinates:*

$$\frac{\mathbb{E}_n(\partial_j \hat{f})^2 + \frac{1}{n}\mathcal{D}_j(\mathbf{1}_d)}{1 + o_{d,\mathbb{P}}(1)} = \underbrace{\mathbb{E}(\partial_j f_{\leq p})^2}_{order\ p} + c \cdot d^{\delta-1} \cdot \underbrace{\mathbb{E}(\partial_j f_{p+1})^2}_{order\ p+1} + \underbrace{O_{d,\mathbb{P}}(d^{-1})}_{noise}. \tag{4}$$

---

[1]The symbol $v \odot w := (v_i w_i)_{i=1}^d$ denotes the Hadamard product of two vectors v and w.

[2]by dropping the term that scales linearly in the ridge parameter

[3]This means that we can write $\hat{f}_w(x) = K_w(x, X)\beta_w$.

[4]We note that since K is a PSD inner-product kernel, the matrix $\frac{\partial K_w(X,X)}{\partial w_j}$ is easily seen to be positive semidefinite and therefore the DN estimator $\mathcal{D}_j(w)$ is nonnegative.

4

Returning to our running example (2), Figure 1d shows that the first three coordinates are already detected by the averaged estimator in the setting $n = d^{0.4}$, but not by the pure empirical weights $\mathbb{E}_n(\partial_j \hat{f})^2$. The reason is that these are the important coordinates for the first-order truncation $f_{\leq 1}(x) = x_1 + x_2 + x_3$. Moreover, note that in this example the higher-order terms of f depend only on x_1, x_2, x_3. In other words, all influential coordinates of the third-order polynomial f, namely those coordinates r that satisfy $\mathbb{E}(\partial_r f)^2 > 0$, are fully captured by its first-order truncation. How can we use the averaged estimator to learn the function f in this example? A natural idea is to introduce a reweighting scheme. Namely, one may draw a new set of samples, rescale the coordinates of the data using averaged estimator, and solve a new kernel regression problem which yields a new kernel estimator $\hat{f}_2(x)$, etc. We summarize this iterative process in the following algorithm, which we call Iteratively Reweighted Kernel Machines (IRKM). Throughout, the symbol $v^{\odot 2} := v \odot v$ denotes the entrywise square.

Algorithm 1 IRKM(α)   **I**teratively **R**eweighted **K**ernel **M**achines

---

1: **Input:** averaging parameter $\alpha \in [0, 1]$.
2: **Initialize:** $w = \mathbf{1}_d$ and fix a safeguard parameter $\epsilon_s > 0$.
3: **for** $t = 1, \ldots, T$ **do**:
4:     Sample fresh data $\{(x^{(i)}, y_i)\}_{i=1}^n$ and define the kernel $K_w(x, z) = K(\sqrt{w} \odot x, \sqrt{w} \odot z)$.
5:     Obtain predictor $\hat{f}_t$ from Kernel Ridge Regression on the data $\{(x^{(i)}, y_i)\}_{i=1}^n$ with the kernel K_w.
6:     Compute the first weight vector: $w^{[1]} = \epsilon_s \mathbf{1}_d + \frac{1}{n} \sum_{i=1}^n [\nabla_x \hat{f}_t(x^{(i)})]^{\odot 2}$.
7:     Compute the second weight vector: $w^{[2]} = \epsilon_s \mathbf{1}_d + \frac{1}{n} \mathcal{D}(w) \odot w$.
8:     Normalize the weights $w^{[k]} = \frac{d}{\|w^{[k]}\|_1} \cdot w^{[k]}, \quad k = 1, 2$.
9:     Update the weight vector $w = (1 - \alpha) \cdot w^{[1]} + \alpha \cdot w^{[2]}$.
10: **end for**

IRKM(α) is closely related to and inspired by existing algorithms, in particular, iteratively reweighted least squares for sparse recovery [29] and low-rank matrix recovery [54] and recursive feature machines [60] for learning multi-index models. We will comment on these connections at the end of the introduction. In the meantime, the main takeaway is that operationally IRKM(α) proceeds by solving a sequence of kernel ridge regression problems on freshly sampled data that is reweighted by the average of the empirical coordinate weights and the DN estimator. More precisely, step (4) samples fresh data and forms a reweighted kernel K_w based on the current weight vector w. Step (5) obtains a KRR predictor based on the kernel K_w. Steps (6) and (7) compute the empirical coordinate weights $\mathbb{E}(\partial f_{\leq p})^2$ and the DN-estimator $\mathcal{D}(w)$. Note that we safeguard the two intermediate weight vectors by adding a small multiple of the all-ones vector and the second weight vector is obtained as a Hadamard product of the DN estimator with the current weight vector w analogously to a multiplicative weight update algorithm. Step (8) normalizes the two intermediate weight vectors and (9) averages them to obtain the weight vector w.

Returning to our example, the green curve in Figure 1a depicts the generalization error of $\hat{f}_2$ for various values of n with the mixing parameter $\alpha = \frac{1}{2}$, which is clearly superior to both the standard kernel estimator and neural networks trained by Adam. Interestingly, iterating IRKM($\frac{1}{2}$) for a few extra steps leads to an even further decrease in generalization error (red) and better identification of relevant coordinates in Figure 1f. The following theorem identifies the reason for this improved performance: by running IRKM($\frac{1}{2}$) for a constant number of iterations, all features with order not exceeding $p + 1$ will be learned. Consequently, as long as all relevant coordinates of f are relevant for $f_{\leq p+1}$, after finitely many steps the predictor returned by IRKM($\frac{1}{2}$) is an asymptotically consistent estimator for f. This is the content of the following theorem.

Theorem 1.3 (Informal). *Suppose that the target function $f(x) = g(x_1, \ldots, x_r)$ is a degree ℓ polynomial on $\mathbb{R}^d$. Suppose that the weights $\mathbb{E}[\partial_j f]^2$ are uniformly bounded away from zero for all $j = 1, \ldots, r$ as $d \to \infty$. Consider the regime $n = d^{p+\delta}$ for any $\delta \in (0, 1)$ and choose any $\alpha \in (0, 1)$. Suppose moreover that the nondegeneracy condition holds:*

* $\mathbb{E}[\partial_j f_{\leq p+1}]^2$ *is bounded away from zero uniformly in $j = 1, \ldots, r$ as $d \to \infty$,*

5

Then there exists a horizon $T = O_d(1)$ such that, under an appropriate choice of the safeguard parameter ϵ_{S}, the predictor $\hat{f}_T$ returned by `IRKM`(α) *satisfies:*

$$\left\| \hat{f}_T - f \right\|_{L_2(\mu)}^2 = o_{d,\mathbb{P}}(1).$$

Interestingly, returning to our running example, Figure 1b shows that `IRKM`$(\frac{1}{2})$ significantly outperforms `IRKM`(0) thereby highlighting the importance of the DN-estimator.

1.1.2 Hierarchical learning.

Theorem 1.3 crucially relies on the assumption that all influential coordinates for f are influential for its truncation $f_{\leq p+1}$; this is the case in the running example (2). As a concrete example where this fails, consider learning the function

$$f(x) = x_1 + x_2 + x_1 x_2 x_3 + x_1 x_2 x_3 x_4 \qquad \text{for } x \in \mathbb{R}^d, \tag{5}$$

with a Laplacian kernel and uniform sampling on a hypercube $\{\pm 1\}^d$. Such staircase-like functions have recently been used as a model for hierarchical structures in statistical and machine learning [1, 2]. Figure 2a depicts the generalization error of kernel regression (blue), one-step `IRKM`$(\frac{1}{2})$ (green), multi-step `IRKM`$(\frac{1}{2})$ (red), and neural networks (orange) trained by Adam when learning the function (5). The superior performance of `IRKM`$(\frac{1}{2})$ is striking; how can we understand this phenomenon? First, observe that in the sub-linear sampling regime $n = d^\delta$ with $\delta \in (0,1)$, ridge regression can only identify the important coordinates for the first-order truncation $f_{\leq 1}$, namely x_1 and x_2. This is supported by Figure 2d, which depicts the empirical coordinate weights using the KRR estimator. On the other hand, looking at Figures 2c and 2e, we see that `IRKM`$(\frac{1}{2})$ quickly identifies the planted coordinates x_3 and x_4 as being significant, all still within the sub-linear sampling regime.

We now explain this phenomenon. To this end, recall that an orthonormal basis for $L_2(\mu)$ is given by Fourier-Walsh monomials:

$$\phi_S(x) = \prod_{i \in S} x_i \qquad \forall S \subset [d].$$

Following [1], we say that a function $f = \sum_{S \subset [d]} b_S \phi_S$ has *leap complexity* k if the nonzero coefficients $\{b_S\}$ can be ordered in such a way that the corresponding sets $\{S_i\}_{i=1}^m$ satisfy

$$|S_1| \leq k \qquad \text{and} \qquad \left| S_i \setminus \bigcup_{1 \leq j < i} S_j \right| \leq k,$$

and k is the smallest integer for which this is possible. In other words, f has leap complexity k if k is the smallest integer such that its nonzero coefficients can be ordered in such a way that the corresponding sets grow by at most k indices. In particular, the function (5) has leap complexity $k = 1$. We use $\mathrm{Leap}(f)$ to denote the leap complexity of f. The following theorem shows that $n = d^{\mathrm{Leap}(f)-1+\delta}$ samples suffices for `IRKM`(α) with $\alpha \in (0,1)$ to learn any function f, where $\delta \in (0,1)$ is arbitrary.

Theorem 1.4 (Informal). *Consider a target function $f(x) = g(x_1, \ldots, x_r)$ that is a degree ℓ polynomial on $\mathbb{R}^d$ with $r = O_d(d^s)$ influential coordinates for some $s \in [0,1)$, meaning that the coordinate weights $\mathbb{E}[\partial_j f]^2$ are uniformly bounded away from zero for all $j = 1, \ldots, r$ as $d \to \infty$. Suppose $\mathrm{Leap}(f) \geq 1$ and consider the regime $n = d^{\mathrm{Leap}(f)-1+\delta}$ with $\delta \in (0,1)$. Suppose, moreover, that the sparsity level satisfies $s \leq \delta/(\ell+1)$. Then for any mixing parameter $\alpha \in (0,1)$ in step (9) and under an appropriate setting of the safeguard ϵ_{s} in steps (6) and (7), there exists a horizon $T = O_d(1)$ such that the predictor $\hat{f}_T$ returned by* `IRKM`(α) *satisfies:*

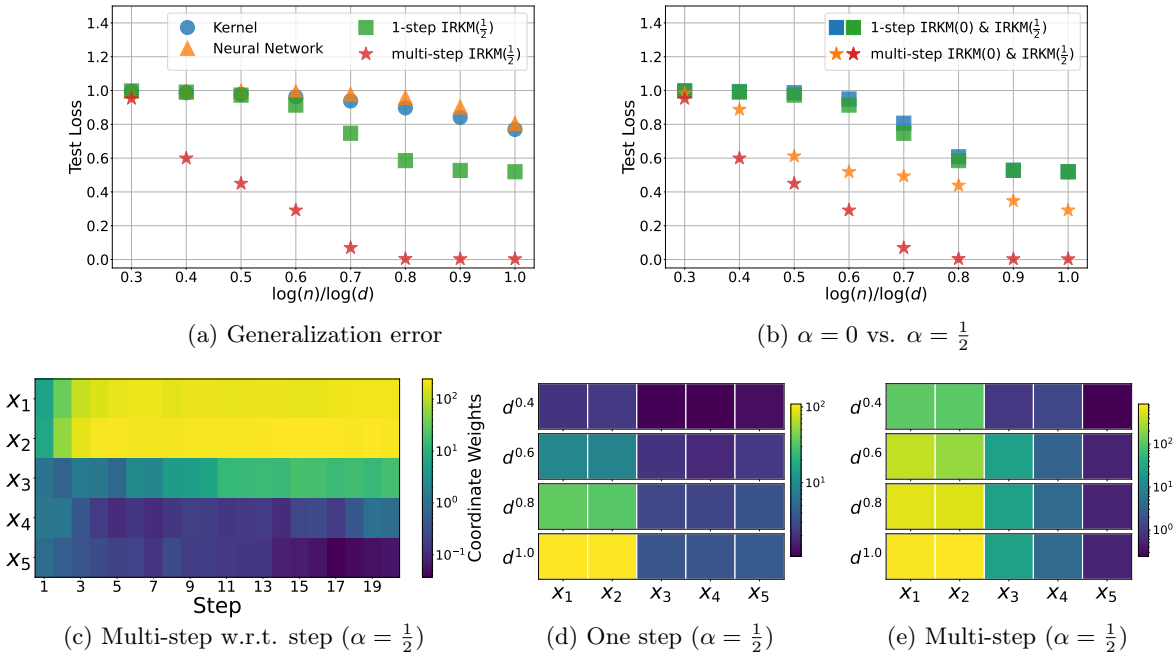$$\left\| \hat{f}_T - f \right\|_{L_2(\mu)}^2 = o_{d,\mathbb{P}}(1).$$

Figure 2: **(a) Test loss for kernel machines, neural networks and** $\mathtt{IRKM}(\frac{1}{2})$**. (b) Comparison of** $\mathtt{IRKM}(\alpha)$ **performance for** $\alpha = 0$ **versus** $\alpha = \frac{1}{2}$**. (c)The empirical coordinate weights of** $\mathtt{IRKM}(\frac{1}{2})$ **at step** $1 - 20$ **in the regime** $n = d^{0.6}$**. (d,e) The empirical coordinate weights of** $\mathtt{IRKM}(\alpha)$ **at step** T **for** $\alpha = 0$ **and** $\frac{1}{2}$ **respectively.** The samples are i.i.d. uniformly drawn from the hypercube $\{-1, 1\}^d$ with $d = 500$ and the label is $y = x_1 + x_2 + x_1 x_2 x_3 + x_1 x_2 x_3 x_4 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. We use Laplacian kernel for both kernel machines and $\mathtt{IRKM}(\alpha)$. The $\mathtt{IRKM}(\alpha)$ algorithm is iterated for at most $T = 20$ steps. We train a fully-connected network by Adam, varying the widths and batch sizes as $\{128, 256, 512\}$, depth as $\{2, 3, 4, 5, 6\}$ and learning rate as $\{10^{-3}, 10^{-4}\}$ and report the smallest test loss. For all experiments, we report an average of 10 runs.

Applying Theorem 1.4 to the example (5) with $s = 0$, we see that $\mathtt{IRKM}(\frac{1}{2})$ in a constant number of steps produces a predictor $\hat{f}_T$ that is asymptotically consistent for estimating f as d tends to infinity. Theorem 1.4 is broader still, as it allows the sparsity level $r = d^s$ of the target function to grow with d.

It is important to compare Theorem 1.4 to the existing results in the literature [1, 2], which analyzed the performance of neural networks trained using (projected) SGD when learning polynomials with bounded leap complexity $\mathrm{Leap}(f^*)$ from Gaussian or hypercube data. In particular, the paper [1] showed that $\tilde{\Theta}_d(d^{\max(\mathrm{Leap}(f^*)-1,1)})$ projected SGD steps suffice for a two-layer neural network to learn a sparse polynomial f^*. In this work, we consider learning with kernel machines and show that $\Theta_d(d^{\mathrm{Leap}(f^*)-1})$ samples suffices to learn f^* after solving finitely many $n \times n$ kernel linear systems. Note that the sample complexity of $\mathtt{IRKM}(\frac{1}{2})$ is better than neural networks by a factor of d when learning leap-one functions and matches neural networks for functions with leap complexity $\mathrm{Leap}(f^*) > 1$. In particular, Figure 1 indeed shows that when trained with ADAM on the leap one function, we see that ADAM performs comparably to kernels until $n = \Omega(d)$. We note that in all of our experiments, $\mathtt{IRKM}(\frac{1}{2})$ typically outperforms neural networks and at worst performs comparably.

Iteratively reweighted least squares and recursive feature machines. $\mathtt{IRKM}(\alpha)$ is closely related to existing algorithms in the literature when the mixing coefficient is $\alpha = 0$. First, if the kernel is linear and the resampling (4) and normalization (8) steps are omitted, then $\mathtt{IRKM}(0)$ reduces exactly to the celebrated Iteratively Re-weighted Least-Squares (IRLS) algorithm [18, 29] for recovering a sparse vector

Initialize: $M = I_d$ and fix a safeguard parameter $\epsilon_s > 0$ and mixing parameter $\alpha \in [0, 1]$.

for $t = 1, \ldots, T$ **do**:

    Sample fresh data $\{(x^{(i)}, y_i)\}_{i=1}^n$ and define the reweighted kernel $K_M(x, z) = K(\sqrt{M}x, \sqrt{M}z)$.

    Obtain predictor $\hat{f}_t(x)$ from Kernel Ridge Regression on the data $\{(x^{(i)}, y_i)\}_{i=1}^n$ with the kernel K_M.

    Update the first weight matrix $M^{[1]} = \epsilon_s I_d + \frac{1}{n} \sum_{i=1}^n \nabla_x \hat{f}_t(x^{(i)}) \nabla_x \hat{f}_t(x^{(i)})^\top$.

    Update the second weight matrix $M^{[2]} = \epsilon_s I_d + \frac{1}{n} M^{\frac{1}{2}} \mathcal{D}(M) M^{\frac{1}{2}}$

    Normalize the eigenvalues $M^{[k]} = \frac{d}{\text{tr}(M^{[k]})} \cdot M^{[k]}, \quad k = 1, 2.$

    Update the weight matrix $M = (1 - \alpha) \cdot M^{[1]} + \alpha \cdot M^{[2]}$.

end for

$z_\sharp \in \mathbb{R}^d$ from finitely many linear measurements $Az_\sharp = b$. In this way, IRKM($\frac{1}{2}$) can be understood as a generalization of IRLS for learning sparse nonlinear functions. We also note that extensions of IRLS for low-rank matrix recovery, where the weights being updated are themselves matrices rather than vectors, have been studied in [36, 54].
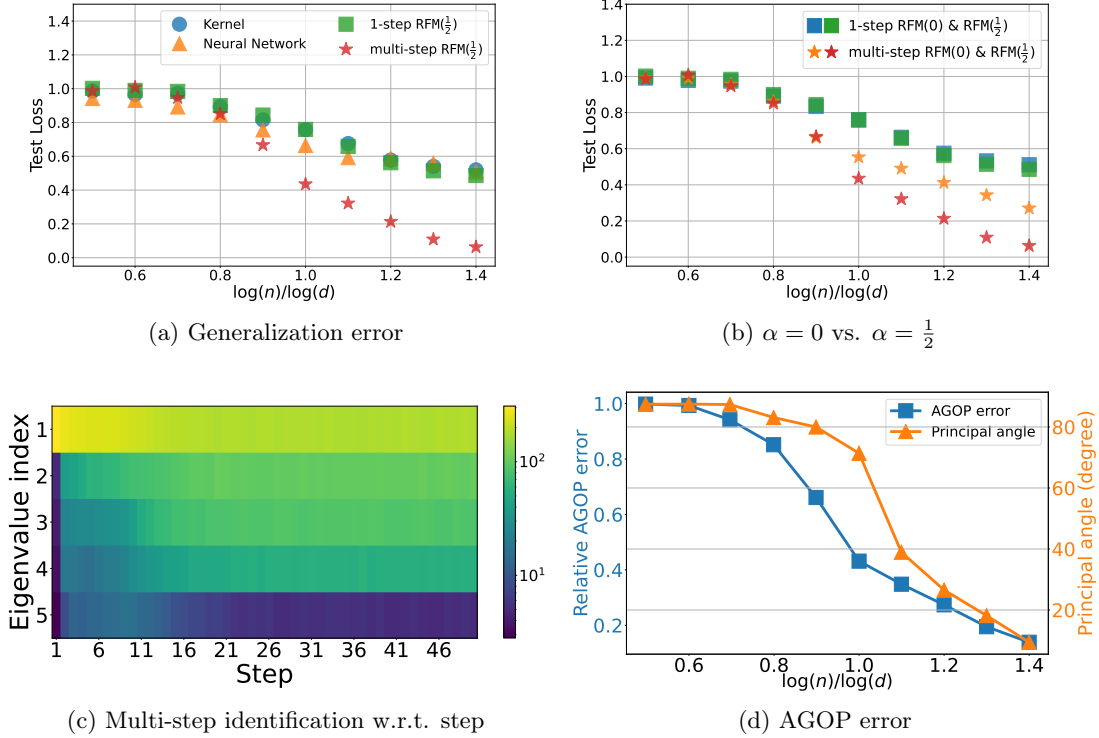

(a) Generalization error

(b) $\alpha = 0$ vs. $\alpha = \frac{1}{2}$

(c) Multi-step identification w.r.t. step

(d) AGOP error

Figure 3: **(a) Test loss for kernel machines, neural networks and** RFM($\frac{1}{2}$) **(Algo. 2). (b) Comparison of** RFM(α) **performance for** $\alpha = 0$ **versus** $\alpha = \frac{1}{2}$. **(c) Eigenvalues of AGOP at step** $1 - 50$ **in the regime** $n = d^{1.4}$. **(d) Relative AGOP error and principal angle between the top-4 eigenspace of AGOPs.** The samples are i.i.d. uniformly drawn from $\{-1, 1\}^d$ with $d = 500$ and label is $y = z_1 + z_2 + z_1 z_2 z_3 + z_1 z_2 z_3 z_4 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.1^2)$ and $z = Ux$ where U is a random rotation matrix. We use the Laplacian kernel for both kernel machines and RFM(α). We train a fully-connected network by Adam, varying the widths and batch sizes as $\{128, 256, 512\}$, depth as $\{2, 3, 4, 5, 6\}$ and learning rate as $\{10^{-3}, 10^{-4}\}$ and report the smallest test loss. For all experiments, we report an average of 10 runs.

IRKM(0) is also closely related to the Recursive Feature Machines (RFM) algorithm introduced in [60] for learning multi-index models. Indeed, if the resampling (4) and normalization (8) steps are omitted and one sets $\epsilon_s = 0$, then IRKM(0) is a variant of RFM [60] where the Average Gradient Outer Product (AGOP) $M = \frac{1}{n} \sum_i \nabla \hat{f}(x^{(i)}) \nabla \hat{f}(x^{(i)})^\top$ is replaced by its diagonal. We stress, however, that the fresh resampling (4) and normalization steps (8), and safeguarding in (6) with $\epsilon_s > 0$ are essential both for our theoretical results to be valid and in our experiments when learning hierarchical functions. Moreover, IRKM(α) suggests a modification of RFM for learning multi-index models (i.e., sparse functions in an unknown basis) online, which draws fresh samples and reweights the data by an average of the AGOP and the DN-estimator, which is naturally defined as

$$\mathcal{D}(M)_{i,j} := \beta_M^\top \frac{\partial K_M(X,X)}{\partial M_{i,j}} \beta_M, \quad \forall i,j \in [d]. \tag{6}$$

We record this modified algorithm as Algorithm 2, and call it RFM(α). We conjecture that many of our theoretical results should extend from learning sparse functions with IRKM(α) to learning multi-index models with RFM(α). As an illustration, we apply RFM(α) for learning the function (5), precomposed by a random rotation. The results are recorded in Figure 3, where we clearly see that the unknown basis is gradually learned by the algorithm, albeit with higher sample complexity and a larger number of steps (20 vs. 50). We measure the learning of the basis using the relative AGOP error and the principal angle against the ground truth AGOP, namely $\mathbb{E}_x \left[\nabla f(x) \nabla f(x)^\top \right]$. As in the vector case, we observe in Figure 3b the superior performance of RFM($\frac{1}{2}$) over RFM(0).

1.2 Further related works

Kernel learning in the polynomial scaling regime. Kernel Ridge Regression is a classical technique in machine learning. Generalization guarantees for kernel ridge regression are by now classical in the regime where the dimension d of the feature vectors is fixed and the number of samples n can grow. See for example the monographs [11, 26, 65] on the subject. More recently, the high-dimensional regime, where d and n can grow simultaneously, has gained much attention. We refer the reader to the representative articles on the subject [13, 16, 20, 33, 51, 53, 59]. The most relevant setting for us is the polynomial scaling regime $n = d^{p+\delta}$ for some $p \in \mathbb{N}$ and $\delta \in (0,1)$. In this regime, it is known that kernel methods can learn (at most) a degree-p polynomial when the data is uniformly distributed on the sphere [39] and the hypercube [53]. Moreover, the subsequent work [30] established analogous upper and lower bounds on the generalization error for a broader class of data distributions and kernels, including radial basis function (RBF) kernels.

Feature learning with neural networks and kernel machines. The learning capacity of infinitely wide neural networks under specific parameterizations is equivalent to that of kernel machines with the Neural Tangent Kernel (NTK) [42]. However, neural networks empirically exhibit superior generalization performance compared to kernel machines, a phenomenon attributed to their ability to implicitly learn low-dimensional features which suffice for accurate prediction. For instance, a series of works has demonstrated a clear separation in generalization performance between neural networks and NTKs [4, 23, 50]. To better understand the feature learning phenomenon, a significant line of research focuses on single- and multi-index models, where the target function takes the form $f^*(x) = g(Ux)$ with $U : \mathbb{R}^d \to \mathbb{R}^r$ and $d \gg r$. It has been shown that neural networks benefit from a small intrinsic dimension r, achieving improved sample complexity, whereas kernel methods do not exhibit similar gains [9, 12, 15, 27, 28, 55].

Recently, kernel-based algorithms that explicitly learn features have gained some attention, with prominent examples including [37, 49, 60, 72]. The most relevant work for our paper is the Recursive Feature Machines algorithm, introduced in [60], which was experimentally shown to efficiently learn multi-index models and achieve impressive performance on tabular datasets and matrix completion tasks [61], among others. RFM learns the features through iteratively learning the so-called Average Gradient Outer Product, a matrix that captures the relevant low-dimensional subspace of the true function. The (implicit) AGOP learning mechanism has also been observed in deep ReLU Neural Networks with linear activation layers [58]. The paper [24] provided theoretical guarantees demonstrating that AGOP learning is inherent

in training two-layer models, where the dimensionality reduction happens in the first linear layer weights. In a parallel line of work, [35] introduces a model combining kernel methods and neural networks to promote feature learning. We note also that the paper [7] proposed a similar normalization scheme for AGOP as ours in the implementation of RFM for better bandwidth selection.

Another interesting line of work studies kernel-based feature selection in nonparametric statistics [43, 63]. In particular, [63] shows that by explicitly penalizing the L_1 norm of the coordinate weights, kernel machines can recover influential coordinates and the hierarchical interactions up to leap complexity one.

Learning with orthogonal polynomials. A recent line of work focuses on learning sparse functions that depend only on a small number of basis elements in the L^2-space induced by the data distribution and analyzes the learning power of neural networks. Note that sparse functions can be viewed as a special case of multi-index models. Neural networks, unlike classical methods, do not require prior knowledge of the functional basis, and were shown to be more sample-efficient than linear methods in a variety of settings; some representative examples of this line of work are [10, 12, 21, 27, 38, 44, 45, 48, 64]. Additionally, the recent work [67] showed that Classification and Regression Trees (CART) can learn leap-1 functions with $O_d(\log(d))$ samples.

1.3 Outline

The outline of the paper is as follows. Section 2 sets forth basic notation that we will use. Section 3 describes the framework of orthogonal polynomials, with a particular focus on Hermite and Fourier-Walsh polynomials. Section 4 develops the main results of the paper. The final Section 6 provides further experimental evidence that illustrates our theoretical results.

2 Notation

We denote the set of natural numbers by $\mathbb{N}$ and let $\mathbb{N}^+$ denote the set $\mathbb{N}/\{0\}$. For a positive integer n, the symbol $[n]$ will stand for the set $\{1, 2, \ldots, n\}$. A vector $\lambda \in \mathbb{N}^d$ with nonnegative entries is called a multi-index and we denote its modulus by $|\lambda| = \lambda_1 + \lambda_2 + \ldots + \lambda_d$. For any vector $x \in \mathbb{R}^d$ and a multi-index $\lambda \in \mathbb{N}^d$, we define the Fourier-Walsh polynomial $x^\lambda = x_1^{\lambda_1} x_2^{\lambda_2} \ldots x_d^{\lambda_d}$. Throughout the paper, given a sequence of vectors $v_i^{(k)} \in \mathbb{R}^d$, the subscript i denotes the coordinate index of a single vector while the superscript (k) denotes the index of the vector in the sequence.

For any vectors $u, v \in \mathbb{R}^d$, we use the symbol $\langle u, v \rangle = \sum_{i=1}^d u_i v_i$ to denote their inner product and we denote the ℓ_p-norm of v by $\|v\|_p = (\sum_{i=1}^d |v_i|^p)^{1/p}$. For any matrix M, we denote its ij'th entry by $M_{i,j}$, its i'th row by $M_{[i,:]}$, and its j'th column by $M_{[:,j]}$. We write $\mathsf{diag}(M)$ for the vector of diagonal entries of M, and for a vector v, the symbol $\mathsf{Diag}(v)$ denotes the diagonal matrix whose diagonal entries are the elements of v. The symbols $\|M\|_{\mathrm{op}}$ and $\|M\|_F$ will stand for the spectral and Frobenius norms of a matrix M, respectively. We use the Hadamard product, denoted by $\odot$, to represent element-wise multiplication between vectors or matrices, meaning $(A \odot B)_{i,j} = A_{i,j} B_{i,j}$.

For any function $f, g \colon \mathbb{R}^d \to \mathbb{R}$, the symbol $f \lesssim g$ means that there exists a constant $C > 0$ satisfying $f(x) \leq Cg(x)$ for all x. The symbol $f \gtrsim g$ is defined similarly. We write $f \asymp g$ to indicate that f and g are of the same order, meaning both $f \lesssim g$ and $f \gtrsim g$ hold. For any functions $f, g \colon \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}$, the asymptotic notation $f = O_d(g)$ means that there exists a constant $C > 0$ and $\bar{d} \in \mathbb{N}$ such that the inequality $|f(x, d)| \leq Cg(x, d)$ holds for all x and all $d \geq \bar{d}$. Similarly, notation $f = o_d(g)$ means that for any $\epsilon > 0$ there exists $d_\epsilon \in \mathbb{N}$ such that the estimate $|f(x, d)| \leq \epsilon g(x, d)$ holds for all x and all $d \geq d_\epsilon$.

Finally, we use $O_{d,\mathbb{P}}(\cdot)$ to denote the big-O in probability relation and similarly $o_{d,\mathbb{P}}(\cdot)$ for the little-o in probability relation. Namely, for two sequences of random variables f_d and g_d, the equation $f_d = O_{d,\mathbb{P}}(g_d)$ means that for any $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ satisfying $\limsup_{d \to \infty} \mathbb{P}(|f_d| > C_\epsilon g_d) \leq \epsilon$. Similarly, the notation $f_d = o_{d,\mathbb{P}}(g_d)$ means that for every $\epsilon > 0$ we have $\lim_{d \to \infty} \mathbb{P}(|f_d| > \epsilon g_d) = 0$.

3 Preliminaries on orthogonal polynomials

In this section, we record notation and preliminaries on orthogonal polynomials, following the standard monographs on the subject [6, 25, 66]. Consider a measure space $(D, \mathcal{A}, \mu)$, where D is a set, $\mathcal{A}$ is a σ-algebra, and μ is a measure on $(D, \mathcal{A})$. We let $L^2(D, \mu)$ denote the Hilbert space of μ square-integrable functions on D equipped with the usual inner product and the induced norm

$$\langle f, g \rangle = \int fg \, d\mu \qquad \|f\| = \sqrt{\langle f, f \rangle}.$$

We will suppress the symbol D from $L^2(D, \mu)$ when it is clear from context. Every Hilbert space $L^2(\mu)$ admits an orthonormal basis $\{\phi_j\}_{j \in J}$, meaning ϕ_j are unit norm, pairwise orthogonal, and their linear span is dense in $L^2(\mu)$. A favorable situation occurs when J is countable, in which case $L^2(\mu)$ is called separable. Any function $f \in L^2(\mu)$ in a separable Hilbert space can be expanded in the orthogonal basis:

$$\left\| f - \sum_{i=0}^{m} \langle f, \phi_i \rangle \phi_i \right\| \to 0 \quad \text{as} \quad m \nearrow |J|.$$

Here, we identify J with the contiguous subset of natural numbers $\mathbb{N}$ starting at zero. We will primarily focus on the following two examples of separable Hilbert spaces along with orthonormal bases.

3.1 Fourier expansion on the boolean Hypercube

As the first example, let $\mathbb{H}^d = \{-1, 1\}^d$ be the hypercube equipped with the uniform measure τ_d. Then the inner product between any two functions $f, g \colon \mathbb{H}^d \to \mathbb{R}$ is simply

$$\langle f, g \rangle = \frac{1}{2^d} \sum_{x \in \mathbb{H}^d} f(x) g(x).$$

An orthonormal basis on $L_2(\mathbb{H}^d)$ is furnished by the multi-linear monomials (called Fourier-Walsh)

$$x^\alpha := \prod_{i=1}^{d} x_i^{\alpha_i} \qquad \text{for each } \alpha \in \{0, 1\}^d.$$

We will denote the degree of the monomial x^α by the symbol $|\alpha| := \sum_{i=1}^{d} \alpha_i$. Notice that there is a one-to-one correspondence between binary vectors $\alpha \in \{0, 1\}^d$ and subsets $S \subset [d]$ (their support). We will therefore often abuse notation and treat α as both a vector and a set whenever convenient.

Thus any polynomial f on the hypercube $\mathbb{H}^d$ can be expanded in the monomial basis:

$$f(x) = \sum_{\alpha \in \{0,1\}^d} b_\alpha x^\alpha \qquad \text{with} \qquad b_\alpha = \langle f, x^\alpha \rangle,$$

where b_α is called the Fourier coefficient of f indexed by α. The p-truncation of f is then defined to be the truncated series

$$f_{\leq p}(x) = \sum_{\alpha \in \{0,1\}^d \colon |\alpha| \leq p} b_\alpha x^\alpha. \tag{7}$$

Equivalently, $f_{\leq p}$ is the projection of f in $L_2(\mathbb{H}^d)$ onto the span of all Fourier-Walsh monomials x^α with $|\alpha| \leq p$. The functions f_p and $f_{>p}$ are defined in an obvious way.

The Fourier coefficients b_α have a convenient interpretation in terms of discrete derivative of f. Namely, the discrete derivative of any function $g \colon \mathbb{H}^d \to \mathbb{R}$ in direction x_i is defined to be

$$\mathsf{D}_i g(x) = \frac{g(x) - g(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_d)}{2 x_i}.$$

The discrete derivative with respect to a set $S = \{i_1, \ldots, i_k\} \subset [d]$ is then defined by iterating:

$$\mathtt{D}_S f(x) = \mathtt{D}_{i_1} \mathtt{D}_{i_2} \ldots \mathtt{D}_{i_k} f(x).$$

Interestingly, Fourier coefficients correspond precisely to expectations of discrete derivatives:

$$b_S = \mathbb{E}[\mathtt{D}_S f]. \tag{8}$$

See for example [56, Section 2.2] for details. Therefore, Fourier coefficients measure the sensitivity of f to coordinate perturbations.

3.2 Hermite expansion in Gauss space

Let γ be the standard Gaussian measure on $\mathbb{R}$. Then the Hilbert space $L_2(\mathbb{R}, \gamma)$ is separable. One convenient orthonormal basis $\{h_k\}_{k \geq 0}$ is obtained by performing the Gram-Schmidt process on the standard monomial basis $\{1, x, x^2, \ldots\}$. The elements of this orthonormal basis are called Hermite polynomials. They can be explicitly described in many different ways. For example, they arise as the higher-order derivatives of the Gaussian density[5]:

$$h_k(x) = \frac{(-1)^k}{\sqrt{k!}} \cdot e^{\frac{x^2}{2}} \cdot p^{(k)}(x) \qquad \text{where} \ \ p(x) := e^{-\frac{x^2}{2}}. \tag{9}$$

The first few Hermite polynomials are

$$h_0(x) = 1, \ \ h_1(x) = x, \ \ h_2(x) = \tfrac{1}{\sqrt{2!}}(x^2 - 1), \ \ h_3(x) = \tfrac{1}{\sqrt{3!}}(x^3 - 3x).$$

Multivariate Hermite polynomials are similarly defined as products of univariate Hermite polynomials along coordinates. Namely, the Hermite polynomial h_α with respect to a multi-index $\alpha \in \mathbb{N}^d$ is defined by

$$h_\alpha(x) = \prod_{i=1}^d h_{\alpha_i}(x_i),$$

where h_{α_i} is the α_i-th univariate Hermite polynomial. The degree of h_α as a polynomial on $\mathbb{R}^d$ is simply the order of the multi-index $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$. Hermite polynomials $\{h_\alpha\}_{\alpha \in \mathbb{N}^d}$ form an orthonormal basis for $L_2(\mathbb{R}^d, \gamma_d)$ where γ_d is the standard Gaussian measure on $\mathbb{R}^d$. Thus any square-integrable function $f \in L_2(\mathbb{R}^d, \gamma_d)$ can be expanded in the Hermite basis:

$$f(x) = \sum_{\alpha \in \mathbb{N}^d} b_\lambda h_\alpha(x) \qquad \text{with} \qquad b_\alpha = \langle f, x^\alpha \rangle,$$

where the quantity b_α is called the Hermite coefficient of f indexed by α. The p-truncation of $f \in L_2(\mathbb{R}^d, \gamma_d)$ is then defined to be the truncated series

$$f_{\leq p} = \sum_{\alpha \in \mathbb{N}^d : \|\alpha\|_1 \leq p} b_\alpha h_\alpha. \tag{10}$$

Equivalently, $f_{\leq p}$ is the projection of f in $L_2(\mathbb{R}^d, \gamma_d)$ onto the span of all Hermite polynomials h_α of degree at most p. The functions f_p and $f_{>p}$ are defined in an obvious way.

Similarly to (8), the Hermite coefficients are closely related to partial derivatives by the formula:

$$b_\alpha = \frac{1}{\sqrt{\alpha!}} \cdot \mathbb{E}\, \partial^{(\alpha)} f, \tag{11}$$

where we use the shorthand notation $\alpha! = \prod_{i=1}^d \alpha_i!$ and $\partial^{(\alpha)}$ is the iterated partial derivative.

---

[5]Note that compared to the "probabilist's Hermite polynomials", we have an extra scaling factor $1/\sqrt{k!}$ to ensure the unit length of the bases elements.

Consider now $\mathbb{R}^d$ equipped with some probability measure μ_d. We will often have to control the L_q-norm $\|f\|_{L_q(\mu_d)} = [\mathbb{E}|f|^q]^{1/q}$ of a polynomial f on $\mathbb{R}^d$. In general, one would expect that the ratio between the $L_q(\mu_d)$ and $L_2(\mu_d)$ norms depends strongly on the dimension d. Interestingly, for a broad class of measures μ and functions f this is not the case. Indeed, the so-called hypercontractivity inequality ensures that any polynomial $f \colon \mathbb{R}^d \to \mathbb{R}$ of degree at most ℓ satisfies the inequality

$$\|f\|_{L_q(\mu_d)} \le (q-1)^{\ell/2} \, \|f\|_{L_2(\mu_d)} \qquad \forall q \ge 2, \tag{12}$$

where μ_d can be the standard Gaussian measure [47, Chapter 3.2] on $\mathbb{R}^d$ or the uniform measure on the hypercube $\mathbb{H}^d$ [56, Chapter 9]. The inequality (12) is widely used in probability and theoretical computer science, and we will use it heavily here as well.

4 Main results

In this section, we formally state the main results of the paper. To this end, fix a measure space $(D, \mathcal{A}, \mu)$, where $D \subset \mathbb{R}^d$ is a set, $\mathcal{A}$ is a σ-algebra, and μ is a measure on $(D, \mathcal{A})$. The two main examples for us will be the uniform measure τ_d on the Hypercube $\mathbb{H}^d = \{-1, 1\}^d$ and the Gaussian measure γ_d on $\mathbb{R}^d$. In both of these cases, the symbol $f_{\le p}$ will refer to the truncated functions (7) and (10), respectively. Fix now a function $f \colon D \to \mathbb{R}$ with finite second moment $\mathbb{E}_\mu f^2 < \infty$ and a data set $\{(x^{(i)}, y_i)\}_{i=1}^n$, where the points $x^{(1)}, \ldots, x^{(n)}$ are sampled independently from μ and are labeled by the function values

$$y_i = f^*(x^{(i)}) + \varepsilon_i,$$

with noise $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ that is independent of x_i. Note that since f^* is defined on $D \subset \mathbb{R}^d$ and we will be interested in asymptotics as d tends to infinity, it would be more accurate to use the notation f_d^* for the sequence of ground truth functions. Although we omit the subscript d to simplify notation, the reader should nonetheless keep in mind the implicit dependence of f^* on the dimension d throughout.

Our goal is to construct a function $\hat{f}$ (the predictor) from the sampled data that approximates the target function f^* in the L_2 sense. We construct a predictor $\hat{f}$ based on Kernel Ridge Regression, henceforth abbreviated as KRR. Throughout the paper we focus on inner product kernels $K \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ having the form

$$K(x, y) = g\left(\frac{\langle x, y \rangle}{d}\right), \tag{13}$$

for all x and y in the support of μ and where $g \colon \mathbb{R} \to \mathbb{R}$ is some function. We then define the following kernel induced by the derivative of g:

$$K'(x, z) := g'\left(\frac{\langle x, z \rangle}{d}\right).$$

Our main results will invoke a combination of the following two regularity assumptions on g.

Assumption 1 (Analytic). *There exists $\varepsilon \in (0, 1)$ such that the function g in (13) is analytic on $(-\varepsilon, \varepsilon)$ and its derivatives satisfy $g^{(k)}(0) \ge 0$ for all $k \ge 0$.*

Assumption 2 (Lipschitz). *There exists $\varepsilon \in (0, 1)$ such that the function g in (13) has Lipschitz continuous derivative g' on $(1 - \varepsilon, 1 + \varepsilon)$.*

Classical results [65, Theorem 13.4] show that non-negativity of derivatives in Assumption 1 implies that the kernels K and K' are positive semidefinite. Inner-product kernels satisfying Assumptions 1 and 2 appear prominently in applications and include for example the polynomial kernel $K(x, y) = (b + \langle x, y \rangle)^m$ and the exponential kernel $K(x, y) = \exp(\alpha \langle x, y \rangle)$ for any $\alpha > 0$. When the vectors in the support of μ have constant norm, as is the case for spherical or hypercube data, the Gaussian kernel $K(x, y) = \exp\left(-\|x - y\|^2 / 2\sigma^2\right)$ satisfies Assumptions 1 and 2, while the the Laplace kernel $K(x, y) = \exp(-\|x - y\|/\sigma)$ satisfies Assumption 1 but not 2.

Given a set of data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ we form the data matrix $X \in \mathbb{R}^{n \times d}$ by stacking $x^{(i)}$ as its rows and define the $n \times n$ similarity matrix $K(X, X) = \{K(x^{(i)}, x^{(j)})\}_{i,j=1}^n$. The kernel ridge regression (KRR) prediction function with regularization parameter $\lambda > 0$ is then given by the expression

$$\hat{f}(z) = K(z, X)\beta \qquad \text{where} \qquad \beta := (K(X, X) + \lambda I_n)^{-1} y, \tag{14}$$

where we define the vector $K(z, X) := [K(z, x^{(1)}), \ldots, K(z, x^{(n)})] \in \mathbb{R}^{1 \times n}$. Throughout, we assume that the regularization parameter satisfies $\lambda = O_d(1)$. Much of our paper focuses on analyzing the partial derivative of the prediction function $\partial_r \hat{f}(x)$, which takes the simple form

$$\partial_r \hat{f}(z) = \frac{1}{d} K'(z, X) X_r \beta, \tag{15}$$

Here, ∂_r denotes the partial derivative with respect to x_r and we define the matrix $X_r := \mathsf{Diag}[x_r^{(1)}, \ldots, x_r^{(n)}]$.

We require one last notational ingredient before stating our results. Namely, all theorems in this section impose a mild nondegeneracy condition of the form:

$$\min_{k=0,\ldots,q} g^{(k)}(0) \text{ and } \lambda + \sum_{k=q+1}^{\infty} g^{(k)}(0) \text{ are strictly positive.} \tag{16}$$

Here, the constant $q \in \mathbb{N}$ will be specified each time that (16) is invoked in theorem statements. In particular, note that if all the derivatives $g^{(k)}(0)$ are strictly positive, then (16) holds even with $\lambda = 0$.

4.1 Results for Gaussian data

Our first main theorem shows that for Gaussian data in the regime $n = d^{p+\delta}$ with $\delta \in (0, 1/2)$ and $p \in \{1, 2\}$, the empirical weights $\mathbb{E}_n(\partial_r \hat{f})^2 := \frac{1}{n} \sum_{i=1}^n (\partial_r \hat{f}(x^{(i)}))^2$ are asymptotically consistent for estimating the weight $\mathbb{E}(\partial_r f_{\leq p}^*)^2$ of the truncated function $f_{\leq p}$. This theorem can be understood as a higher-order analogue of the bound on generalization error established in [39, 53]. Although we only prove Theorem 4.5 in the setting $p \in \{1, 2\}$, we conjecture that the analogous result holds for all $p \in \mathbb{N}$. The proof of the theorem appears in Appendix B.1.

Theorem 4.5 (Gradient consistency with Gaussian data). *Suppose that Assumptions 1 and 2 hold. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0, 1/2)$ is a constant and $p \in \{1, 2\}$. Suppose moreover that condition (16) holds with $q = p$. Then for any $\epsilon > 0$ and $\eta > 0$, the KRR predictor $\hat{f}$ satisfies:*

$$\max_{r=1,\ldots,d} \left| \mathbb{E}_n(\partial_r \hat{f})^2 - \mathbb{E}(\partial_r f_{\leq p}^*)^2 \right| = O_{d,\mathbb{P}} \left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon} \right) \cdot \left(\|f^*\|_{L_2}^2 + \|f_{>2}^*\|_{L_{2+\eta}}^2 + \sigma_\varepsilon^2 \right). \tag{17}$$

A few comments are in order. First, the bound (17) depends on the target function f^* only through the moments, $\|f^*\|_{L_2}^2$ and $\|f_{>2}^*\|_{L_{2+\eta}}^2$, where $\eta > 0$ is arbitrary. In the particular case when f^* is a finite degree polynomial with a bounded second moment $\|f^*\|_{L_2}^2 < \kappa$ uniformly in dimension, then the higher-order moment $\|f_{>2}^*\|_{L_{2+\eta}}^2$ is automatically uniformly bounded due to the Hypercontractivity inequality (12). Consequently, in this case the right side of (17) tends to zero in probability as d tends to infinity. The assumptions on the derivatives of g are mild and hold for example if K is an exponential kernel or a polynomial kernel with degree $m \geq 3$. In both of these examples, one can even set $\lambda = 0$.

It is worthwhile to note that the quantities $\mathbb{E}(\partial_r f_{\leq p}^*)^2$ have an intuitive variational interpretation in terms of the derivatives of f^* itself. Namely, suppose that we may write $f^* = \sum_{\alpha \in \mathbb{N}^d} b_\alpha h_\alpha$, where h_α are the Hermite polynomials. Then for $p = 1$, we have the expression

$$\mathbb{E}(\partial_r f_{\leq 1}^*)^2 = b_{e_r}^2 = (\mathbb{E} f^*(x) x_r)^2 = (\mathbb{E} \partial_r f^*)^2,$$

14

where the last equality follows from (11). Similarly, for the case $p = 2$ simple arithmetic shows the expression

$$\mathbb{E}\big(\partial_r f_{\leq 2}^*\big)^2 = \mathbb{E}(b_{e_r} + \sqrt{2}b_{2e_r}x_r + \sum_{i \neq r} b_{e_i+e_r}x_i)^2$$

$$= b_{e_r}^2 + 2\alpha_{2e_r}^2 + \sum_{j \neq r} b_{e_j+e_r}^2,$$

$$= (\mathbb{E}\partial_r f^*)^2 + \sum_{j \in [d]} (\mathbb{E}\partial^{(rj)}f^*)^2,$$

where again the last equality follows from (11). Thus in both cases, the quantity $\mathbb{E}\big(\partial_r f_{\leq p}^*\big)^2$ measures the total magnitude of the expected derivatives of the target function f^* with respect to x_r.

4.2 Results for hypercube data

For the rest of the section, we focus entirely on learning polynomials from data that is sampled uniformly over the hypercube. We formalize this setting in the following assumption.

Assumption 3 (Hypercube data). *The measure μ is the uniform measure on the hypercube $\{-1, 1\}^d$ and the underlying true function $f^* = \sum_{S \subset [d]} b_S \phi_S$ is a polynomial of constant degree ℓ.*

We will be interested in learning functions on $\mathbb{R}^d$ that only depend on a small portion of all coordinates. With this in mind, we introduce the following definition of "influential coordinates". Conceptually, x_r is an influential coordinate of g if variations in x_r lead to nontrivial variations in g.

Definition 1 (Influential coordinates). *For a function $g \in L_2(\mu)$, a coordinate x_r is said to be influential if the inequality $\mathbb{E}(\partial_r g)^2 > 0$ holds.*

The following is the analogue of Theorem 4.5 for hypercube data and which holds for all values of $p \in \mathbb{N}$. Indeed, we prove an even stronger estimate. In the regime $n = d^{p+\delta}$, the error $\mathbb{E}_n(\partial_r \hat{f})^2 - \mathbb{E}\big(\partial_r f_{\leq p}^*\big)^2$ decays at slow rate $\approx d^{-\delta/2}$ when r is an influential coordinate and at a fast rate d^{-1} if r is not influential! We will see that the improved rate for non-influential coordinates is the key reason why reweighted kernel machines can efficiently learn hierarchical polynomials. The proof of the theorem appears in Appendix B.2.

Theorem 4.6 (Gradient consistency on hypercube). *Suppose that Assumptions 1 and 3 hold. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0, 1)$ is a constant and suppose that condition (16) holds with $q = p$. Then for any $\epsilon > 0$, the KRR predictor $\hat{f}$ satisfies:*

$$\left| \mathbb{E}_n(\partial_r \hat{f})^2 - \mathbb{E}\big(\partial_r f_{\leq p}^*\big)^2 - c \cdot (d^{2\delta-2}) \cdot \mathbb{E}(\partial_r f_{p+1}^*)^2 \right|$$

$$\leq O_{d,\mathbb{P}}(1) \cdot \left(\frac{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2}{d} + \frac{\big\|\partial_r f_{\leq p}^*\big\|_{L_2}^2 + d^{2\delta-2}\big\|\partial_r f_{p+1}^*\big\|_{L_2}^2}{d^{\max(\delta, 1-\delta)/2-\epsilon}} + \frac{\|f^*\|_{L_2}\big\|\partial_r f_{\leq p}^*\big\|_{L_2}}{d^{1/2}} + \frac{(\|f^*\|_{L_2} + \sigma_\varepsilon)\big\|\partial_r f_{p+1}^*\big\|_{L_2}}{d^{3/2-\delta}} \right).$$

simultaneously for all $r = 1, \ldots, d$ where $c > 0$ is a constant.

Unpacking Theorem 4.6 we see that there are two interesting regimes corresponding to whether δ is greater or less than $1/2$. Namely, the following estimate holds

$$\mathbb{E}_n(\partial_j \hat{f})^2 \approx \begin{cases} \mathbb{E}(\partial_j f_{\leq p}^*)^2 & \text{if } \delta \in (0, \frac{1}{2}) \\ \mathbb{E}(\partial_j f_{\leq p}^*)^2 + c \cdot d^{-2(1-\delta)}\mathbb{E}(\partial_j f_{p+1}^*)^2 & \text{if } \delta \in (\frac{1}{2}, 1) \end{cases},$$

uniformly over all the coordinates up to a small order error. In words, this means that in the small oversampling regime $\delta \in (0, \frac{1}{2})$, the KRR estimator learns the coordinate weights of the p'th truncation $\mathbb{E}(\partial_j f_{\leq p}^*)^2$. Interestingly, in the larger oversampling regime $(\frac{1}{2}, 1)$, the KRR estimator can also detect the coordinate weights of the $(p+1)$-order term $\mathbb{E}(\partial_j f_{p+1}^*)^2$ when the lower-order weight $\mathbb{E}(\partial_j f_{\leq p}^*)^2$ is zero.

15

We note that Theorem 4.6 requires f^* to be a polynomial of constant degree per Assumption 3. This requirement was only needed to establish the fast rate d^{-1} for learning non-influential coordinates, namely the norm $\left\|\partial_r f^*_{\leq p}\right\|_{L_2} = \left\|\partial_r f^*_{p+1}\right\|_{L_2} = 0$. The slow rate $O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon})$ is valid without this requirement, thereby directly paralleling Theorem 4.5. We also note that as in the Gaussian case, the quantities $\mathbb{E}(\partial_r f^*_{\leq p})^2$ have an intuitive variational interpretation in terms of the derivatives of f^* itself. Namely, we have

$$\mathbb{E}\big(\partial_r f^*_{\leq p}\big)^2 = \mathbb{E}\left(\sum_{S:\, r\in S,\, |S|\leq p} b_S x^{S\setminus\{r\}}\right)^2 = \sum_{S:\, r\in S,\, |S|\leq p} b_S^2 = \sum_{S:\, r\in S,\, |S|\leq p} (\mathbb{E}[\mathsf{D}_S f^*])^2, \tag{18}$$

where the last inequality follows from (8).

Notice that the noise/error term in the theorem scales as d^{-1} and therefore the empirical weights estimator $\mathbb{E}_n(\partial_r \hat{f})^2$ fails to distinguish the coordinate weight of the $(p+1)$-order term $\mathbb{E}(\partial_j f^*_{p+1})^2$ from the noise when $\delta \in (0, \frac{1}{2})$. To address this limitation, we introduce the Derivative Norm (DN) estimator $\mathcal{D}(w)$, defined as

$$\mathcal{D}_j(w) := \beta_w^\top \frac{\partial K_w(X,X)}{\partial w_j} \beta_w, \tag{19}$$

where $\beta_w \in \mathbb{R}^n$ denotes the dual coefficients $\beta_w \in \mathbb{R}^n$ of the KRR predictor f_w induced by the reweighted kernel K_w. Analogous to Theorem 4.6, the following result shows that $\mathcal{D}(\mathbf{1}_d)$ provides an estimate of the population coordinate weights. The proof, which also extends to general weights, appears in Appendix B.6.

Theorem 4.7 (Derivative norm estimator on hypercube). *Suppose that Assumptions 1 and 3 hold. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant and suppose that condition (16) holds with $q = p+1$. Then for any $\epsilon > 0$, the following holds:*

$$\left| \frac{1}{n}\mathcal{D}_r(\mathbf{1}_d) - \sum_{k=0}^{p+1} \Theta_d(d^{-|k-p-\delta|}) \cdot \mathbb{E}(\partial_r f^*_k)^2 \right|$$

$$\leq O_{d,\mathbb{P}}(1) \cdot \left(\frac{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2}{d^{1-\epsilon}} + \frac{\sum_{k=0}^{p+1} d^{-|k-p-\delta|}\|\partial_r f^*_k\|_{L_2}^2}{d^{\max(\delta, 1-\delta)/2-\epsilon}} + \frac{\sum_{k=0}^{p+1} d^{-|k-p-\delta|}\|\partial_r f^*_k\|_{L_2}(\|f^*\|_{L_2}+\sigma_\varepsilon)}{d^{1/2-\epsilon}} \right).$$

simultaneously for all $r = 1, \ldots, d$.

The estimator $\mathcal{D}_j(w)$ addresses the limitation of $\mathbb{E}_n(\partial_j \hat{f})^2$ in estimating the coordinate weights of the $(p+1)$-order term when $\delta \in (0, \frac{1}{2})$. Indeed, in this regime, the last term in the sum is $\Theta_d(d^{1-\delta})\mathbb{E}(\partial_j f_{p+1})^2$ and the leading coefficient $d^{1-\delta}$ is large compared to the error term $O_{d,\mathbb{P}}(d^{-1})$. In contrast, for the coordinate weights of the p'th truncation $\mathbb{E}(\partial_j f_{\leq p})^2$, this estimator systematically underestimates the true coordinate weights compared to $\mathbb{E}_n(\partial_j \hat{f})^2$. A simple average of the empirical weight estimator and the DN estimator enjoys the best properties of both, as the following corollary shows.

Corollary 4.8. *Suppose that Assumptions 1 and 3 hold. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant and suppose that condition (16) holds with $q = p+1$. Then the following holds for any average parameter $\alpha \in [0,1]$:*

$$\left| (1-\alpha) \cdot \mathbb{E}_n(\partial_j \hat{f})^2 + \alpha \cdot \frac{1}{n}\mathcal{D}_j(\mathbf{1}_d) - (1 + o_{d,\mathbb{P}}(1))\left((1-\alpha)\mathbb{E}(\partial_j f_{\leq p})^2 + c \cdot \alpha d^{\delta-1}\mathbb{E}(\partial_j f_{p+1})^2\right) \right| = O_{d,\mathbb{P}}(d^{-1}). \tag{20}$$

simultaneously for all $r = 1, \ldots, d$ where $c > 0$ is a constant.

In light of Corollary 4.8, an appealing strategy to learn the function f^* is to iteratively reweight the data using a combination of the empirical weights $\mathbb{E}_n(\partial_r \hat{f})^2$ and DN-estimator $\mathcal{D}_r(w)$ corresponding to each coordinate $r = 1, \ldots, d$, then refit the data with kernel ridge regression, and so forth. This retraining procedure was informally summarized in Algorithm 1 and we formalize it here as Algorithm 3.

16

Algorithm 3 IRKM(α)   **I**teratively **R**eweighted **K**ernel **M**achines

1: **Require:** Iteration count $T \in \mathbb{N}$, kernel function K on $\mathbb{R}^d$, ridge parameter $\lambda \geq 0$, safeguard $\epsilon_{\mathrm{s}} > 0$, average parameter $\alpha \in [0, 1]$.

2: **Initialize:** $w = \mathbf{1}_d$.

3: **for** $t = 1, \ldots, T$ **do**:

4:     Sample fresh data $\{(x^{(i)}, y_i)\}_{i=1}^n$ and define the kernel $K_w(x, z) = K(\sqrt{w} \odot x, \sqrt{w} \odot z)$.

5:     Obtain predictor $\hat{f}_t$ from Kernel Ridge Regression:

$$\hat{f}_t(z) := K_w(z, X)\beta \qquad \text{where} \qquad \beta := (K_w(X, X) + \lambda I)^{-1}y.$$

6:     Compute the first weight vector: $w^{[1]} = \epsilon_{\mathrm{s}}\mathbf{1}_d + \cdot\frac{1}{n}\sum_{i=1}^n [\nabla_x \hat{f}_t(x^{(i)})]^{\odot 2}$.

7:     Compute the second weight vector: $w^{[2]} = \epsilon_{\mathrm{s}}\mathbf{1}_d + \frac{1}{n}\mathcal{D}(w) \odot w$.

8:     Normalize the weights $w^{[k]} = \frac{d}{\|w^{[k]}\|_1} \cdot w^{[k]}$, $k = 1, 2$.

9:     Update the weight vector $w = (1 - \alpha) \cdot w^{[1]} + \alpha \cdot w^{[2]}$.

10: **end for**

Let us make a few comments. The algorithm maintains a sequence of prediction functions $\hat{f}$ and weights w. Each iteration t of the algorithm begins by drawing fresh data $\{(x^{(i)}, y_i)\}_{i=1}^n$ with $x^{(1)}, \ldots, x^{(n)} \overset{\mathrm{iid}}{\sim} \mu$. The algorithm then forms a reweighted kernel $K_w(x, z)$, which in essence amounts to reweighing the feature vectors. The next prediction function $\hat{f}_t$ is then defined to be the KRR predictor obtained by fitting the data $\{(x^{(i)}, y_i)\}_{i=1}^n$ with the kernel K_w. The weight vector w is then updated using the average of the empirical coordinate weights and the DN estimator, safeguarded by adding a small positive constant $\epsilon_{\mathrm{s}} > 0$. This safeguard turns out to be important both in theory and in numerical experiments. The reason is intuitively clear: in the extreme case, a weight w_r on coordinate x_r that is set to zero in iteration t will remain zero for all future iterations. In other words, without the safeguard it would be impossible for the algorithm to recover from incorrectly classifying a coordinate as noninfluential. In the update, we obtain the second weight vector $w^{[2]}$ by taking the Hadamard product of the DN estimator with the current weight vector w analogously to a multiplicative weight update algorithm. Finally, we normalize the two intermediate weight vectors and combine them using an averaging parameter α to obtain the new weight vector w.

The analysis of Algorithm 3 proceeds in two stages. First, we describe which coordinates can be learned in a single step of IRKM(α) for a fixed weight vector w (regardless of how w originated). Secondly, we monitor the evolution of the weights and inductively apply the one-step guarantees in order to describe all the coordinate that are learned along with the generalization error of the final predictor.

The first stage, which states which coordinates can be learned in a single step of IRKM(α), is summarized in Theorem B.27. We defer this theorem to Appendix B.3 since its statement is quite technical. For the second stage, we introduce the following definition from [1] to describe the class of functions that IRKM(α) is capable of learning.

Definition 2 (Leap complexity). A function $f = \sum_{S \subset [d]} b_S \phi_S$ has *leap complexity* k, denoted $\mathrm{Leap}(f) = k$, if the nonzero coefficients $\{b_S\}$ can be ordered in such a way that the corresponding sets $\{S_i\}_{i=1}^m$ satisfy

$$|S_1| \leq k \qquad \text{and} \qquad \left| S_i \setminus \bigcup_{1 \leq j < i} S_j \right| \leq k,$$

and k is the smallest integer for which this is possible.

In other words, f has leap complexity k if k is the smallest integer such that its nonzero coefficients can be ordered in a way that the corresponding sets grow by at most k indices. As a concrete example, the polynomial

$$f^*(x) = x_1 + x_2 + x_1 x_2 x_3 + x_1 x_2 x_3 x_4, \tag{21}$$

17

which figured prominently in the introduction (5), has leap complexity $\text{Leap}(f^*) = 1$ since we may define $S_1 = \{1, 2\}$, $S_2 = \{1, 2, 3\}$, and $S_3 = \{1, 2, 3, 4\}$. We note that in the earlier work [2], functions with leap complexity one were said to satisfy the "merged staircase property".

In the same way that we may isolate a low-degree component of a function, we may define a low-leap complexity component. This is the content of the following definition.

Definition 3 (Maximum k-leap complexity component). For any function $f = \sum_{\mathcal{I} \subset [d]} b_{\mathcal{I}} \phi_{\mathcal{I}}$, its *maximum k-leap complexity component*, denoted by $\mathsf{L}_k f$, is defined as

$$\mathsf{L}_k f := \sum_{\mathcal{I} \in \mathcal{S}} b_{\mathcal{I}} \phi_{\mathcal{I}}, \quad \text{where} \quad \mathcal{S} = \underset{\mathcal{S}' \subseteq [d]}{\arg\max} \left\{ |\mathcal{S}'| : \text{Leap}\left(\sum_{\mathcal{I} \in \mathcal{S}'} b_{\mathcal{I}} \phi_{\mathcal{I}} \right) \leq k \right\}. \tag{22}$$

For example, the maximum 1-leap component of the function $g(x) = x_1 + x_2 + x_1 x_2 x_3 + x_3 x_4 x_6$ is $\mathsf{L}_1 f = x_1 + x_2 + x_1 x_2 x_3$. The following theorem shows that in the sampling regime $n = d^{p+\delta}$, a constant number of iterations of $\text{IRKM}(\frac{1}{2})$ learns the maximum $(p+1)$-leap component of a sparse polynomial. The proof appears in Appendix B.4.

Theorem 4.9 (Learning the maximal leap component). *Suppose that Assumptions 1 and 3 hold and that the number of influential coordinates of f^* scales as $\Theta(d^s)$ for some constant $s \in [0, 1)$. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0, 1)$ is a constant and suppose that condition (16) holds with $q = \ell$. Suppose the sparsity level satisfies $s < \frac{\delta}{\ell+1}$, choose any $\eta \in \left(1 - \frac{\delta}{\ell+1}, 1 - s\right)$ and mixing parameter $\alpha \in (0, 1)$. Then there exists a horizon $T = O_d(1)$ such that the predictor $\hat{f}_T$ returned by $\text{IRKM}(\alpha)$ (Algorithm 1) with safeguard $\epsilon_s = d^{-\eta}$ satisfies:*

$$\left\| \hat{f}_T - \mathsf{L}_{p+1} f^* \right\|_{L_2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2).$$

Let us illustrate Theorem 4.9 by applying it to the function (21). Recall that this function has leap complexity $p = 1$. Then applying Theorem 4.9 with $s = 0$ and $n = d^\delta$ for any $\delta \in (0, 1)$, we deduce that after a constant number of steps $T = O_d(1)$ the $\text{IRKM}(\frac{1}{2})$ algorithm returns a predictor $\hat{f}_T$ that is a consistent estimator of f^* in the sense that $\|\hat{f}_T - f^*\|_{L_2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$.

An important special case occurs when $f^*_{\leq p+1}$ and f^* share the same set of influential coordinates meaning that $\mathbb{E}(\partial_r f^*)^2$ is zero if and only if $\mathbb{E}(\partial_r f^*_{\leq p+1})^2$ is zero. In this situation, we have $\mathsf{L}_{p+1} f^* = f^*$ and therefore Theorem 4.9 reduces to Theorem 1.3 stated in the introduction. As a concrete example, consider the function $f(x) = x_1 + x_2 + x_3 + x_1 x_2 x_3$ (Eq. (2)) discussed in the introduction. Setting $p = 0$ gives $\mathsf{L}_1 f^* = f^*_{\leq 1} = x_1 + x_2 + x_3$. Note that Corollary 4.8 shows that the linear component $f_{\leq 1}(x) = x_1 + x_2 + x_3$ can be weakly learned (with coefficient $d^{\delta-1}$) for any average parameter $\alpha \in (0, 1)$ in the regime $n = d^\delta$. Thus we may instantiate Theorem 4.9 by setting $s = 0$, $p = 0$, and choosing any $\eta \in (1 - \frac{\delta}{4}, 1)$. Theorem 4.9 then implies that in a constant number of steps T the $\text{IRKM}(\frac{1}{2})$ algorithm return a predictor $\hat{f}_T$ that learns the third-order component $x_1 x_2 x_3$ and becomes an asymptotically consistent estimator for f^* in the sense that $\|\hat{f}_T - f^*\|_{L_2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$.

5 Proof sketches of Theorems 4.6 and 4.9

In this section, we outline the key ingredients underpinning our results. We will focus on the case of hypercube data as this is the setting when our strongest results are applicable.

5.1 Basic notation

For each $\lambda \in \{0, 1\}^d$, let $\phi_\lambda(x) = x^\lambda$ be the Fourier orthonormal basis element. For a subset $\mathcal{S} \subset \{0, 1\}^d$ define the feature vector $\phi_{\mathcal{S}}(x) := \{\phi_\lambda(x)\}_{\lambda \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, and the feature matrix $\Phi(X) \in \mathbb{R}^{n \times |\mathcal{S}|}$ whose i-th row is $\phi_{\mathcal{S}}(x^{(i)})$. We write $\mathcal{S}_k := \{\lambda : |\lambda| = k\}$ for the degree-k index set, and use $\Phi_{\mathcal{S}_k}(X)$ and $\Phi_{\leq k}(X)$ for the corresponding submatrices.

An important observation is that multiplication of x^S by a coordinate adds or removes that coordinate from the support:

$$x_r x^S = \begin{cases} x^{S \cup \{r\}}, & r \notin S, \\ x^{S \setminus \{r\}}, & r \in S. \end{cases}$$

We let $\mathcal{S}_j$ consist of all subsets of $[d]$ of cardinality j. Giving a coordinate r, we decompose each $\mathcal{S}_j$ as

$$\mathcal{S}_j^0 = \{S \in \mathcal{S}_j : r \notin S\} \qquad \text{and} \qquad \mathcal{S}_j^1 = \{S \in \mathcal{S}_j : r \in S\}. \tag{23}$$

5.2 Kernel and feature decompositions

Now define the matrices $K_{ij} = g\left(\langle x^{(i)}, x^{(j)} \rangle / d\right)$, $K'_{ij} = g'\left(\langle x^{(i)}, x^{(j)} \rangle / d\right)$, and $K_\lambda := K + \lambda I_n$. Direct algebraic manipulations show the convenient expression for the coordinate weights of the KRR predictor:

$$\mathbb{E}_n(\partial_r \hat{f})^2 = \frac{1}{nd^2} \|K' X_r \beta\|_2^2, \tag{24}$$

where $X_r \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entry $(X_r)_{i,i} = x_r^{(i)}$ for $i \in [d]$ and $\beta = K_\lambda^{-1} y$ are the dual coefficients. Now a natural idea is to rewrite K and K' in the Fourier basis in order to identify the dominant terms in this expression. Interestingly, in this basis the two matrices are nearly diagonal up to an identity perturbation:

$$K = \Phi_{\leq p} D \Phi_{\leq p}^\top + \rho I_n + \Delta_1, \tag{25}$$

$$K' = \Phi_{\leq p} D' \Phi_{\leq p}^\top + \rho' I_n + \Delta, \tag{26}$$

where $\rho, \rho' > 0$ are some constants, $\|\Delta\|_{\text{op}}, \|\Delta_1\|_{\text{op}} = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\varepsilon})$ are error terms, and D and D' are diagonal matrices satisfying

$$D_{\mathcal{S}_k}, D'_{\mathcal{S}_k} = \Theta_d(d^{-k}) I_{|\mathcal{S}_k|}.$$

5.3 Empirical weights estimator

Inserting (26) into (24) and carefully bounding the error terms gives the intuitive expression

$$\mathbb{E}_n(\partial_r \hat{f})^2 = \frac{1 + o_{d,\mathbb{P}}(1)}{nd^2} \left\|\beta^\top X_r \Phi_{\leq p} D' \Phi_{\leq p}^\top\right\|_2^2 + O_{d,\mathbb{P}}(d^{-1}\sigma_\varepsilon^2).$$

Now one can show that $\frac{1}{n} \Phi_{\leq p}^\top \Phi_{\leq p}$ is almost an isometry in the sense that $\left\|\frac{1}{n} \Phi_{\leq p}^\top \Phi_{\leq p} - I_n\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon})$. Therefore, after controlling the error terms, we can further simplify to

$$\mathbb{E}_n(\partial_r \hat{f})^2 = \frac{1 + o_{d,\mathbb{P}}(1)}{d^2} \cdot \|\beta^\top X_r \Phi_{\leq p} D'\|_2^2. \tag{27}$$

Next, let us focus on rewriting the product $X_r \Phi_{\leq p}$ in the Fourier basis. To this end, notice that multiplication by x_r shifts degrees. Simple algebra and orthogonality considerations show that the right-side of (32) is proportional to

$$\sum_{j=1}^{p+1} \|\beta^\top \Phi_{\mathcal{S}_j^1} (\tfrac{1}{d} D'_{\mathcal{S}_{j-1}^0})\|_2^2 + \sum_{j=0}^{p-1} \|\beta^\top \Phi_{\mathcal{S}_j^0} (\tfrac{1}{d} D'_{\mathcal{S}_{j+1}^1})\|_2^2. \tag{28}$$

Writing $y = f^*(x) + \varepsilon = \sum_S b_S x^S + \varepsilon$ in the Fourier basis, we decompose $\beta = K_\lambda^{-1} y$ as:

$$\beta = \sum_{j=1}^{p+1} K_\lambda^{-1} \Phi_{\mathcal{S}_j^1} b_{\mathcal{S}_j^1} + \sum_{j=0}^{p-1} K_\lambda^{-1} \Phi_{\mathcal{S}_j^0} b_{\mathcal{S}_j^0} + K_\lambda^{-1}(\Phi_{\mathcal{C}} b_{\mathcal{C}} + \varepsilon) \tag{29}$$

where $\mathcal{C} := 2^{[d]} / \left(\left(\bigsqcup_{j=0}^{p-1} S_j^0\right) \cup \left(\bigsqcup_{j=1}^{p+1} S_j^1\right)\right)$ and $\varepsilon \in \mathbb{R}^n$ denotes the noise vector with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Now the key observation is $K_\lambda^{-1} \Phi_{\mathcal{S}_k}$ is essentially a multiple of $\Phi_{\mathcal{S}_k}$, where the multiple depends on whether k is smaller or greater than p. This is the content of the following proposition, which follows from the arguments in the proof of Theorem 4.6.

Proposition 5.10. *In the regime $n = d^{p+\delta}$, the following estimate holds*

$$K_\lambda^{-1} \Phi_{\mathcal{S}_k} b_{\mathcal{S}_k} \approx \begin{cases} \Phi_{\mathcal{S}_k}(nD_{\mathcal{S}_k})^{-1} b_{\mathcal{S}_k}, & \text{if } k \leq p, \\ \Phi_{\mathcal{S}_k} b_{\mathcal{S}_k}, & \text{if } k > p, \end{cases} \tag{30}$$

up to an $o_{d,\mathbb{P}}(\|\cdot\|_2)$ error.

Note that $\frac{n}{d} D_{\mathcal{S}_k} = \Theta_d(d^{\delta-1}) I_{|\mathcal{S}_k|}$ for $k \leq p$ and therefore K_λ^{-1} roughly acts on $\Phi_{\mathcal{S}_k}$ as a contraction. Combining (28), (29), and Proposition 5.10 and using orthogonality of Fourier blocks, we can obtain

$$\|\beta^\top X_r \Phi_{\leq p}(\tfrac{1}{d}D')\|_2^2 = A + B + O_{d,\mathbb{P}}(d^{-1}), \tag{31}$$

where A collects the $\mathcal{S}_j^1$ terms and B the $\mathcal{S}_j^0$ terms. A detailed calculation yields

$$A \approx \left\| b_{\mathcal{S}_{p+1}^1}^\top \underbrace{\left(\tfrac{1}{n} \Phi_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1} \right)}_{\approx I_{|\mathcal{S}_{p+1}^1|}} (\tfrac{n}{d} D'_{\mathcal{S}_p^0}) \right\|_2^2 + \sum_{j=1}^p \left\| b_{\mathcal{S}_j^1}^\top D_{\mathcal{S}_j^1} \underbrace{\left(\tfrac{1}{n} \Phi_{\mathcal{S}_j^1}^\top \Phi_{\mathcal{S}_j^1} \right)}_{\approx I_{|\mathcal{S}_j^1|}} (\tfrac{1}{d} D'_{\mathcal{S}_{j-1}^0}) \right\|_2^2 .$$

Noting that $(\tfrac{n}{d} D'_{\mathcal{S}_p^0}) = \Theta_d(d^{\delta-1}) I_{|\mathcal{S}_p^0|}$ in the first term and that $D_{\mathcal{S}_j^1}$ cancels out with $(\tfrac{1}{d} D'_{\mathcal{S}_{j-1}^0})$ in the second term, the above equation reduces to

$$A \approx \Theta_d(d^{2\delta-2}) \|b_{\mathcal{S}_{p+1}^1}\|_2^2 + \sum_{j=1}^p \|b_{\mathcal{S}_j^1}\|_2^2 .$$

A similar argument can be applied to show $B = O_{d,\mathbb{P}}(d^{-2})$. Combining all the ingredients yields the desired expression

$$\mathbb{E}_n(\partial_r \hat{f})^2 \approx \Theta_d(d^{2\delta-2}) \|b_{\mathcal{S}_{p+1}^1}\|_2^2 + \sum_{j=1}^p \|b_{\mathcal{S}_j^1}\|_2^2 + O_{d,\mathbb{P}}(d^{-1} \sigma_\varepsilon^2). \tag{32}$$

5.4 DN estimator for $w = \mathbf{1}_d$

The computations for the DN estimator with $w = \mathbf{1}_d$ proceed similarly. Namely, elementary algebra shows:

$$\mathcal{D}_r(\mathbf{1}_d) = \frac{1}{nd} \|(K')^{\frac{1}{2}} X_r \beta\|_2^2.$$

Analogously to the argument that led to (27), the estimator $\mathcal{D}_r$ can be approximated by

$$\mathcal{D}_r(\mathbf{1}_d) \approx \left\| \beta^\top \Phi_{\leq p} X_r \left(\frac{1}{nd} D' \right)^{\frac{1}{2}} \right\|_2^2.$$

Compared to Eq. (27), we observe that the only difference is that $(d^{-1}D')$ is replaced by $\left(\frac{1}{nd}D'\right)^{\frac{1}{2}}$. Continuing the same argument, but with this different scaling yields the approximation

$$\mathcal{D}_r(\mathbf{1}_d) \approx \sum_{j=1}^{p+1} \Theta_{d,\mathbb{P}}(d^{-|j-p-\delta|}) \left\| b_{\mathcal{S}_j^1} \right\|_2^2 + O_{d,\mathbb{P}}(d^{-1+\epsilon} \sigma_\varepsilon^2). \tag{33}$$

5.5 Extension to general weights

Now we extend the proof to the case where weights are present and highlight how the iterative process IRKM(α) can learn hierarchical functions. We consider a weight $w \in \mathbb{R}_+^d$ that satisfies $\|w\|_1 = d$. Along the iterations of IRKM(α), the coordinates of w may exhibit different scalings in d. It will be important to monitor these dimension dependencies. To this end, fix a partition $\pi_N = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$ of $[d]$ with

$$|\mathcal{T}_k| = \Theta_d(d^{s_k}), \quad w_r = \Theta_d(d^{1-\lambda_k}), \quad \forall r \in \mathcal{T}_k. \tag{34}$$

For a tuple $\boldsymbol{T} = (T_1, T_2, \ldots, T_N)$ with $T_k \in 2^{\mathcal{T}_k}$ for $k \in [N]$, we define its corresponding Fourier feature

$$\phi_{\boldsymbol{T}} := \phi_{\cup_{k=1}^N T_k} = \prod_{k=1}^N x^{T_k}. \tag{35}$$

and we define its *effective degree* with respect to w as

$$p_{\text{eff}}(\boldsymbol{T}) := \sum_{k=1}^N \lambda_k \cdot |T_k|, \tag{36}$$

As a concrete example for the initial weights $w = \mathbf{1}_d$, we have $k = 1$, $\mathcal{T}_1 = [d]$, and $\lambda_1 = 1$. Then the effective degree of any $\boldsymbol{T}$ is simply $|T_1|$, which is exactly the degree of the feature. In the more general setting, the effective degree replaces the role of the polynomial degree p. Understanding how this quantity evolves along the iterations is key to obtaining guarantees for IRKM(α).

Consider the empirical weights for the inner product kernel $(K_w)_{ij} = g\left(\langle \sqrt{w} \odot x^{(i)}, \sqrt{w} \odot x^{(j)} \rangle / d\right)$ and set $K_{w,\lambda} := K_w + \lambda I_n$. As in Eq. (24), direct manipulations show

$$\mathbb{E}_n\left(\partial_r \hat{f}_w\right)^2 = \frac{w_r^2}{nd^2}\left\|K_w' X_r K_{w,\lambda}^{-1} y\right\|_2^2. \tag{37}$$

Next, in parallel to Eq. (25), we can decompose K_w on Fourier blocks truncated now by effective degree:

$$K_w = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + \rho I_n + \Delta_1, \tag{38}$$

where

$$\mathcal{A}_{\leq p_\delta} := \left\{\boldsymbol{T} = (T_1, T_2, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k} \text{ and } p_{\text{eff}}(\boldsymbol{T}) \leq p + \delta\right\},$$

The diagonal entries of $\hat{D}$ scale as

$$\widehat{D}_{\boldsymbol{T},\boldsymbol{T}} = \Theta_d(d^{-p_{\text{eff}}(\boldsymbol{T})}),$$

which is a direct analog of D. Note that for the initial weights $w = \mathbf{1}_d$, since the effective degree is simply the degree of the feature, the set $\mathcal{A}_{\leq p_\delta}$ contains all Fourier features with degree not exceeding p. Thus $\mathcal{A}_{\leq p_\delta}$ is the analogues of all Fourier features of degree bounded by p.

Define the dual coordinates $\beta_w := K_{w,\lambda}^{-1} y$. In parallel to Eq. (27), the following term will dominate the empirical weights:

$$\mathbb{E}_n\left(\partial_r \hat{f}_w\right)^2 \approx \left\|\beta_w^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}}\left(\frac{w_r}{d} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}}\right)\right\|_2^2,$$

where $\widehat{D}'$ is the diagonal matrix appearing in the decomposition of K_w'. Fix now a coordinate index r and let ι_r be the unique index satisfying $r \in \mathcal{T}_{\iota_r}$. We now define

$$\mathcal{A}_{p_\delta}^+(r) := \left\{\boldsymbol{T} = (T_1, T_2, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k} \text{ and } 0 < p_{\text{eff}}(\boldsymbol{T}) - (p + \delta) < \lambda_{\iota_r}\right\}.$$

Henceforth, we will suppress the subscript r on ι_r if it is clear from context. We will see that $\mathcal{A}_{p_\delta}^+(r)$ plays the role of degree $p+1$ feasrures in the equiweighted case. Analogous to (32) we can derive the estimates for the general weights:

$$\mathbb{E}_n(\partial_r \hat{f}_w)^2 \approx \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} \Theta_d(d^{-2|p_{\text{eff}}(\boldsymbol{T})-(p+\delta)|})\mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta}} \mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota}\sigma_\varepsilon^2). \quad (39)$$

The DN-estimator for general weights can be analogously derived:

$$w_r \cdot \mathcal{D}_r(w) \approx \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r) \sqcup \mathcal{A}_{\leq p_\delta}} \Theta_d(d^{-|p_{\text{eff}}(\boldsymbol{T})-(p+\delta)|})\mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota+\epsilon}\sigma_\varepsilon^2). \quad (40)$$

Here for simplicity of presentation, we assume there is no $\boldsymbol{T}$ such that $p_{\text{eff}}(\boldsymbol{T}) = p + \delta$, since this case has to be treated separately.

Algorithm IRKM(α) learns leap $(p+1)$ functions. Suppose now that f^* has leap complexity $p+1$. For any averaging parameters $\alpha \in (0,1)$, the two estimators (39) and (40) together give

$$(1-\alpha)\mathbb{E}_n(\partial_r \hat{f}_w)^2 + \alpha w_r \cdot \mathcal{D}_r(w)$$
$$\approx (1-\alpha) \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta}} \mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 + \alpha \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} \Theta_d(d^{-|p_{\text{eff}}(\boldsymbol{T})-(p+\delta)|})\mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota+\epsilon}\sigma_\varepsilon^2).$$

From the estimate we conclude that a feature $x_{\boldsymbol{T}}$ that depends on coordinate x_r can be identified when its effective degree (36) does not exceed $p + \delta + \lambda_{\iota_r}$, where λ_{ι_r} is initially 1. As the iterations proceed, one can show that any coordinate r that is learned—meaning that one of the summands in the expression is nonzero—will be assigned a large weight after the normalization step (Line 8 in Algorithm 1), that is of order d^η with $\eta \in (0,1)$. This reweighting reduces the effective degree of every feature that depends on x_r, thereby enabling those features to be learned in subsequent iterations.

Consider any feature that is the product of learned and unlearned coordinates. A careful accounting of normalization, safeguarding and sparsity shows that (1) each unlearned coordinate retains weight of the order $\Theta_d(1)$; hence the corresponding $\lambda_\iota = 1$ and it contributes one to the effective degree, and (2) the combined contribution of the learned coordinates can be controlled to remain below δ since these coordinates are assigned large weights. Consequently, in the regime $n = d^{p+\delta}$, any feature whose effective degree is less than $p + \delta + 1$ can be identified as discussed, and can contain at most $p + 1$ new coordinates (see Figure 4 for a visual illustration). Since f^* is assumed to be finite degree, this careful accounting also shows that a constant number of iterations of IRKM(α) suffices to learn leap $(p+1)$ component of f^*.
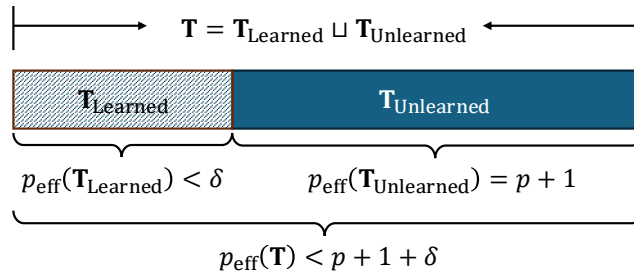

Figure 4: Illustration of the effective degree of a feature as the product of learned and unlearned coordinates.

6 Numerical experiments

In this section, we demonstrate the efficiency of IRKM(α) (Algorithm 1) when learning sparse functions across diverse data distributions and with higher leap complexity than considered in the introduction.

Our experimental results show that IRKM(α) often outperforms neural networks on these tasks including epistasis detection, and achieves comparable performance to state-of-the-art methods on UCI classification tasks [34]. The code for IRKM(α) that we use is available at https://github.com/zhm1995/IRKM.

For fixed training size n, we train both neural networks and kernel machines using $2n$ samples, drawing fresh n samples at each IRKM(α) training step. All neural networks reported in this section are fully connected and trained with Adam, using widths and batch sizes in $\{128, 256, 512\}$, depth in $\{2, 3, 4, 5, 6\}$ and learning rate in $\{10^{-3}, 10^{-4}\}$. We report the smallest test loss across these configurations.

Generally speaking, in our experiments, wider and deeper networks tend to perform better. As a concrete illustration, Figure 5 presents the test performance of fully connected neural networks across depths from 2 to 6 and width in $\{128, 256, 512\}$ for the experimental settings in Figure 1, 2 and 3 of the introduction. We observe that the test performance is improved as the depth or the width increases, and this trend persists in all our experiments.
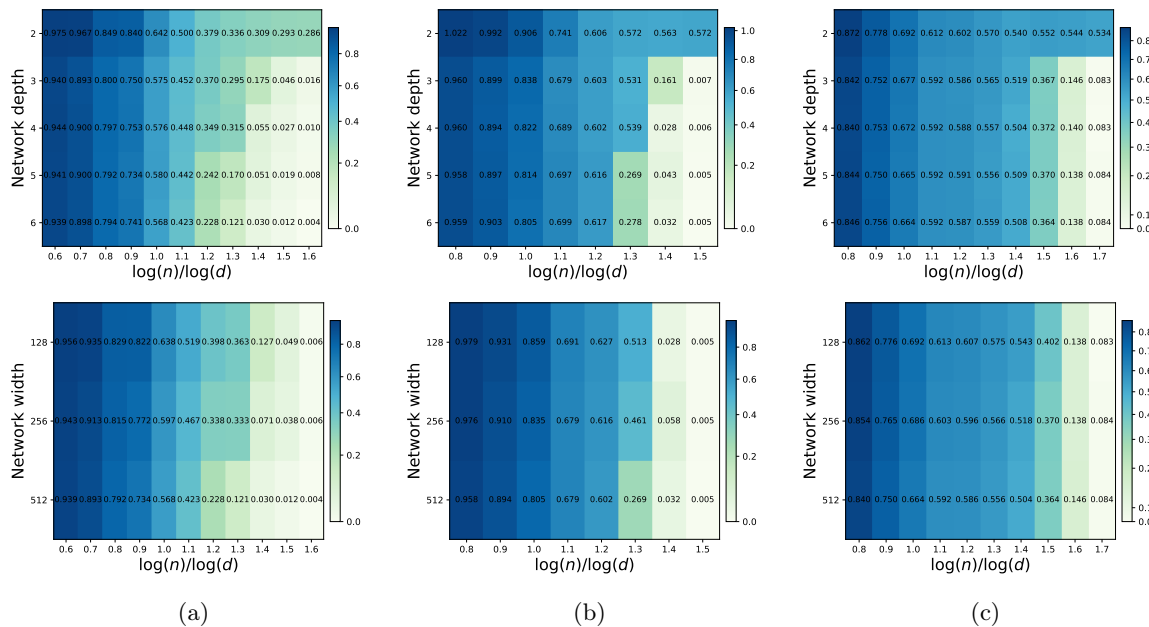

Figure 5: **Best test performance of neural networks across various widths and depths.** Columns (a), (b) and (c) correspond to the experimental setting in Figure 1, 2 and 3 respectively. We train a fully-connected network by Adam, varying the widths and batch sizes as $\{128, 256, 512\}$, depth as $\{2, 3, 4, 5, 6\}$ and learning rate as $\{10^{-3}, 10^{-4}\}$ and report the smallest test loss. For all experiments, we report an average of 5 runs.

## 6.1 Evaluation of IRKM(α) on Gaussian and CIFAR-10 data

Though our theoretical analysis of IRKM(α) focuses on hypercube data, we now empirically demonstrate its impressive performance across multiple input distributions. Beyond the hypercube experiments in the introduction, we evaluate IRKM(α) for learning sparse functions with inputs drawn from Gaussian distributions and the CIFAR-10 dataset [46].

Analogously to Figures 1, 2 and 3 for hypercube data, we run IRKM($\frac{1}{2}$) on samples from Gaussian distributions to learn the same target functions. The results are reported in Figures 6A, 6B and 7A, respectively. For the CIFAR-10 dataset, we randomly select 200 pixels from each image and learn sparse functions and multi-index models based on these selected pixels. See the results in Figure 6C and 7B respectively. The conclusion is exactly the same in all cases as for hypercube data.
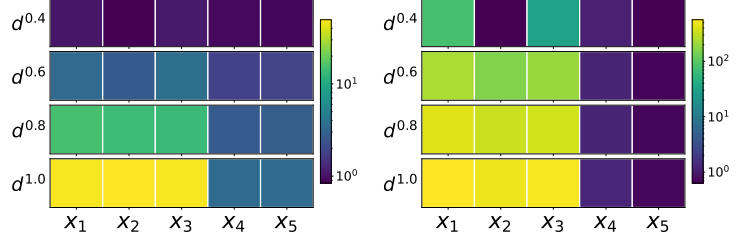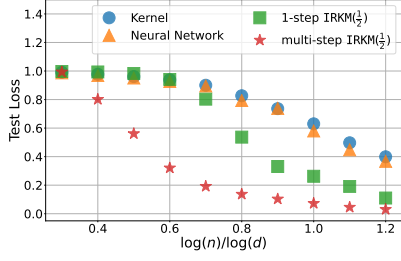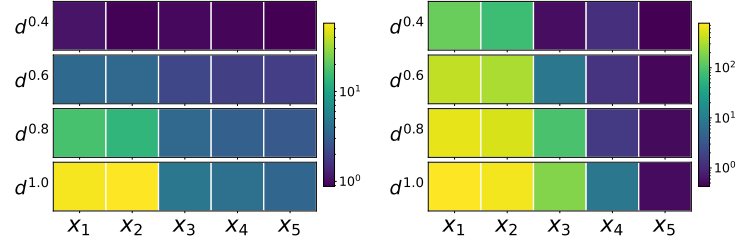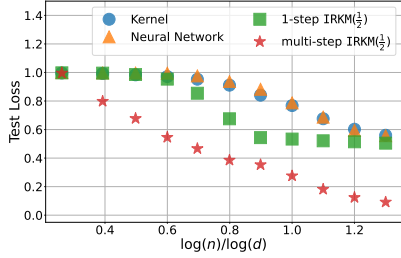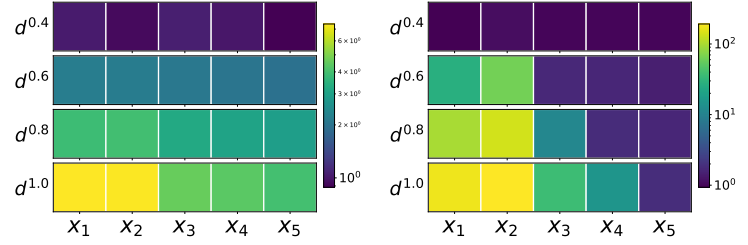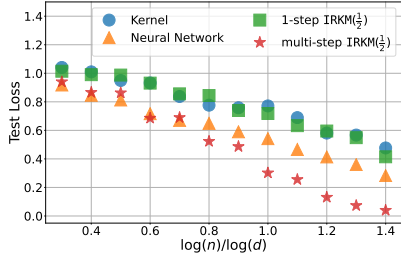
(1) Generalization error   (2) One-step identification   (3) Multi-step identification

Figure 6: **Column. (1)** Test loss for kernel machines, neural networks and $\mathtt{IRKM}(\frac{1}{2})$. **(2)** The empirical coordinate weights at the first step. **(3)** The empirical coordinate weights at step T.
Row. A: synthetic data $x \sim \mathcal{N}(0, I_{500})$, label $y = x_1 + x_2 + x_3 + x_1 x_2 x_3 + \varepsilon$. **B:** synthetic data $x \sim \mathcal{N}(0, I_{500})$, label $y = x_1 + x_2 + x_1 x_2 x_3 + x_1 x_2 x_3 x_4 + \varepsilon$. **C:** 200 random pixels from CIFAR-10, same label rule as row B. The noise ε is i.i.d. drawn from $\mathcal{N}(0, 0.1^2)$. We use the Laplacian kernel for both kernel machines and $\mathtt{IRKM}(\frac{1}{2})$. The $\mathtt{IRKM}(\frac{1}{2})$ algorithm is iterated for at most $T = 20$ steps for row A, and 10 steps for row B and C. We train a fully-connected network by Adam. For all experiments, we report an average of 10 runs.

A

(A1) Generalization error   (A2) AGOP error   (A3) Multi-step identification

B
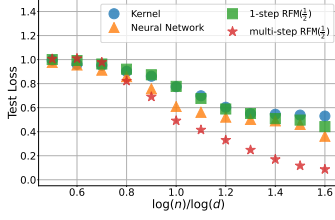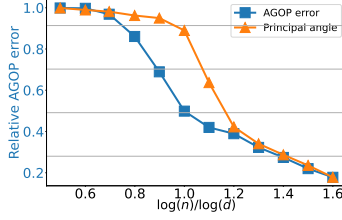
(B1) Generalization error   (B2) AGOP error   (B3) Multi-step identification
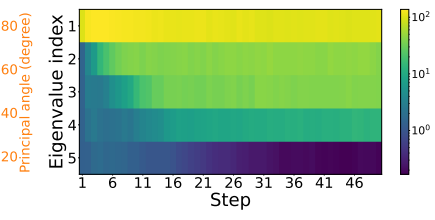
Figure 7: **(1) Test loss for kernel machines, neural networks and** $\mathtt{RFM}(\frac{1}{2})$**) (Algo. 2). (2) Relative AGOP error and the principal angle between the top-4 eigenspace of AGOPs. (3) Eigenvalues of AGOP at step** $1-50$ **in the regime** $n = d^{1.4}$**. A:** synthetic data $x \sim \mathcal{N}(0, I_{500})$, label $y = z_1 + z_2 + z_3 + z_1 z_2 z_3 + \varepsilon$. **B:** 200 random pixels from ZCA whitened CIFAR-10, same label rule as row A. The noise ε is i.i.d. drawn from $\mathcal{N}(0, 0.1^2)$ and $z = Ux$ where U is a random rotation matrix. We use the Laplacian kernel for both kernel machines and $\mathtt{RFM}(\frac{1}{2})$. We train a fully-connected network by Adam. For all experiments, we report an average of 10 runs.

6.2 Learning functions with leap complexity two

So far, we have experimented only with functions that have leap complexity one. We now consider target functions with leap complexity two:

$$f^*(x) = x_1 x_2 + x_1 x_2 x_3 x_4,$$

and show that $\mathtt{RFM}(\alpha)$ outperforms neural networks on both hypercube and Gaussian synthetic datasets. Our experimental results appear in Figures 8.

A
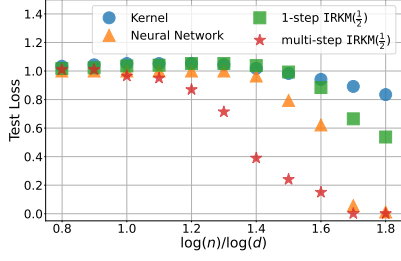
(A1) Generalization error
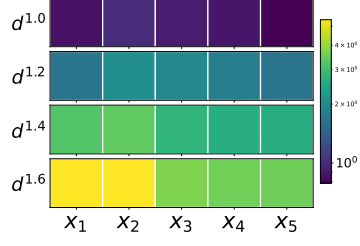
(A2) One step identification

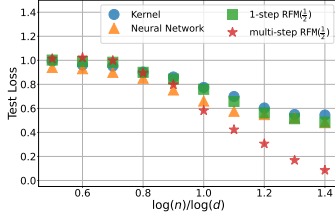(A3) Multi-step identification

B
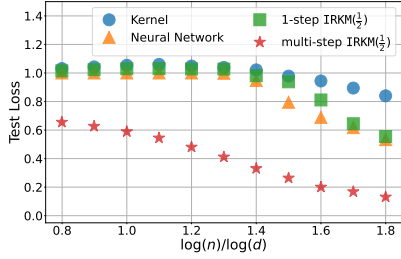
(B1) Generalization error

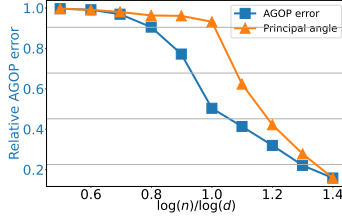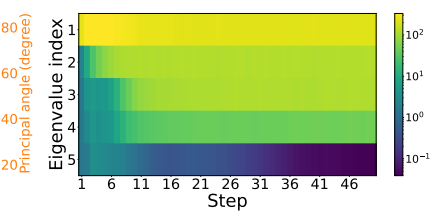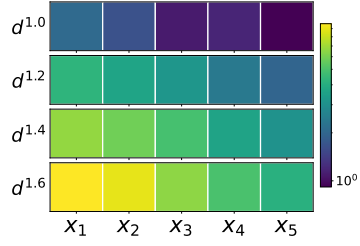(B2) One step identification
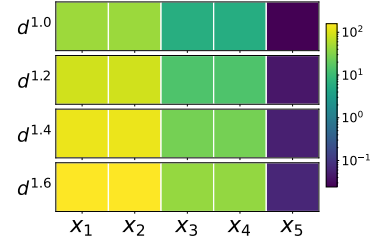
(B3) Multi-step identification

Figure 8: **(A1,B1) Test loss for kernel machines, neural networks and** $\mathtt{RFM}(\frac{1}{2})$**. (A2,B2) The empirical coordinate weights at the first step. (A3,B3) The empirical coordinate weights at step** T**.** The samples are i.i.d. uniformly drawn from the hypercube $\{\pm 1\}^d$ (Row A) and $\mathcal{N}(0, I_d)$ (Row B) with $d = 100$. The label is $y = x_1 x_2 + x_1 x_2 x_3 x_4 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. We use the Laplacian kernel for both kernel machines and $\mathtt{RFM}(\frac{1}{2})$. The $\mathtt{RFM}(\alpha)$ algorithm is iterated for at most $T = 20$ steps. We train a fully-connected network by Adam. For all experiments, we report an average of 10 runs.

6.3 Effect of the mixing parameter α on $\mathtt{IRKM}(\alpha)$ performance

We vary the value of the mixing parameter α and evaluate the test performance of $\mathtt{IRKM}(\alpha)$. Overall, the results indicate that setting $\alpha \approx \frac{1}{2}$ yields the best performance. The corresponding experimental results are presented in Figure 9.
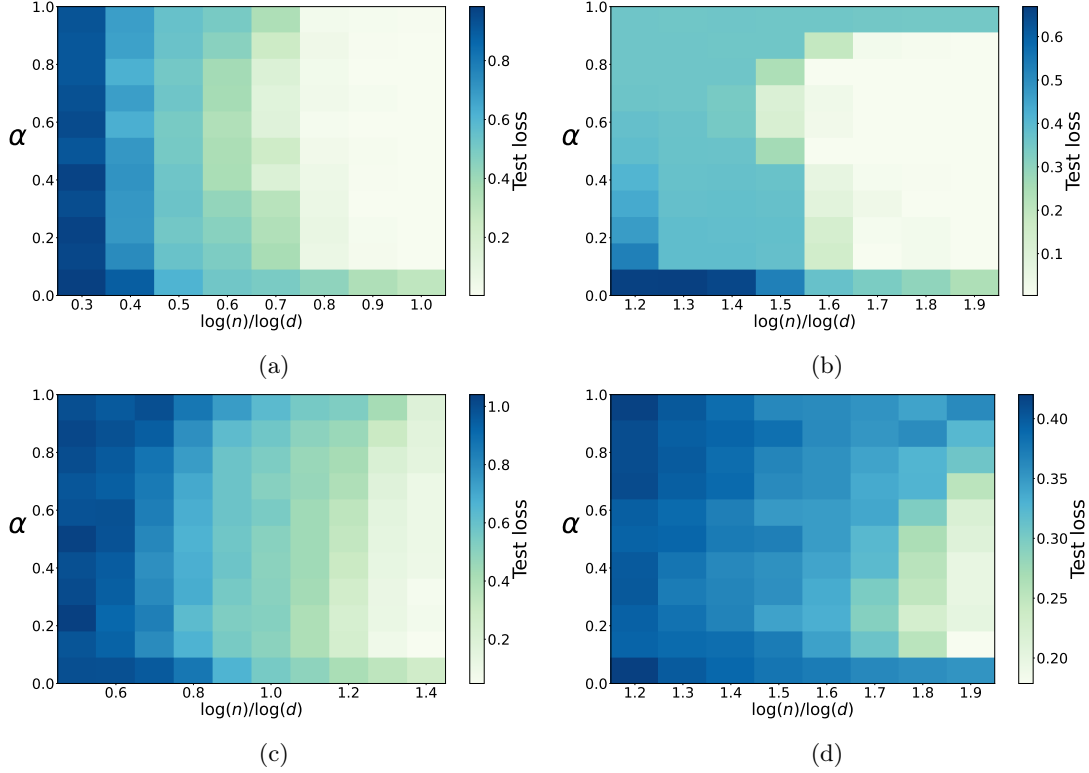
Figure 9: **Test loss of** `IRKM`(α)/`RFM`(α) **with varying mixing parameter** α**.** Panel (a): $y = x_1 + x_2 + x_1x_2x_3 + x_1x_2x_3x_4 + \varepsilon$, $x \sim \{\pm 1\}^{500}$. Panel (b): $y = x_1 + x_2x_3 + x_1x_2x_3x_4x_5 + \varepsilon$, $x \sim \{\pm 1\}^{100}$. Panel (c): $y = z_1 + z_2 + z_1z_2z_3 + z_1z_2z_3z_4 + \varepsilon$, $z = Ux$ where $x \sim \{\pm 1\}^{500}$ and U is a random rotation matrix. Panel (d): $y = x_1 + x_2x_3 + x_1x_2x_3x_4x_5 + \varepsilon$, $x \sim \mathcal{N}(0, I_{100})$. The noise ε is i.i.d. drawn from $\mathcal{N}(0, 0.1^2)$. `IRKM`(α)/`RFM`(α) uses the Laplacian kernel, is run for $T = 20$ steps for Panel (a,b,d) and for $T = 50$ steps for Panel (c), and results are averaged over 10 runs.

6.4   Performance on tabular datasets

We evaluate `IRKM`(α) on tabular datasets that potentially require learning sparse functions. Using the benchmark from [34], which compares 179 machine learning methods across 121 tabular classification tasks, we benchmark `IRKM`(α) against neural networks, random forests, and RFM, previously reported as state-of-the-art for these datasets [60]. Note that in each iteration of `IRKM`(α), we reuse the entire training set instead of sampling fresh data. We evaluate the performance based on the following standard metrics:

- Average accuracy: The average accuracy of the classifier across all datasets.

- P90/P95: The percentage of datasets where the classifier achieved accuracy within 90%/95% of the best model.

- Friedman rank: The average rank of the classifier across all datasets.

Table 1 shows that `IRKM`(0) achieves comparable results to RFM, while both significantly outperform random forests and neural networks. This in particular suggests that the prediction functions in these data sets tend to be sparse; that is, many feature coordinates do not significantly influence the label. We report only the results for `IRKM`(0), as this setting achieves the highest average accuracy.

Classifier (λ)	Avg. Accuracy (%)	P90	P95	Friedman Rank
IRKM (ours)	84.69	92.56	79.34	25.80
RFM	85.15	93.39	78.51	20.37
Random Forest	80.01	75.43	64.46	62.21
Neural Network	78.46	64.46	46.28	64.71

Table 1: Test performance of `IRKM`(0) against RFM, Random Forest, and Neural Networks.

6.5 Performance on Epistatic Interactions

In Genome-Wide Association Studies (GWAS), epistasis detection involves identifying genetic variant interactions that jointly influence disease phenotypes. This task is important in biology as it can reveal the underlying mechanisms of complex diseases. The challenge is difficult because combinations of SNPs may contribute to disease risk while individual SNPs show no detectable effects. This creates nonlinear sparse functions in high-dimensional data where only a small subset of genetic variants are truly relevant.

We evaluate our approach using the GAMETES [68] simulator. GAMETES generates pure epistatic interactions through penetrance tables where only multi-locus genotype combinations (AA, Aa, aa) predict phenotype, and individual loci show no main effects in these models. This provides controlled ground truth for assessing our algorithm's ability to identify relevant feature subsets among genetic variants. Each genetic variant (i.e., SNP) takes values $\{0, 1, 2\}$ corresponding to the count of minor alleles, and the classification target is binary disease status.

We conduct experiments with 2-way epistatic interactions, namely the label depends on 2 SNPs, varying the heritability parameter across $\{0.05, 0.1, 0.2\}$. Higher heritability values correspond to stronger genetic effects and improved signal-to-noise ratios. Classification performance is evaluated using the AUC metric (Area Under the ROC Curve), comparing `IRKM`($\frac{1}{2}$) against neural networks and random forests. We consider fully-connected networks with depth 6 and width 512, trained by Adam with mini-batch size 128. Note that in each iteration of `IRKM`($\frac{1}{2}$), we reuse the entire training set instead of sampling fresh data. The results are presented in Figure 10.

(a) Heritability = 0.05   (b) Heritability = 0.1   (c) Heritability = 0.2
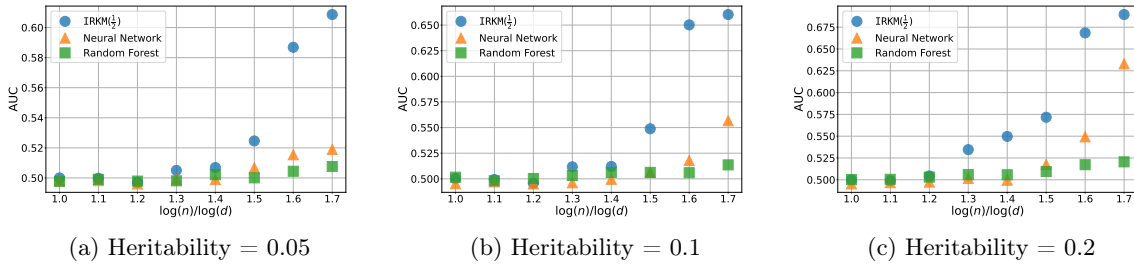
Figure 10: **AUC for `IRKM`($\frac{1}{2}$), neural networks and random forest.** The datasets are generated by GAMETES simulator with 100 SNPs, where the minor allele frequency is set to 0.2 for both interacting loci. We set the number of epistatic attributes to 2 and vary heritability across 0.05, 0.1, and 0.2. We report the average of 10 independent runs.

7 Conclusion

In this work, we have demonstrated that iteratively retrained kernel machines can efficiently learn coordinate features and hierarchical structures, thereby challenging the prevailing belief that these phenomena are unique to neural networks. Specifically, we established two key results: (1) the empirical coordinate weights of the KRR predictor can serve as an asymptotically consistent estimator for the population coordinate weights, even while the generalization error remains large; and (2) IRKM can learn functions

with leap complexity p using roughly d^{p-1} samples, thereby demonstrating that kernel methods effectively learn hierarchical structure. These findings suggest that kernel-based methods are much more powerful than previously thought and their feature-learning ability deserves further investigation. More broadly, our work contributes to understanding what makes learning algorithms effective, suggesting that adapting to the intrinsic structure of data may be more fundamental than the specific model class that is employed.

References

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. "Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 2552–2623.

[2] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. "The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4782–4887.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. "Backward feature correction: How deep learning performs deep (hierarchical) learning". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 4598–4598.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. "What can resnet learn efficiently, going beyond kernels?" In: *Advances in Neural Information Processing Systems* 32 (2019).

[5] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. "Living on the edge: Phase transitions in convex programs with random data". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 224–294.

[6] George E Andrews and Richard Askey. "Classical orthogonal polynomials". In: *Polynômes Orthogonaux et Applications: Proceedings of the Laguerre Symposium held at Bar-le-Duc, October 15–18, 1984*. Springer. 2006, pp. 36–62.

[7] David Aristoff, Mats Johnson, Gideon Simpson, and Robert J Webber. "The fast committor machine: Interpretable prediction with kernels". In: *The Journal of chemical physics* 161.8 (2024).

[8] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. "Learning in the presence of low-dimensional structure: a spiked random matrix perspective". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 17420–17449.

[9] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. "High-dimensional asymptotics of feature learning: How one gradient step improves the representation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 37932–37946.

[10] Francis Bach. "Breaking the curse of dimensionality with convex neural networks". In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[11] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[12] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. "Hidden progress in deep learning: Sgd learns parities near the computational limit". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 21750–21764.

[13] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. DOI: 10.1073/pnas.1903070116.

[14] Y Bengio. *Learning Deep Architectures for AI*. 2009.

[15] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. "Learning single-index models with shallow neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 9768–9783.

[16] Brendan Bordelon, Abla Canatar, and Cengiz Pehlevan. "Spectrum of kernel random matrices at finite rank". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 10332–10343. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/2c04c6bd3a84db6d8a0d8f6d984c0fb0-Paper.pdf.

[17] Emmanuel J Candes, Justin K Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.8 (2006), pp. 1207–1223.

[18] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. "Enhancing sparsity by reweighted ℓ 1 minimization". In: *Journal of Fourier analysis and applications* 14 (2008), pp. 877–905.

[19] Emmanuel J Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on information theory* 52.2 (2006), pp. 489–509.

[20] Yandong Cao and Quanquan Gu. "Generalization Bounds for Overparameterized Closed-form Ridge Regression". In: *Conference on Learning Theory (COLT)*. 2021, pp. 1063–1102. URL: https://proceedings.mlr.press/v134/cao21a.html.

[21] Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari. "Minimum complexity interpolation in random features models". In: *arXiv preprint arXiv:2103.15996* (2021).

[22] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. "The convex geometry of linear inverse problems". In: *Foundations of Computational mathematics* 12.6 (2012), pp. 805–849.

[23] Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. "Towards understanding hierarchical learning: Benefits of neural representations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22134–22145.

[24] Yunlu Chen, Yang Li, Keli Liu, and Feng Ruan. "Kernel Learning in Ridge Regression" Automatically" Yields Exact Low Rank Solution". In: *arXiv preprint arXiv:2310.11736* (2023).

[25] Theodore S Chihara. *An introduction to orthogonal polynomials*. Courier Corporation, 2011.

[26] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*. Vol. 24. Cambridge University Press, 2007.

[27] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. "Neural networks can learn representations with gradient descent". In: *Conference on Learning Theory*. PMLR. 2022, pp. 5413–5452.

[28] Amit Daniely and Eran Malach. "Learning parities with neural networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20356–20365.

[29] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. "Iteratively reweighted least squares minimization for sparse recovery". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 63.1 (2010), pp. 1–38.

[30] Konstantin Donhauser, Mingqi Wu, and Fanny Yang. "How rotational invariance of common kernels prevents generalization in high dimensions". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2804–2814.

[31] David L Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[32] David L Donoho. "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.6 (2006), pp. 797–829.

[33] Noureddine El Karoui. "The spectrum of kernel random matrices". In: *The Annals of Statistics* 38.1 (2010), pp. 1–50. DOI: 10.1214/09-AOS715.

[34] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. "Do we need hundreds of classifiers to solve real world classification problems?" In: *The journal of machine learning research* 15.1 (2014), pp. 3133–3181.

[35] Bertille Follain and Francis Bach. "Enhanced Feature Learning via Regularisation: Integrating Neural Networks and Kernel Methods". In: *arXiv preprint arXiv:2407.17280* (2024).

[36] Massimo Fornasier, Holger Rauhut, and Rachel Ward. "Low-rank matrix recovery via iteratively reweighted least squares minimization". In: *SIAM Journal on Optimization* 21.4 (2011), pp. 1614–1640.

[37] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. "Kernel dimension reduction in regression". In: *The Annals of Statistics* 37.4 (2009), pp. 1871–1905. DOI: 10.1214/08-AOS637.

[38] Rong Ge, Jason D Lee, and Tengyu Ma. "Learning one-hidden-layer neural networks with landscape design". In: *arXiv preprint arXiv:1711.00501* (2017).

[39] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Linearized two-layers neural networks in high dimension". In: *The Annals of Statistics* 49.2 (2021).

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[41] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[42] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems*. 2018, pp. 8571–8580.

[43] Michael I Jordan, Keli Liu, and Feng Ruan. "On the self-penalization phenomenon in feature selection". In: *arXiv preprint arXiv:2110.05852* (2021).

[44] Nirmit Joshi, Theodor Misiakiewicz, and Nati Srebro. "On the complexity of learning sparse functions with statistical and gradient queries". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 103198–103241.

[45] Yiwen Kou, Zixiang Chen, Quanquan Gu, and Sham Kakade. "Matching the statistical query lower bound for k-sparse parity problems with sign stochastic gradient descent". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 113001–113037.

[46] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[47] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[48] Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. "Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 58716–58756.

[49] Yang Li and Feng Ruan. "On the kernel learning problem". In: *arXiv preprint arXiv:2502.11665* (2025).

[50] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. "Learning over-parametrized two-layer neural networks beyond ntk". In: *Conference on learning theory*. PMLR. 2020, pp. 2613–2682.

[51] Tengyuan Liang and Alexander Rakhlin. "Just Interpolate: Kernel 'Ridgeless' Regression Can Generalize". In: *The Annals of Statistics* 48.3 (2020), pp. 1329–1347. DOI: 10.1214/19-AOS1850.

[52] Yishay Mansour. "Learning Boolean functions via the Fourier transform". In: *Theoretical advances in neural computation and learning*. Springer, 1994, pp. 391–424.

[53] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration". In: *Applied and Computational Harmonic Analysis* 59 (2022), pp. 3–84.

[54] Karthik Mohan and Maryam Fazel. "Iterative reweighted algorithms for matrix rank minimization". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 3441–3473.

[55] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. "Neural Networks Efficiently Learn Low-Dimensional Representations with SGD". In: *The Eleventh International Conference on Learning Representations*.

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

[57] Parthe Pandit, Zhichao Wang, and Yizhe Zhu. "Universality of kernel random matrices and kernel regression in the quadratic regime". In: *arXiv preprint arXiv:2408.01062* (2024).

[58] Suzanna Parkinson, Greg Ongie, and Rebecca Willett. "Relu neural networks with linear layers are biased towards single-and multi-index models". In: *arXiv preprint arXiv:2305.15598* (2023).

[59] Fabian Pedregosa and Rachael Tappenden. "On the Convergence Rate of Kernel Ridge Regression with Random Features". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2021, pp. 16259–16270. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/6e8502051cd8b74e2b4a5a44d7275dc4-Paper.pdf.

[60] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. "Mechanism for feature learning in neural networks and backpropagation-free machine learning models". In: *Science* 383.6690 (2024), pp. 1461–1467. DOI: 10.1126/science.adi5639. eprint: https://www.science.org/doi/pdf/10.1126/science.adi5639. URL: https://www.science.org/doi/abs/10.1126/science.adi5639.

[61] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. "Linear Recursive Feature Machines provably recover low-rank matrices". In: *arXiv preprint arXiv:2401.04553* (2024).

[62] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[63] Feng Ruan, Keli Liu, and Michael I Jordan. "Taming Nonconvexity in Kernel Feature Selection–Favorable Properties of the Laplace Kernel". In: *arXiv preprint arXiv:2106.09387* (2021).

[64] Johannes Schmidt-Hieber. "Nonparametric regression using deep neural networks with ReLU activation function". In: *The Annals of Statistics* 48.4 (2020), pp. 1875–1897. DOI: 10.1214/19-AOS1875. URL: https://doi.org/10.1214/19-AOS1875.

[65] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[66] Gabor Szeg. *Orthogonal polynomials*. Vol. 23. American Mathematical Soc., 1939.

[67] Yan Shuo Tan, Jason M Klusowski, and Krishnakumar Balasubramanian. "Statistical-computational trade-offs for greedy recursive partitioning estimators". In: *arXiv preprint arXiv:2411.04394* (2024).

[68] Ryan J Urbanowicz, Jeff Kiralis, Nicholas A Sinnott-Armstrong, Tamra Heberling, Jonathan M Fisher, and Jason H Moore. "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures". In: *BioData mining* 5 (2012), pp. 1–14.

[69] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[70] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027* (2010).

[71] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[72] Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. "An adaptive estimation of dimension reduction space". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3 (2002), pp. 363–410.

[73]   Weihang Xu and Simon Du. "Over-parameterization exponentially slows down gradient descent for learning a single neuron". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 1155–1198.

[74]   Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[75]   Libin Zhu, Damek Davis, Dmitriy Drusvyatskiy, and Maryam Fazel. "Spectral norm bound for the product of random Fourier-Walsh matrices". In: *arXiv preprint arXiv:2504.03148* (2025).

Appendix

Contents

A   Orthogonal polynomials: Isometric properties and kernel approximation

A.1   Isometric properties of orthogonal polynomials

In this section, we establish some isometric properties of covariance-like matrices induced by orthogonal polynomials. Setting the stage, fix a probability space $(D, \mathcal{A}, \mu_d)$ with $D \subset \mathbb{R}^d$ and a set $\{\phi_j\}_{j \in \bar{\mathcal{S}}}$ of

orthonormal polynomials with respect to μ_d, indexed by some set $\bar{\mathcal{S}}$. For any finite set $\mathcal{S} \subset \bar{\mathcal{S}}$ and a point $x \in D$, we define the concatenated vector:

$$\phi_{\mathcal{S}}(x) := (\phi_j(x))_{j \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}.$$

Finally given a sequence of points $X = (x^{(1)}, \ldots, x^{(n)})$ we may stack $\phi_{\mathcal{S}}(x^{(i)})$ as rows to form the matrix

$$\Phi_{\mathcal{S}}(X) = [\phi_j(x^{(i)})] \in \mathbb{R}^{n \times |\mathcal{S}|}.$$

Note that the rows of $\Phi_{\mathcal{S}}(X)$ are indexed by the data points and the columns by the polynomials in $\mathcal{S}$. To simplify notation, we will often omit X from the symbol $\Phi_{\mathcal{S}}(X)$, whenever it is clear from context. Throughout, we will assume that the data points $x^{(1)}, \ldots, x^{(n)}$ are sampled independently from μ.

We impose the following hypercontractivity assumption for the rest of the section, mimicking equation (12) for our two running examples. More precisely, we will assume hypercontractivity of the product measure $\mu_d \times \mu_d$ on $\mathbb{R}^d \times \mathbb{R}^d$. Hypercontractivity of the original measure μ_d then follows trivially.

Assumption 4 (Hypercontractivity). There exist constants $C_{l,q} > 0$ indexed by integers $l, q \in \mathbb{N}$, such that any polynomial f on $\mathbb{R}^d \times \mathbb{R}^d$ of degree at most ℓ satisfies:

$$\|f\|_{L_q(\mu_d \times \mu_d)} \leq C_{\ell,q} \cdot \|f\|_{L_2(\mu_d \times \mu_d)} \qquad \forall q \geq 2.$$

In particular, Assumption 4 holds in our two running examples—Gaussian and uniform on hypercube—with $C_{l,q} = (q-1)^{l/2}$, since the product measure of Gaussians (respectively uniform on hypercube) is Gaussian (respectively uniform on hypercube). The rest of the section is devoted to establishing isometry properties of the matrix $\Phi_{\mathcal{S}}$ in the high-dimensional regime $n, d \to \infty$. We begin by estimating the operator norm of $\Phi_{\mathcal{S}}$.

Lemma A.11 (Norm control). *Suppose Assumption 4 holds. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant. Fix a set $\mathcal{S} \subseteq \bar{\mathcal{S}}$ indexing polynomials of degree at most l. Then the following are true:*

1. *in the regime $|\mathcal{S}| \leq Cd^{p+\delta_0}$ with $\delta_0 \in (-\infty, \delta)$ the estimate $\|\Phi_{\mathcal{S}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ holds.*

2. *in the regime $|\mathcal{S}| \geq Cd^{p+\delta_0}$ with $\delta_0 \in (\delta, \infty)$ the estimate $\|\Phi_{\mathcal{S}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{|\mathcal{S}|})$ holds.*

The exact asymptotics of $\Phi_{\mathcal{S}}$ depend on the relative scale of n to $|\mathcal{S}|$, meaning:

$$\frac{n}{|\mathcal{S}|} \asymp d^q \text{ with } \underbrace{q > 0}_{\text{Case I}} \text{ or } \underbrace{q < 0}_{\text{Case II}}.$$

The following two theorems show (respectively) that in the first case $\lambda > 0$, the empirical covariance matrix $\frac{1}{n}\Phi_{\mathcal{S}}^{\top}\Phi_{\mathcal{S}}$ is asymptotically equal to the identity $I_{|\mathcal{S}|}$, while in the second case $\lambda < 0$, the Gram matrix $\frac{1}{|\mathcal{S}|}\Phi_{\mathcal{S}}\Phi_{\mathcal{S}}^{\top}$ is asymptotically equal to the identity I_n. Both Lemma A.11 and Theorem A.12 are proved in Appendix C.1.1 and closely follow the proof template of [53, Theorem 6(b)].

Theorem A.12 (Asymptotics in Case I). *Suppose that Assumption 4 holds and consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant. Fix a set $\mathcal{S} \subseteq \bar{\mathcal{S}}$ of cardinality $|\mathcal{S}| \leq Cd^{p+\delta_0}$, with $\delta_0 \in (-\infty, \delta)$, indexing polynomials of degree at most ℓ. Then for any $\delta' \in (0, \delta - \delta_0)$ there exists a constant C' satisfying*

$$\mathbb{E}\left[\left\|\frac{1}{n}\Phi_{\mathcal{S}}^{\top}\Phi_{\mathcal{S}} - I_{|\mathcal{S}|}\right\|_{\mathrm{op}}\right] \leq C'd^{-\delta'/2}\sqrt{\log(n)}.$$

Consequently, with probability at least $1 - \frac{C'}{\sqrt{\log(n)}}$, we have $\left\|\Phi_{\mathcal{S}}(X)^{\top}\Phi_{\mathcal{S}}(X)/n - I_{|\mathcal{S}|}\right\|_{\mathrm{op}} \leq d^{-\delta'/2+\epsilon}$.

Next, we argue that in the complementary setting when n is small compared to $|\mathcal{S}|$, the Gram matrix $\Phi_{\mathcal{S}}\Phi_{\mathcal{S}}^{\top} = [\langle \phi_{\mathcal{S}}(x^{(i)}), \phi_{\mathcal{S}}(x^{(j)})\rangle]_{i,j=1}^{n}$ is close to a multiple of the identity. The proof appears in Appendix C.1.2 and closely follows the proof template of [53, Lemma 3] based on matrix decoupling.

Theorem A.13 (Asymptotics in Case II). *Suppose that Assumption 4 holds and consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant. Fix a set $\mathcal{S} \subseteq \bar{\mathcal{S}}$ of cardinality $|\mathcal{S}| \geq Cd^{p+\delta_0}$, with $\delta_0 \in (\delta, \infty)$, indexing polynomials of degree at most ℓ. Then for any $\epsilon > 0$ there exists a constant C' satisfying:*

$$\mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}|}\Phi_{\mathcal{S}}\Phi_{\mathcal{S}}^{\top} - I_n\right\|_{\mathrm{op}}\right] \leq C'\left(d^{-\frac{p+\delta}{2}+2\epsilon} + \sqrt{\log(n)}\cdot d^{-\frac{\delta_0-\delta}{2}+\epsilon}\right), \tag{41}$$

Consequently, in the case $p \geq 1$, $\delta_0 = 1$, and $\epsilon \in (0, \delta)$ with probability at least $1 - \frac{C'}{\sqrt{\log(n)}}$, we have $\left\|\Phi_{\mathcal{S}}\Phi_{\mathcal{S}}^{\top}/|\mathcal{S}| - I_n\right\|_{\mathrm{op}} \leq d^{-\frac{1-\delta}{2}+\epsilon}$.

A.2 Approximating kernels by orthogonal polynomials

In this section, we discuss approximating kernels by quadratic forms in orthogonal polynomials. Throughout, we fix an inner-product kernel (13) satisfying the following regularity assumption. Note that Assumption 1 implies that both kernels $K(z, x)$ and $K'(z, x)$ satisfy Assumption 5. This is important because later parts of the paper will invoke the results of this section with K replaced by K'.

Assumption 5 (Regularity of the kernel). *There exists a constant $\varepsilon \in (0, 1)$ such that the function g in (13) is Lipschitz continuous on $(-1 - \varepsilon, 1 + \varepsilon)$ and is analytic on $(-\varepsilon, \varepsilon)$.*

Our goal is to approximate $K(x, y)$ by a quadratic form $K(x, y) \approx \phi(x)^{\top} D\phi(y)$ where $\phi\colon \mathbb{R}^d \to \mathbb{R}^p$ has orthonormal polynomials as its coordinate functions. Ideally, D should be a relatively simple matrix (e.g., diagonal). We proceed by first forming a Taylor approximation of g at the origin and forming the approximate kernel:

$$g_m(t) := \sum_{k=0}^{m} \frac{g^{(k)}(0)}{k!}t^k \qquad \text{and} \qquad K_m(x, y) := g_m\left(\frac{\langle x, y\rangle}{d}\right).$$

The following theorem shows that the deviation $K_m(X, X) - K(X, X)$ is asymptotically equivalent to a multiple of the identity under favorable conditions. Namely in the regime, $n = d^{p+\delta}$ with $\delta \in (0, 1/2)$ the right side of (42) tends to zero for any approximation order $m \geq 2p$. The proof appears in Appendix C.1.3. The approximation of kernels has been previously studied in the linear regime (i.e., $n \asymp d$) by [33] and in the quadratic regime (i.e., $n \asymp d^2$) by [57].

Theorem A.14 (Taylor approximation of kernels with Gaussian data). *Consider independent, mean zero, isotropic random vectors $x^{(1)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$ and suppose that the coordinates of each vector $x^{(i)}$ are independent and σ-subGaussian. Suppose that we are in the regime $n = d^{p+\delta}$, where $\delta \in (0, 1/2)$ is a constant. Fix an arbitrary constant $C > 2$ and approximation order $m \geq 2p$. Then if $d/\log(d) > c_\epsilon$, the estimate*

$$\|K(X, X) - K_m(X, X) - (g(1) - g_m(1))I_n\|_{\mathrm{op}} \leq c_{g,m}\left(\sqrt{\frac{(C\log(n))^{m+1}}{d^{m-2p+1-2\delta}}} + \sqrt{\frac{C\log(n)}{d}}\right), \tag{42}$$

holds with probability at least $1 - \frac{4}{n^{C-2}}$, where $c_{g,m}, c_\epsilon < \infty$ are constants that depend only on m, $\sup_{k \geq m+1} g^{(k)}(0)$, σ^2, ϵ, and the Lipschitz constant of g.

In light of Theorem A.14, we may now approximate the kernel matrix by $K_m(X, X)$ plus a multiple of the identity. Note that each entry of $K_m(X, X)$ is linear in the powers of inner products $(\langle x^{(i)}, x^{(j)}\rangle/d)^k$. It remains therefore to express these powers as a quadratic form in an orthogonal polynomial basis. We do so in our two running examples in the following two sections.

37

A.2.1   Uniform measure on the hypercube

In this section, we focus on the uniform measure τ_d over the hypercube $\mathbb{H}^d = \{-1, 1\}^d$ as discussed in Section 3.1. We consider the Fourier orthogonal basis $\phi_\lambda := x^\lambda$ for each $\lambda \in \{0, 1\}^d$. As usual, for any set $\mathcal{S} \subset \{0, 1\}^d$, we define the concatenated vector $\phi_S(x) = \{\phi_\lambda(x)\}_{\lambda \in S} \in \mathbb{R}^{|\mathcal{S}|}$. Stacking the vectors $\Phi_\lambda(x^{(i)})$ as rows yields a matrix $\Phi(X) \in \mathbb{R}^{n \times |\mathcal{S}|}$. It will be useful to isolate degree k basis elements:

$$\mathcal{S}_k = \{\lambda \in \{0, 1\}^d : |\lambda| = k\}.$$

We set $\Phi_{\mathcal{S}_k}(X)$ and $\Phi_{\leq k}(X)$ to be, respectively, the submatrices of Φ indexed by degree k and degree at most k basis elements. We will suppress the symbol X from $K(X, X)$ and $\Phi(X)$ throughout the section in order to simplify the notation. We stress that all probabilistic statements are in reference to the random vectors $\{x^{(i)}\}_{i=1}^n$ sampled independently from τ_d.

In parallel to Theorem A.14, we present the Taylor approximation of kernels for data on the hypercube. We provide a proof for a more general parameterized kernel $K_w(x, y) = g(\langle \sqrt{w} \odot x, \sqrt{w} \odot y \rangle / d)$, which encompasses the special case $w = \mathbf{1}_d$ below. The proof appears in Appendix C.4.2.

Theorem A.15 (Taylor approximation of kernels on hypercube). *Consider independent, mean zero, isotropic random vectors $x^{(1)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$ and suppose that the coordinates of each vector $x^{(i)}$ are independent and are uniformly drawn from $\mathbb{H}^d$. Suppose that we are in the regime $n = d^{p+\delta}$, where $\delta \in (0, 1)$ is a constant. Fix an arbitrary constant $C > 2$ and approximation order $m > 2p$. Then, then the estimate*

$$\|K(X, X) - K_m(X, X) - (g(1) - g_m(1))I_n\|_{\mathrm{op}} \leq c_{g,m}\sqrt{\frac{(C\log(n))^{m+1}}{d^{m-2p+1-2\delta}}} \tag{43}$$

holds with probability at least $1 - \frac{4}{n^{C-2}}$, where $c_{g,m} < \infty$ is a constant that depends only on m, $\{g^{(k)}(0)\}_{k=1}^m$, and L_m.

We begin with the following (deterministic) lemma that expresses the matrix $\left(\frac{XX^\top}{d}\right)^{\odot k}$ as a diagonal quadratic form acting on the Fourier basis. We further decompose the diagonal matrix into the dominant part, corresponding to the degree k basis elements, and the small error part, corresponding to the lower-order basis elements.

Lemma A.16 (Conversion from a polynomial kernel to a Fourier basis). *For any $k > 0$, the following holds*

$$\frac{1}{k!}\left(\frac{XX^\top}{d}\right)^{\odot k} = d^{-k} \cdot \Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top + \sum_{\substack{j:\, 0 \leq j < k, \\ k-j \text{ is even}}} \Phi_{\mathcal{S}_j}\widetilde{D}_{\mathcal{S}_j}\Phi_{\mathcal{S}_j}^\top, \tag{44}$$

where each $\widetilde{D}_{\mathcal{S}_j}$ is a diagonal matrix with all entries on the order of $\Theta_d(d^{-(j+k)/2})$.

The following lemma—the main part of the section—shows the asymptotic equivalence of $K(X, X)$ to a sum of a multiple of the identity and a quadratic form in Fourier basis elements of degree at most p. The proof of the lemma appears in Section C.1.5.

Lemma A.17 (Fourier basis approximation of a kernel). *Consider the regime $n = d^{p+\delta}$ where $\delta \in (0, 1)$ is a constant. Then for any $\epsilon \in (0, \delta)$, the following holds:*

$$\left\|K - \Phi_{\leq p}D\Phi_{\leq p}^\top - (g(1) - g_p(1))I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(\frac{\log(n)}{d^{\frac{1-\delta}{2}-\epsilon}}\right), \tag{45}$$

$$\left\|K - \Phi_{\leq p}D\Phi_{\leq p}^\top - \frac{g^{(p+1)}(0)}{d^{p+1}}\mathrm{off}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right) - (g(1) - g_p(1))I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(\frac{\log(n)}{d^{\frac{2-\delta}{2}-\epsilon}}\right), \tag{46}$$

where D is a diagonal matrix satisfying $\|D_{\mathcal{S}_k} - g^{(k)}(0)d^{-k}I_{|\mathcal{S}_k|}\|_{\mathrm{op}} = O_d(d^{-k-1})$ for $k = 0, \ldots, p$.

Note that the estimate (45) allows us to approximate K by the simple expression $\Phi_{\leq p} D \Phi_{\leq p}^\top +$ $(g(1) - g_p(1)) I_n$ up to an error $\approx d^{-(1-\delta)/2}$. In contrast, (46) provides a more sophisticated approximation of K but with a much better error rate $\approx d^{-(2-\delta)/2}$. This improved error rate will be important for some of our arguments.

A.2.2 Hermite expansion of polynomial kernel

Next, we focus on the Gaussian measure γ_d on $\mathbb{R}^d$ as discussed in Section 3.2. We consider the Hermite orthogonal basis $\phi_\alpha = h_\alpha$ for each multi-index $\alpha \in \mathbb{N}^d$. As usual, for any set $\mathcal{S} \subset \mathbb{N}^d$, we define the concatenated vector $\Phi_S(x) = \{\phi_\alpha(x)\}_{\alpha \in S} \in \mathbb{R}^{|\mathcal{S}|}$. Stacking the vectors $\Phi_\lambda(x^{(i)})$ as rows yields a matrix $\Phi(X) \in \mathbb{R}^{n \times |\mathcal{S}|}$. It will be useful to isolate degree k basis elements:

$$\mathcal{S}_k = \left\{ \alpha \in \mathbb{N}^d : \sum_{i=1}^d \alpha_i = k \right\}.$$

We set $\Phi_{\mathcal{S}_k}(X)$ and $\Phi_{\leq k}(X)$ to be, respectively, the submatrices of Φ indexed by degree k and degree at most k basis elements. We will suppress the symbol X from $K(X,X)$ and $\Phi(X)$ throughout the section in order to simplify the notation. All probabilistic statements in this section are in reference to the random vectors $\{x^{(i)}\}_{i=1}^n$ sampled independently from γ_d.

As in the hypercube setting of the previous section, we estimate the matrix $\left(\frac{XX^\top}{d}\right)^{\odot k}$ by a quadratic form acting on the Hermite basis. The distinction is now that this quadratic form is not diagonal in the simplest case of $k = 2$.[6] Namely, in the hypercube case, we saw that $\frac{1}{k!}\left(\frac{XX^\top}{d}\right)^{\odot k}$ is well-approximated by $d^{-k}\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top$. The analogous statement is not true in the Gaussian setting and instead the correct approximation for $k = 2$ takes the form in Lemma A.18. The proof of the lemma is elementary and appears in Section C.1.6. In the lemma and in the rest of the section, we will order the second-order Hermite polynomials $\phi_{\mathcal{S}_2}$ so that the pure second-order parts $\{h_{2e_i}\}_{i=1}^d$ appear first and the mixed parts $\{h_{e_i+e_j}\}_{1 \leq i < j \leq d}$ appear second.

Lemma A.18 (Conversion from quadratic kernel to a Hermite basis). *The equality holds:*

$$\frac{1}{2}\left(\frac{XX^\top}{d}\right)^{\odot 2} = \Phi_{\leq 2} \cdot \begin{pmatrix} \frac{1}{2d} & 0 & D_{13}^\top \\ 0 & 0 & 0 \\ D_{13} & 0 & \frac{1}{d^2}I_{|\mathcal{S}_2|} \end{pmatrix} \cdot \Phi_{\leq 2}^\top, \tag{47}$$

where we define the matrix $D_{13} := (\frac{1}{\sqrt{2d^2}}\mathbf{1}_d; 0_{|\mathcal{S}_2|-d}) \in \mathbb{R}^{|\mathcal{S}_2|}$.

The case $k = 2$, described in Lemma A.18, appears to be exceptional. Indeed, the expansions of $\left(\frac{XX^\top}{d}\right)^{\odot k}$ for the larger values of $k = 3, 4$ are diagonal, as the following lemma shows, and we conjecture this to be true for all $k > 4$. We note, however, that the combinatorial argument we use in the case $k = 3, 4$ appears to be difficult to generalize. The proof of Lemma A.19 appears in Appendix C.1.7.

Lemma A.19 (Conversion from a polynomial kernel to a Hermite basis for $k = 3, 4$). *Consider the regime* $n = d^{2+\delta}$ *where* $\delta \in (0, 1/2)$. *Then the following estimates hold:*

$$\left\| \frac{1}{3!}\left(\frac{XX^\top}{d}\right)^{\odot 3} - \frac{1}{d^3}\Phi_{\mathcal{S}_3}\Phi_{\mathcal{S}_3}^\top - \frac{1}{2d^2}\Phi_{\mathcal{S}_1}\Phi_{\mathcal{S}_1}^\top \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta-1/2}), \tag{48}$$

$$\left\| \frac{1}{4!}\left(\frac{XX^\top}{d}\right)^{\odot 4} - \frac{1}{d^4}\Phi_{\mathcal{S}_4}\Phi_{\mathcal{S}_4}^\top - \frac{1}{4d^2}\Phi_{\mathcal{S}_0}\Phi_{\mathcal{S}_0}^\top \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta-1/2}). \tag{49}$$

---

[6]Note that the claim in [8] that the quadratic form in the Hermite basis is diagonal is inaccurate as can be seen from (47).

The following lemma—the main result of the section—parallels Lemma A.17 for the uniform distribution on the hypercube. We note, however, the key difference is that the polynomial kernel is well-approximated by a sum of the identity and a quadratic form in the Hermite basis with a non-diagonal arrowhead matrix. The proof appears in Appendix C.1.8.

Lemma A.20 (Kernel to Hermite conversion). *Consider the regime $n = d^{2+\delta}$ where $\delta \in (0, 1/2)$ is a constant. Then for any $\delta_1 \in (\delta, 1/2)$, the following holds:*

$$\left\| K - \Phi_{\leq 2} D \Phi_{\leq 2}^\top - (g(1) - g_2(1)) I_n \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta_1 - 1/2}\right), \tag{50}$$

where

$$D = \begin{pmatrix} g(0) + \frac{g^{(2)}(0)}{2d} + \frac{g^{(4)}(0)}{4d^2} & 0 & D_{13}^\top \\ 0 & \left(\frac{g^{(1)}(0)}{d} + \frac{g^{(3)}(0)}{2d^2}\right) I_{|\mathcal{S}_1|} & 0 \\ D_{13} & 0 & \frac{1}{d^2}g^{(2)}(0)I_{|\mathcal{S}_2|} \end{pmatrix}, \tag{51}$$

and $D_{13} = \left(\frac{g^{(2)}(0)}{\sqrt{2}d^2} \mathbf{1}_d; \mathbf{0}_{|\mathcal{S}_2|-d}\right) \in \mathbb{R}^{|\mathcal{S}_2|}$.

B Proofs of main results

B.1 Proof of Theorem 4.5

We prove the theorem only in the case $p = 2$, since the case $p = 1$ follows by similar (and simpler) arguments. We analyze each coordinate $r = 1, \ldots, d$ individually. For simplicity of notation, we omit the input X in $K(X, X)$ and $K'(X, X)$ throughout and set $K_\lambda = K + \lambda I$. Summing (15) with $z = x^{(i)}$ yields the convenient expression:

$$\frac{1}{n}\sum_{i=1}^n \left(\partial_r f_{\mathrm{KRR}}(x^{(i)})\right)^2 = \frac{1}{nd^2}\|K'X_r K_\lambda^{-1}y\|_2^2, \tag{52}$$

where $X_r \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entry $(X_r)_{i,i} = x_r^{(i)}$ for $i \in [d]$. By the assumption $f^* \in L_2(\gamma)$, we can expand f^* into Hermite polynomials:

$$y_i = f^*(x^{(i)}) = \phi(x^{(i)})^\top b = \sum_{\lambda \in \mathbb{N}^d} b_\lambda \phi_\lambda(x^{(i)}), \tag{53}$$

for some coefficients b_λ satisfying $\|b\|_2 < \infty$. Since $\mathbb{E}[y_i^2] = \mathbb{E}|f^*|^2 + \sigma_\varepsilon^2$ we deduce the estimate $\mathbb{E}[\|y\|_2^2] = n(\mathbb{E}|f^*|^2 + \sigma_\varepsilon^2) = n(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$. Markov's inequality subsequently shows $\|y\|^2 = O_{d,\mathbb{P}}(n\log(n) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2))$. We summarize this observation in the following proposition.

Proposition B.21. *With probability at least $1 - 1/\log(n)$ we have $\|y\|^2 \leq n\log(n)(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$.*

Applying Lemma A.20 with the two Kernels K and K' shows that for any $\epsilon > 0$ there exist matrices Δ, Δ_1 satisfying $\|\Delta\|_{\mathrm{op}}, \|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon})$ and

$$K = \Phi_{\leq 2} D \Phi_{\leq 2}^\top + \rho I_n + \Delta_1, \tag{54}$$
$$K' = \Phi_{\leq 2} D' \Phi_{\leq 2}^\top + \rho' I_n + \Delta, \tag{55}$$

where we define $\rho := g(1) - g_2(1)$ and $\rho' := g'(1) - g_3(1)$. With the expression (55) in place of K', equation (52) becomes

$$\frac{1}{n}\sum_{i=1}^n \left(\partial_r f_{\mathrm{KRR}}(x^{(i)})\right)^2 = \frac{1}{nd^2} \left\| \left(\underbrace{\Phi_{\leq 2} D' \Phi_{\leq 2}^\top X_r K_\lambda^{-1}}_{=:T_1} + \underbrace{(\rho' I_n + \Delta)X_r K_\lambda^{-1}}_{=:T_2}\right) y \right\|_2^2.$$

Expanding the square, we conclude

$$\frac{1}{n}\sum_{i=1}^{n}\left(\partial_r f_{\mathrm{KRR}}(x^{(i)})\right)^2 = \frac{1}{nd^2}\left(\|T_1 y\|_2^2 + \|T_2 y\|_2^2 + 2\langle T_1 y, T_2 y\rangle\right). \tag{56}$$

We will analyze each summand separately. Specifically, we will verify the following claim, from which Theorem 4.5 follows immediately.

Claim 1. *The following estimates hold for any $\epsilon > 0$ uniformly over all coordinates r:*

(a) $\frac{1}{nd^2}\|T_2 y\|_2^2 = O_{d,\mathbb{P}}\left(d^{-2+\epsilon}\right)\cdot\left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2\right)$

(b) $\frac{1}{nd^2}\|T_1 y\|_2^2 = \mathbb{E}\left[\left(\partial_r f_{\leq 2}^*(x)\right)^2\right] + O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right)\cdot\|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right)\cdot\|f^*_{>2}\|_{L_{2+\eta}}^2$
$+ O_{d,\mathbb{P}}(d^{\delta-1/2+\epsilon})\cdot\sigma_\varepsilon^2,$

(c) $\frac{1}{nd^2}\langle T_1 y, T_2 y\rangle = O_{d,\mathbb{P}}(d^{-1+\epsilon/2})\cdot\|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{\delta/2-5/4+\epsilon})\cdot\sigma_\varepsilon^2 + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right)\cdot\|f^*_{>2}\|_{L_{2+\eta}}^2.$

The rest of the section is devoted to proving Claim 1. Note that Item (c) follows immediately from Item (a) and (b) and the Cauchy-Schwartz inequality. Throughout the proof, all asymptotic terms $o_{d,\mathbb{P}}$ and $O_{d,\mathbb{P}}$ that appear will always be uniform over the coordinates $r = 1, \ldots, d$.

B.1.1 Proof of Item (a) of Claim 1

Sub-multiplicativity of the operator norm directly implies

$$\|T_2 y\| \leq (|\rho'| + \|\Delta\|_{\mathrm{op}})\|X_r\|_{\mathrm{op}}\left\|K_\lambda^{-1}y\right\|_2.$$

Then the expression (54) directly implies

$$\left\|K_\lambda^{-1}\right\|_{\mathrm{op}} \leq 1/(\rho + \lambda - \|\Delta_1\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(1). \tag{57}$$

where we use the fact that D is PSD since D is diagonally dominant. Therefore, taking into account Proposition B.21, we deduce

$$\left\|K_\lambda^{-1}y\right\|_2 \leq \left\|K_\lambda^{-1}\right\|_{\mathrm{op}}\|y\|_2 \leq O_{d,\mathbb{P}}\left(\sqrt{n\log(n)\left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2\right)}\right) \tag{58}$$

Since the diagonal of X_r has the same distribution as a standard Gaussian vector $v^{(r)}$ in $\mathbb{R}^n$, we have

$$\sup_r\|X_r\|_{\mathrm{op}} = \sup_r\|v^{(r)}\|_\infty = \max_{i,r}|v_i^{(r)}| = O_{d,\mathbb{P}}(\sqrt{\log(nd)}). \tag{59}$$

Consequently, we conclude $\frac{1}{nd^2}\|T_2 y\|_2^2 = O_{d,\mathbb{P}}\left(\frac{\log^2(d)}{d^2}\cdot\left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2\right)\right) = O_{d,\mathbb{P}}\left(d^{-2+\epsilon}\cdot\left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2\right)\right)$ for any $\epsilon > 0$ as desired.

B.1.2 Proof of Item (b) of Claim 1

We first simplify $T_1 T_1^\top$. By Theorem A.12, there exists a matrix $\widetilde{\Delta}$ satisfying

$$\frac{1}{n}\Phi_{\leq 2}^\top\Phi_{\leq 2} = I + \widetilde{\Delta} \qquad \text{and} \qquad \left\|\widetilde{\Delta}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon}\log(d)).$$

Using this approximation and defining $\beta = K_\lambda^{-1}y$, we therefore expand the square to obtain

$$\frac{1}{nd^2}\|T_1 y\|^2 = \frac{1}{nd^2}\beta^\top[X_r\Phi_{\leq 2}D'\Phi_{\leq 2}^\top\Phi_{\leq 2}D'\Phi_{\leq 2}^\top X_r]\beta$$
$$= \frac{1}{d^2}\beta^\top[X_r\Phi_{\leq 2}D'(I + \widetilde{\Delta})D'\Phi_{\leq 2}^\top X_r]\beta$$
$$= \frac{1+O(\|\widetilde{\Delta}\|_{\mathrm{op}})}{d^2}\cdot\|\beta^\top X_r\Phi_{\leq 2}D'\|_2^2. \tag{60}$$

41

Next, we argue that the contribution in this expression due to second-order terms $\Phi_{\mathcal{S}_2}$ is negligible. Namely, let us decompose $\Phi_{\leq 2}$ and D' into blocks as follows:

$$\Phi_{\leq 2} = [\Phi_{\leq 1}, \Phi_{\mathcal{S}_2}] \qquad \text{and} \qquad D' = \begin{bmatrix} D'_{\leq 1} \\ D'_{\mathcal{S}_2} \end{bmatrix} = \begin{bmatrix} D'_{\leq 1, \leq 1} & D'_{\leq 1, \mathcal{S}_2} \\ D'_{\mathcal{S}_2, \leq 1} & D'_{\mathcal{S}_2, \mathcal{S}_2} \end{bmatrix}.$$

Using this decomposition, we successively compute

$$\begin{aligned}
\frac{1}{d^2} \|\beta^\top X_r \Phi_{\leq 2} D'\|_2^2 &= \frac{1}{d^2} \|\beta^\top X_r \left(\Phi_{\leq 1} D'_{\leq 1} + \Phi_{\mathcal{S}_2} D'_{\mathcal{S}_2} \right)\|_2^2 \\
&= \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1} \right\|_2^2 + \left\| \frac{1}{d} \beta^\top X_r \Phi_{\mathcal{S}_2} D'_{\mathcal{S}_2} \right\|_2^2 \\
&\quad + \frac{2}{d^2} \langle \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1}, \beta^\top X_r \Phi_{\mathcal{S}_2} D'_{\mathcal{S}_2} \rangle.
\end{aligned} \tag{61}$$

We will show that only the first term on the right side of (61) does not vanish as d tends to infinity. To this end, the following proposition shows that the second term on the right of (61) is negligible. The proof appears in Appendix C.2.1.

Proposition B.22. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$\left\| \frac{1}{d} \beta^\top X_r \Phi_{\mathcal{S}_2} D'_{\mathcal{S}_2} \right\|_2^2 = O_{d,\mathbb{P}}(d^{2\delta - 1 + \epsilon}) \cdot \left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2 \right).$$

In order to show that the last term on the right side of (61) is negligible, it will be useful to control the first term and then apply Cauchy-Schwarz. To this end, we decompose the first term as

$$\left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1} \right\|_2^2 = \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} \right\|_2^2 + \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \mathcal{S}_2} \right\|_2^2. \tag{62}$$

The following proposition shows that the second term on the right side of (62) is negligible. The proof appears in Appendix C.2.2

Proposition B.23. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$\left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \mathcal{S}_2} \right\|_2^2 = O_{d,\mathbb{P}}(d^{2\delta - 1 + \epsilon}) \cdot \left(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2 \right). \tag{63}$$

Thus, it remains to analyze the first term on the right side of (62). This is the main part of the entire argument. The proof of the proposition appears in Appendix C.2.3.

Proposition B.24. *For any $\epsilon > 0$, the following estimate holds uniformly over all coordinates r:*

$$\begin{aligned}
\left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} \right\|_2^2 &= \mathbb{E}\left[(\partial_r f^*_{\leq 2}(x))^2 \right] + O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon} \right) \cdot \|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}\left(d^{-\delta/2} \right) \cdot \|f^*_{>2}\|_{L_{2+\eta}}^2 \\
&\quad + O_{d,\mathbb{P}}(d^{-(\delta+1)/2 + \epsilon}) \cdot \sigma_\varepsilon^2.
\end{aligned}$$

In particular, Propositions B.22 and B.24 together with the Cauchy Schwartz inequality imply that the last term in (61) is bounded by $O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon/2}) \cdot \left(\|f^*\|_{L_2} + \sigma_\varepsilon \right) \cdot \left\| \partial_r f^*_{\leq 2} \right\|_{L_2}$ and therefore we arrive at the bound

$$\begin{aligned}
\frac{1}{d^2} \|\beta^\top X_r \Phi_{\leq 2} D'\|_2^2 &= \mathbb{E}\left[(\partial_r f^*_{\leq 2}(x))^2 \right] + O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon} \right) \cdot \|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}\left(d^{-\delta/2} \right) \cdot \|f^*_{>2}\|_{L_{2+\eta}}^2 \\
&\quad + O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon}) \cdot \sigma_\varepsilon^2,
\end{aligned}$$

where we apply AM-GM inequality for the cross terms.

Lastly, combining this expression with (60) completes the proof.

B.2 Proof of Theorem 4.6

The proof follows along similar lines as the proof of Theorem 4.5. Therefore, we will omit some of the common steps in the argument when they are clear and focus mostly on the key differences.

By the assumption $f^* \in L_2(\tau)$ where τ is the uniform measure over the hypercude $\mathbb{H}^d$, we can expand f^* into Fourier orthogonal basis:

$$y_i = f^*(x^{(i)}) = \phi(x^{(i)})^\top b = \sum_{\lambda \in \mathbb{N}^d} b_\lambda \phi_\lambda(x^{(i)}), \tag{64}$$

for some coefficients b_λ satisfying $\|b\|_2 < \infty$.

First, observe that equation (52) and Proposition B.21 hold verbatim. Applying Lemma A.17 with the two kernels K and K' shows that for any $\epsilon > 0$ there exist matrices Δ, Δ_1 satisfying $\|\Delta\|_{\mathrm{op}}, \|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon})$ and

$$K = \Phi_{\leq p} D \Phi_{\leq p}^\top + \rho I_n + \Delta_1, \tag{65}$$

$$K' = \Phi_{\leq p} D' \Phi_{\leq p}^\top + \rho' I_n + \Delta, \tag{66}$$

where we define $\rho := g(1) - g_p(1)$ and $\rho' := g'(1) - g_{p+1}(1)$, and the diagonal matrices D and D' satisfy $\|D_{\mathcal{S}_k} - g^{(k)}(0)d^{-k} I_{|\mathcal{S}_k|}\|_{\mathrm{op}} = O_d(d^{-k-1})$ and $\|D'_{\mathcal{S}_k} - g^{(k+1)}(0)d^{-k} I_{|\mathcal{S}_k|}\|_{\mathrm{op}} = O_d(d^{-k-1})$ for $k = 0, \ldots, p$.

With the expression (66) in place of K', equation (52) becomes

$$\frac{1}{n} \sum_{i=1}^n \left(\partial_r \hat{f}(x^{(i)})\right)^2 = \frac{1}{nd^2} \left\| \left(\underbrace{\Phi_{\leq p} D' \Phi_{\leq p}^\top X_r K_\lambda^{-1}}_{=:T_1} + \underbrace{(\rho' I_n + \Delta) X_r K_\lambda^{-1}}_{=:T_2} \right) y \right\|_2^2.$$

Expanding the square, we conclude

$$\frac{1}{n} \sum_{i=1}^n \left(\partial_r \hat{f}(x^{(i)})\right)^2 = \frac{1}{nd^2} \left(\|T_1 y\|_2^2 + \|T_2 y\|_2^2 + 2\langle T_1 y, T_2 y \rangle \right). \tag{67}$$

We verify the following claim from which Theorem 4.6 follows immediately.

Claim 2. *The following estimates hold for any $\epsilon > 0$ uniformly over all coordinates r:*

(a) $\frac{1}{nd^2} \|T_2 y\|_2^2 \leq O_{d,\mathbb{P}}(d^{-2+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$

(b) $\left| \frac{1}{nd^2} \|T_1 y\|_2^2 - \|\partial_r f_{\leq p}^*\|_{L_2}^2 - \frac{d^{2\delta-2} \cdot \left(g^{(p+1)}(0)\right)^2}{(\rho+\lambda)^2} \cdot \|\partial_r f_{p+1}^*\|_{L_2}^2 \right|$

$\leq O_{d,\mathbb{P}}(1) \cdot \left(\frac{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2}{d} + \frac{\|\partial_r f_{\leq p}^*\|_{L_2}^2 + d^{2\delta-2} \|\partial_r f_{p+1}^*\|_{L_2}^2}{d^{\delta/2-\epsilon} + d^{(1-\delta)/2-\epsilon}} + \frac{\|f^*\|_{L_2} \|\partial_r f_{\leq p}^*\|_{L_2}}{d^{1/2}} + \frac{(\|f^*\|_{L_2} + \sigma_\varepsilon) \|\partial_r f_{p+1}^*\|_{L_2}}{d^{3/2-\delta}} \right).$

(c) $\frac{1}{nd^2} |\langle T_1 y, T_2 y \rangle|$

$\leq O_{d,\mathbb{P}}(1) \cdot \left(\frac{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2}{d^{3/2-\epsilon}} + \frac{\|f^*\|_{L_2} \cdot \|\partial_r f_{\leq p}^*\|_{L_2}}{d^{1-\epsilon}} + \frac{(\|f^*\|_{L_2} + \sigma_\varepsilon) \cdot \|\partial_r f_{p+1}^*\|_{L_2}}{d^{2-\delta-\epsilon}} \right).$

Plugging Claim 2 (a,b,c) into Eq. (67) yields:

$$\left| \mathbb{E}_n(\partial_r \hat{f})^2 - \mathbb{E}\left(\partial_r f_{\leq p}^*\right)^2 - \frac{d^{2\delta-2} \cdot \left(g^{(p+1)}(0)\right)^2}{(\rho+\lambda)^2} \mathbb{E}(\partial_r f_{p+1}^*)^2 \right|$$

$$\leq O_{d,\mathbb{P}}(1) \cdot \left(\frac{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2}{d} + \frac{\|\partial_r f_{\leq p}^*\|_{L_2}^2 + d^{2\delta-2} \|\partial_r f_{p+1}^*\|_{L_2}^2}{d^{\delta/2-\epsilon} + d^{(1-\delta)/2-\epsilon}} + \frac{\|f^*\|_{L_2} \|\partial_r f_{\leq p}^*\|_{L_2}}{d^{1/2}} + \frac{(\|f^*\|_{L_2} + \sigma_\varepsilon) \cdot \|\partial_r f_{p+1}^*\|_{L_2}}{d^{3/2-\delta}} \right).$$

The rest of the section is devoted to proving Claim 2. Note that Item (c) follows immediately from Items (a) and (b) and the Cauchy-Schwartz and Young's inequality. Throughout the proof, all asymptotic terms $o_{d,\mathbb{P}}$ and $O_{d,\mathbb{P}}$ that appear will always be uniform over the coordinates $r = 1, \ldots, d$.

B.2.1 Proof of Item (a) of Claim 2

The proof is identical to the proof of Item (a) of Claim 1 with (59) replaced by the trivial bound $\sup_r \|X_r\|_{\mathrm{op}} \le 1$. We refer the readers to Section B.1.1.

B.2.2 Proof of Item (b) of Claim 2

Similar to Eq. (60), we can write

$$
\begin{aligned}
\frac{1}{nd^2}\|T_1 y\|^2 &= \frac{1}{nd^2}\beta^\top [X_r \Phi_{\le p} D' \Phi_{\le p}^\top \Phi_{\le p} D' \Phi_{\le p}^\top X_r]\beta \\
&= \frac{1}{d^2}\beta^\top [X_r \Phi_{\le p} D' (I + \widetilde{\Delta}) D' \Phi_{\le p}^\top X_r]\beta \\
&= \frac{1 + O(\|\widetilde{\Delta}\|_{\mathrm{op}})}{d^2} \cdot \|\beta^\top X_r \Phi_{\le p} D'\|_2^2,
\end{aligned}
\tag{68}
$$

where $\left\|\widetilde{\Delta}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2 + \epsilon})$.

We next decompose $\left\|\frac{1}{d}\beta^\top X_r \Phi_{\le p} D'\right\|_2^2$. First notice that the following holds for any set $S \subseteq [d]$:

$$
x_r x^S = \begin{cases} x^{S \cup \{r\}}, & \text{if } r \notin S; \\ x^{S/r}, & \text{otherwise.} \end{cases}
$$

Define now the two sets:

$$
\mathcal{S}_j^0 = \{S \in \mathcal{S}_j : r \notin S\} \qquad \text{and} \qquad \mathcal{S}_j^1 = \{S \in \mathcal{S}_j : r \in S\}.
\tag{69}
$$

It follows that every column of $X_r \Phi_{\le p}$ has the form ϕ_S for some $S \in \left(\bigsqcup_{j=0}^{p-1} S_j^0\right) \cup \left(\bigsqcup_{j=1}^{p+1} S_j^1\right)$. Consequently, there exist diagonal matrices R_j^0 and R_j^1 such that the following holds:

$$
\left\|\beta^\top X_r \Phi_{\le p}\left(\frac{1}{d}D'\right)\right\|_2^2 = \sum_{j=0}^{p-1} \|\beta^\top \Phi_{\mathcal{S}_j^0} R_j^0\|_2^2 + \sum_{j=1}^{p+1} \|\beta^\top \Phi_{\mathcal{S}_j^1} R_j^1\|_2^2.
\tag{70}
$$

Claim 3. *The following estimate holds:*

$$
\left\|R_j^0 - g^{(j+2)}(0)d^{-j-2}\cdot I\right\|_{\mathrm{op}} = O_d(d^{-j-3}), \quad \text{for } j = 0,\ldots,p-1,
\tag{71}
$$

$$
\left\|R_j^1 - g^{(j)}(0)d^{-j}\cdot I\right\|_{\mathrm{op}} = O_d(d^{-j-1}), \quad \text{for } j = 1,\ldots,p+1.
\tag{72}
$$

Proof. Recall that D' is a diagonal matrix with entry $D'_{S,S} = g^{(|S|+1)}(0)d^{-|S|} + O_d(d^{-|S|-1})$. It is straightforward to see that R_j^0 is a diagonal matrix with entry $g^{(j+2)}(0)d^{-j-2} + O_d(d^{-j-3})$, which establishes (71). Similar reasoning yeilds Eq. (72). $\square$

We now group all the p-th order terms in (70). Namely, define the block-diagonal matrix R with blocks $R_{\mathcal{S}_j^1, \mathcal{S}_j^1} = R_j^1$ for $j \le p$, $R_{\mathcal{S}_j^0, \mathcal{S}_j^0} = R_j^0$ for $j \le p-1$, and $R_{\mathcal{S}_p^0, \mathcal{S}_p^0} = 0$. Then equation (70) becomes

$$
\left\|\beta^\top X_r \Phi_{\le p}\left(\frac{1}{d}D'\right)\right\|_2^2 = \left\|\beta^\top \Phi_{\le p} R\right\|_2^2 + \left\|\beta^\top \Phi_{\mathcal{S}_{p+1}^1} R_{p+1}^1\right\|_2^2.
\tag{73}
$$

We now estimate the first term on the right side of (73). The proof of the proposition appears in Appendix C.3.1.

44

Proposition B.25. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$
\left| \left\| \beta^\top \Phi_{\leq p} R \right\|_2^2 - \left\| \partial_r f_{\leq p}^* \right\|_{L_2}^2 \right|
$$
$$
= O_{d,\mathbb{P}}\left(d^{-1} \right) \cdot \left(\| f^* \|_{L_2}^2 + \sigma_\varepsilon^2 \right) + O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\| \partial_r f_{\leq p}^* \right\|_{L_2}^2
$$
$$
+ O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \left\| \partial_r f_{\leq p}^* \right\|_{L_2} \| f^* \|_{L_2}. \tag{74}
$$

We next analyze the second term on the right side of (73). We defer the proof to Appendix C.3.2.

Proposition B.26. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$
\left| \left\| \beta^\top \Phi_{\mathcal{S}_{p+1}^1} R_{p+1}^1 \right\|_2^2 - \frac{d^{2\delta-2} \cdot \left(g^{(p+1)}(0) \right)^2}{(\rho + \lambda)^2} \left\| \partial_r f_{p+1}^* \right\|_{L_2}^2 \right|
$$
$$
= O_{d,\mathbb{P}}(d^{-1}) \cdot (\| f^* \|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{\frac{3}{2}\delta-2+\epsilon} + d^{\frac{5}{2}(\delta-1)+\epsilon}) \cdot \left\| \partial_r f_{p+1}^* \right\|_{L_2}^2
$$
$$
+ O_{d,\mathbb{P}}(d^{\delta-3/2}) \cdot \left\| \partial_r f_{p+1}^* \right\|_{L_2} \cdot (\| f^* \|_{L_2} + \sigma_\varepsilon). \tag{75}
$$

Propositions B.25 and B.26 together give the bound

$$
\left| \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq p} D' \right\|_2^2 - \left\| \partial_r f_{\leq p}^* \right\|_{L_2}^2 - \frac{d^{2\delta-2} \cdot \left(g^{(p+1)}(0) \right)^2}{(\rho + \lambda)^2} \left\| \partial_r f_{p+1}^* \right\|_{L_2}^2 \right|
$$
$$
= O_{d,\mathbb{P}}(d^{-1})(\| f^* \|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left(\left\| \partial_r f_{\leq p}^* \right\|_{L_2}^2 + d^{2\delta-2} \left\| \partial_r f_{p+1}^* \right\|_{L_2}^2 \right)
$$
$$
+ O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \left\| \partial_r f_{\leq p}^* \right\|_{L_2} \| f^* \|_{L_2} + O_{d,\mathbb{P}}(d^{\delta-3/2}) \cdot \left\| \partial_r f_{p+1}^* \right\|_{L_2} \cdot (\| f^* \|_{L_2} + \sigma_\varepsilon). \tag{76}
$$

Combining this expression with (68) completes the proof.

B.3   One-step learning progress and proof of Theorem 4.9

We state the paper's main result, which characterizes each step's learning progress in the implementation of IRKM (Algorithm 1). This result generalizes Theorem 4.6, which can be regarded as the first step of IRKM where $w = \mathbf{1}_d$. The proof of Theorem 4.9 follows directly from the result of this theorem.

We begin with some notation. Given a nonnegative vector $w \in \mathbb{R}_+^d$, we define the parameterized kernel K_w by the expression

$$
K_w(x, y) = g\left(\frac{\langle \sqrt{w} \odot x, \sqrt{w} \odot y \rangle}{d} \right). \tag{77}
$$

The KRR prediction function induced by K_w then is given by

$$
\hat{f}_w(z) = K_w(z, X)\beta \qquad \text{where} \qquad \beta := \left(K_w(X, X) + \lambda I_n \right)^{-1} y. \tag{78}
$$

We now define the active coordinates for a weight vector w as follows.

Definition 4 (Active indices). Consider a weight vector $w \in \mathbb{R}_+^d$ with $\|w\|_1 = d$ and fix two vectors $\boldsymbol{s}, \boldsymbol{\gamma} \in \mathbb{R}_{\geq 0}^N$. A partition $\pi_N = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$ of $[d]$ is called $(\boldsymbol{s}, \boldsymbol{\gamma})$-*active* for w if equalities $|\mathcal{T}_k| = \Theta_d(d^{s_k})$ hold and

$$
w_r = \Theta_d(d^{\gamma_k}) \quad \forall r \in \mathcal{T}_k.
$$

We define the function $\pi_N : [d] \to N$ that maps each element to its corresponding partition index

$$
\pi_N(r) = \iota \iff r \in \mathcal{T}_\iota. \tag{79}
$$

45

Given a tuple $\boldsymbol{T} = (T_1, T_2, \ldots, T_N)$ where $T_k \in 2^{\mathcal{T}_k}$ for $k \in [N]$, we define its corresponding Fourier feature

$$\phi_{\boldsymbol{T}} := \phi_{\cup_{k=1}^N T_k} = \prod_{k=1}^N \underbrace{\prod_{j \in T_k} x_j}_{:=\phi_{T_k}}. \tag{80}$$

Then we proceed to define the effective degree of $\phi_{\boldsymbol{T}}$:

Definition 5 (Effective degree). *Let $\pi_N = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$ be $(\boldsymbol{s}, \boldsymbol{\gamma})$-active for vector $w \in \mathbb{R}_+^d$. For any tuple $\boldsymbol{T} = (T_1, T_2, \ldots, T_N)$ where $T_k \in 2^{\mathcal{T}_k}$ for $k \in [N]$, the effective degree of its corresponding Fourier feature $\phi_{\boldsymbol{T}}$ with respect to w is*

$$p_{\text{eff}}(\boldsymbol{T}) := \sum_{k=1}^N (1 - \gamma_k) \cdot |T_k|. \tag{81}$$

Now we are ready to present the theorem. We defer the proof to Appendix B.5 and B.6 for the empirical weights estimator (87) and DN-estimator (88), respectively. The proof for the generalization bound (86) appears in Appendix B.7.

Theorem B.27 (One-step learning progress). *Suppose that Assumptions 1 and 3 hold. Consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant. Consider a weight vector $w \in \mathbb{R}_+^d$ satisfying $\|w\|_1 = d$ and let $\pi_N = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$ be $(\boldsymbol{s}, \boldsymbol{\gamma})$-active for w. Suppose moreover for all $k \in [N]$, equality $s_k + \gamma_k = 1$ holds if and only if $s_k = 1$. Define the family of sets*

$$\mathcal{A}_{<p_\delta} := \left\{ \boldsymbol{T} = (T_1, T_2, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k} \text{ and } p_{\text{eff}}(\boldsymbol{T}) < p + \delta \right\}, \tag{82}$$

$$\mathcal{A}_{=p_\delta} := \left\{ \boldsymbol{T} = (T_1, T_2, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k} \text{ and } p_{\text{eff}}(\boldsymbol{T}) = p + \delta \right\}, \tag{83}$$

$$\mathcal{A}_{p_\delta}^+(r) := \left\{ \boldsymbol{T} = (T_1, T_2, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k} \text{ and } p_{\text{eff}}(\boldsymbol{T}) \in (p + \delta, p + \delta + 1 - \gamma_{\pi_N(r)}] \right\}, \tag{84}$$

for $r \in [d]$, and the truncation of the target function

$$f_{\boldsymbol{T}}^* := b_{\boldsymbol{T}} \phi_{\boldsymbol{T}}, \quad f_{\mathcal{A}}^* := \sum_{\boldsymbol{T} \in \mathcal{A}} f_{\boldsymbol{T}}^*, \quad \text{for} \quad \mathcal{A} \in \{\mathcal{A}_{<p_\delta}, \mathcal{A}_{=p_\delta}, \mathcal{A}_{p_\delta}^+(r)\}.$$

Given $\boldsymbol{T}$, define a function

$$\kappa(\boldsymbol{T}) := p_{\text{eff}}(\boldsymbol{T}) - (p + \delta). \tag{85}$$

Define now the effective sampling power $\widetilde{p} = \min(\ell, \lfloor \frac{p+\delta}{1-\max_{k \in [N]} \gamma_k} \rfloor)$, and suppose that condition (16) holds with $q = \widetilde{p}$. Then the parameterized KRR predictor $\hat{f}_w$ admits the generalization bound

$$\left\| \hat{f}_w - f_{\mathcal{A}_{<p}}^* - \Theta_d(1) \cdot f_{\mathcal{A}_{=p}}^* \right\|_{L_2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2), \tag{86}$$

and for any $\epsilon > 0$, the bound on the empirical weight estimator error

$$\left| \mathbb{E}_n (\partial_r \hat{f}_w)^2 - (1 + o_{d,\mathbb{P}}(1)) \left(\mathbb{E} \left(\partial_r f_{\mathcal{A}_{<p}}^* \right)^2 + \Theta_d(1) \cdot \mathbb{E} \left(\partial_r f_{\mathcal{A}_{=p}}^* \right)^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} \Theta_d(1) \cdot d^{-2\kappa(\boldsymbol{T})} \mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 \right) \right|$$

$$\leq O_{d,\mathbb{P}} \left(d^{\gamma_{\pi_N(r)} - 1} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2), \tag{87}$$

and the derivative norm estimator error:

$$\left| w_r \cdot \mathcal{D}_r(w) - (1 + o_{d,\mathbb{P}}(1)) \left(\sum_{\boldsymbol{T} \in \{\mathcal{A}_{\leq p_\delta}, \mathcal{A}_{p_\delta}^+(r)\}} \Theta_d(1) \cdot d^{-\kappa(\boldsymbol{T})} \mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 \right) \right|$$

$$\leq O_{d,\mathbb{P}} \left(d^{\gamma_{\pi_N(r)} - 1 + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2), \tag{88}$$

simultaneously over all $r = 1, \ldots, d$.

B.4 Proof of Theorem 4.9

We will prove the results by iteratively applying Theorem B.27. Throughout the proof, we will use the following notation for the estimated coordinate weights at iteration t:

$$w^{[1]}(t) = \underbrace{\mathbb{E}_n(\partial_w \hat{f}_w(t))^2}_{:=\hat{w}} + \mathbf{1}_d \epsilon_{\mathrm{s}},$$

$$w^{[2]}(t) = \mathcal{D}(w(t)) \odot w(t) + \mathbf{1}_d \epsilon_{\mathrm{s}},$$

and

$$\bar{w}^{[1]}(t) = d \cdot \frac{w^{[1]}(t)}{\left\| w^{[1]}(t) \right\|_1}, \quad \bar{w}^{[2]}(t) = d \cdot \frac{w^{[1]}(t)}{\left\| w^{[2]}(t) \right\|_1}. \tag{89}$$

Additionally, we denote

$$\bar{w}(t) = (1 - \alpha) \cdot \bar{w}^{[1]}(t) + \alpha \cdot \bar{w}^{[2]}(t).$$

For convenience, we denote by $w(t)$ the coordinate weights before the update at step t. Accordingly, after applying the update rule, we have the equality

$$w(t + 1) = \bar{w}(t).$$

Since the target function f^* is assumed to be of degree at most ℓ (Assumption 3), we have $\|f^*\|_{L_2}^2 = O_d(1)$ therefore we omit the norm terms in the bound for simplicity of the notation.

In the proof, we will show that within $O_d(1)$ many iterations of the algorithm, the set $\mathcal{A}_{<p_\delta}$ will contain all $\boldsymbol{T}$ where features $\phi_{\boldsymbol{T}}$ have non-zero Fourier coefficients, i.e., $b_{\boldsymbol{T}} \neq 0$ for $\mathsf{L}_{p+1}f^*$. The desired generalization bound can be immediately obtained following Eq. (86) in Theorem B.27.

B.4.1 Supporting Claims and Propositions

We start with a useful claim that shows that the empirical coordinate weights for non-influential coordinates remain of the order $\Theta_{d,\mathbb{P}}(1)$ after normalization during the whole iterations.

Claim 4. *For any coordinate r that is not influential for f^*, the following holds for all $t \geq 0$*

$$\bar{w}_r^{[1]}(t), \quad \bar{w}_r^{[2]}(t) = \Theta_{d,\mathbb{P}}(1). \tag{90}$$

Proof. We prove the result by induction. Observe that for any t, we have

$$\left\| \mathbb{E}_n(\partial_w \hat{f}_w(t))^2 \right\|_1, \quad \|\mathcal{D}(w(t)) \odot w(t)\|_1 \lesssim d^s,$$

where s is the sparsity.

With the assumption $s \leq 1 - \eta$, we immediately have

$$\left\| w^{[1]}(t) \right\|_1, \quad \left\| w^{[2]}(t) \right\|_1 = \Theta_{d,\mathbb{P}}(d^{1-\eta}).$$

At step $t = 0$, since $\gamma_{\pi_N(r)} = 0$, we have

$$w^{[1]}(t) = \Theta_{d,\mathbb{P}}(d^{-1}) + d^{-\eta} = \Theta_{d,\mathbb{P}}(d^{-\eta}),$$
$$w^{[2]}(t) = \Theta_{d,\mathbb{P}}(d^{-1+\epsilon}) + d^{-\eta} = \Theta_{d,\mathbb{P}}(d^{-\eta}).$$

Then applying the normalization (89) gives the claimed bound.

Suppose that at step $t = \tau$ the claimed bound holds. At step $t = \tau + 1$, the inductive hypothesis guarantees that $\gamma_{\pi_N(r)} = 0$. Therefore, by applying the same analysis as in the previous step, we obtain the claimed bound at $t = \tau + 1$. This completes the inductive step and thus the proof. $\square$

We define an auxiliary estimate

$$w_r^{\mathrm{aux}} := (1 - \alpha) \cdot \hat{w}_r + \alpha \cdot \mathcal{D}_r(w) \odot w. \tag{91}$$

Combining the bounds (87) and (88), the estimate w^{aux} satisfies:

$$\left| w_r^{\mathrm{aux}} - (1 + o_{d,\mathbb{P}}(1)) \left(\mathbb{E}\left(\partial_r f_{\mathcal{A}_{<p}}^* \right)^2 + \Theta_d(1) \cdot \mathbb{E}\left(\partial_r f_{\mathcal{A}_{=p}}^* \right)^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} \Theta_d(1) \cdot d^{-\kappa(\boldsymbol{T})} \mathbb{E}(\partial_r f_{\boldsymbol{T}}^*)^2 \right) \right|$$
$$\leq O_{d,\mathbb{P}}\left(d^{\gamma_{\pi_N(r)} - 1 + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2). \tag{92}$$

Since the non-influential coordinates dominate the ℓ_1-norm, normalizing the weights jointly, i.e., normalizing w^{aux}, induces the same scaling across all coordinates as normalizing the weights separately, as performed in Algorithm 1. We formalize the result below and omit the proof since it follows directly from the arguments in the proof of Claim 4.

Claim 5. *For any $t \geq 0$ and $r \in [d]$, the following equality holds:*

$$\bar{w}_r^{\mathrm{aux}}(t) \asymp \bar{w}_r(t). \tag{93}$$

Therefore, for simplicity of presentation, we establish the learning result by analyzing the following update rule, which replaces lines 6–9 in Algorithm 1:

$$w = \frac{w^{\mathrm{aux}} + \epsilon_{\mathrm{s}}}{\|w^{\mathrm{aux}} + \mathbf{1}_d \epsilon_{\mathrm{s}}\|_1}. \tag{94}$$

We similarly denote the normalized w_r^{aux} for each $r \in [d]$ by $\bar{w}_r^{\mathrm{aux}}$.

Now we define fully learned, weakly learned and unlearned coordinates based on the magnitude of w_r^{aux} in Table 2:

Learning status of coordinate r	Magnitude of w^{aux}	Magnitude of $w^{\mathrm{aux}} + \epsilon_{\mathrm{s}}$ ($\epsilon_{\mathrm{s}} = d^{-\eta}$)	Magnitude of $\bar{w}^{\mathrm{aux}}$
Fully learned	$\Theta_d(1)$	$\Theta_d(1)$	$\Theta_d(d^\eta)$
Weakly learned	$\Theta_d(d^{-\eta'}), \eta' \in (0, \eta)$	$\Theta_d(d^{-\eta'})$	$\Theta_d(d^{\eta-\eta'})$
Unlearned	$O_d(d^{-\eta})$	$\Theta_d(d^{-\eta})$	$\Theta_d(d^0)$

Table 2: Magnitude of coordinate weights during a single iteration.

Next we define a feature being learned by $\hat{f}$:

Definition 6 (Learned feature). *A feature $\phi_{\boldsymbol{T}}$ is said to be learned by $\hat{f}$, if the following holds:*

$$\left| \left\langle \phi_{\boldsymbol{T}}, \hat{f} \right\rangle_{L_2} - b_{\boldsymbol{T}} \right| = o_{d,\mathbb{P}}(1).$$

The following proposition shows that the order of $\bar{w}_r^{\mathrm{aux}}$ is non-decreasing, which implies that the learned features remain learned and new features will be learned as the interaction step t increases.

Claim 6. *For each empirical coordinate weight $\bar{w}_r$, its order of magnitude with respect to d is non-decreasing as a function of the iteration step t throughout the first $O_d(1)$ iterations of Algorithm 1.*

Proof. Based on the update rule (87) and (88), we can deduce that for any unlearned coordinate r, the empirical weight $\bar{w}_r$ can increase the order only when they become fully or weakly learned, otherwise $\bar{w}_r(t)$ remains at the order $\Theta_d(1)$ with $\gamma_{\pi_N(r)} = 1$. For fully learned and weakly learned coordinates, the results are proved in Claim 7 and 8 respectively. $\square$

Next we give a claim for the fully learned coordinates:

Claim 7. *For any fully learned coordinate r, it will remain fully learned until the algorithm terminates.*

Proof. Note that when coordinate r is fully learned, there must exist a feature $\phi_{\boldsymbol{T}}$ that depends on x_r such that $p_{\text{eff}}(\boldsymbol{T}) \leq p + \delta$ and $\|\partial_r f_{\boldsymbol{T}}^*\|_{L_2} > 0$. Furthermore, any coordinate upon which $\phi_{\boldsymbol{T}}$ depends is also fully learned. For this specific feature, a recursive argument shows that the corresponding weight $\bar{w}$ of all coordinates on which $\phi_{\boldsymbol{T}}$ depends remains the order unchanged, specifically $\Theta_{d,\mathbb{P}}(d^\eta)$, and the condition $p_{\text{eff}}(\boldsymbol{T}) \leq p + \delta$ remains satisfied until algorithm termination. This establishes that coordinate r remains fully learned throughout the process. $\qquad\square$

To prove the desired result, i.e., $\left\|\hat{f} - \mathsf{L}_{p+1}f^*\right\|_{L_2} = o_{d,\mathbb{P}}(1)$, we will show for any feature $\phi_{\boldsymbol{T}}$ that is learned by $\hat{f}$, feature $\phi_{\boldsymbol{T}'}$ with $|\boldsymbol{T}' \setminus \boldsymbol{T}| \leq p + 1$ is also learned when the algorithm terminates. Here we abuse the notation by treating the tuple $\boldsymbol{T} = (T_1, \ldots, T_N)$ as a set $\cup_{j=1}^N T_k$. To prove the result, it suffices to show that $\phi_{\boldsymbol{T}'} := \phi_{\boldsymbol{T}}\phi_S$ for any $S \in 2^{[d]}$ with $|S| \leq p + 1$ is learned by $\hat{f}$.

B.4.2 Main proof

The following claim shows that the order of the weight of weakly learned coordinates will monotonically increase until it reaches the maximum order η:

Claim 8. *For any coordinate r that is weakly learned, we have the estimate*

$$\bar{w}_r^{\text{aux}} = \Theta_{d,\mathbb{P}}(d^{\eta-\kappa(\boldsymbol{T}_r')}) \cdot \mathbb{E}\left(\partial_r f_{\boldsymbol{T}_r'}^*\right)^2, \tag{95}$$

where

$$\boldsymbol{T}_r' = \underset{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r) \,:\, \mathbb{E}\left(\partial_r f_{\boldsymbol{T}}^*\right)^2 > 0}{\arg\max} \kappa(\boldsymbol{T}). \tag{96}$$

Furthermore, there exists a constant $\sigma > 0$ independent of the step t such that the following holds:

$$\lambda_{\pi_N(r)}(t+1) \leq \max(1 - \eta, \lambda_{\pi_N(r)}(t) - \sigma). \tag{97}$$

Proof. We prove the result by induction. Note that there exists a time step where the coordinate r changes from unlearned to weakly learned after this step. Without loss of generality, we assume this step is $t = 1$. Correspondingly, we have $\lambda_{\pi_N(r)}(1) = 1$ and $\bar{w}_r^{\text{aux}}(1) = \Theta_d(d^\sigma)$ with some constant $\sigma \in (0, \eta)$, where

$$\sigma := \eta - \kappa(\boldsymbol{T}_r'(1)).$$

Since $\kappa(\boldsymbol{T}_r'(1)) < \lambda_{\pi_N(r)}(1) = 1$, the error term in Eq. (92) is strictly smaller than w_r^{aux} hence we have the estimate (95). Consequently, we have $w_r(2) = \bar{w}_r(1) \gtrsim \Theta_d(d^\sigma)$ which implies

$$\lambda_{\pi_N(r)}(2) \leq 1 - \sigma.$$

Then we complete the bound (97) at $t = 1$.

Now we suppose that at step $t = \tau$ the estimate (95) and the inequality (97) hold and we consider $t = \tau + 1$. A simple calculation shows

$$\lambda_{\pi_N(r)}(\tau + 1) = 1 - \eta + \kappa(\boldsymbol{T}_r'(\tau)).$$

We assume $\kappa(\boldsymbol{T}_r'(\tau)) > 0$ otherwise coordinate r has been fully learned at step $\tau + 1$ which finishes the proof. Notice that the inductive hypothesis indicates that $\lambda_{\pi_N(r)}(\tau)$ will decrease by at least σ:

$$\lambda_{\pi_N(r)}(\tau + 1) \leq \lambda_{\pi_N(r)}(\tau) - \sigma.$$

49

Therefore, considering unlearned and fully learned coordinates, we can deduce that until step $\tau + 1$, the empirical coordinates weight w is non-decreasing in terms of the order of the magnitude. Consequently, the effective dimension will also decrease by at least σ:

$$p_{\text{eff}}(\boldsymbol{T}'_r(\tau + 1)) \leq p_{\text{eff}}(\boldsymbol{T}'_r(\tau)) - \sigma$$

which leads to the decrease of $\kappa(\boldsymbol{T}'_r(\tau + 1))$:

$$\kappa(\boldsymbol{T}'_r(\tau + 1)) \leq \kappa(\boldsymbol{T}'_r(\tau)) - \sigma. \tag{98}$$

If $\kappa(\boldsymbol{T}'_r(\tau + 1)) \leq 0$, then coordinate r is fully learned at step $\tau + 2$, therefore we only need to consider the case when $\kappa(\boldsymbol{T}'_r(\tau + 1)) > 0$.

From Eq. (98) and the condition $\sigma \in (0, \eta)$, we immediately have

$$\lambda_{\pi_N(r)}(\tau + 1) = 1 - \eta + \kappa(\boldsymbol{T}'_r(\tau)) \geq 1 - \eta + \kappa(\boldsymbol{T}'_r(\tau + 1)) + \sigma > \kappa(\boldsymbol{T}'_r(\tau + 1)),$$

which completes the inductive step for Eq. (95) at step $\tau + 1$. Additionally, with $\lambda_{\pi_N(r)}(\tau + 2) = 1 - \eta + \kappa(\boldsymbol{T}'_r(\tau + 1))$, we can deduce

$$\lambda_k(\tau + 2) \leq \max(1 - \eta, \lambda_k(\tau + 1) - \sigma)$$

which completes the inductive step for Eq. (97) thus completing the proof. □

With the results of Claim 8, we can deduce that for any coordinate that is weakly learned, it will be fully learned in $O_d(1)$ many steps. Next, we show that at most $p + 1$ unlearned coordinates can be learned (fully or weakly) in a single step, which then indicates that the algorithm can learn functions with leap complexity $p + 1$. To that end, we consider the possible tuple $\boldsymbol{T}$ that satisfies:

$$\underset{\boldsymbol{T} = (T_1, \ldots, T_N), \ \lambda_1 = 1}{\arg\max} |T_1|, \quad s.t. \quad p_{\text{eff}}(\boldsymbol{T}) < p + \delta + \eta.$$

Note that it is sufficient to consider $\boldsymbol{T}$ with only two entries, i.e., $\boldsymbol{T} = (T_1, T_2)$ with $\lambda_1 = 1$ and $\lambda_2 = 1 - \eta$, i.e., fully learned and unlearned coordinates. For those weakly learned coordinates, Claim 8 ensures that they will ultimately be fully learned, so the format of $\boldsymbol{T}$ reduces to one with two entries. Given the assumption $\eta > 1 - \frac{\delta}{\ell + 1}$, we can deduce the effective degree satisfies

$$\lambda_2 \cdot \ell + \lambda_1 \cdot (p + 1) = (1 - \eta) \cdot \ell + 1 \cdot (p + 1) < p + \delta + \eta, \tag{99}$$

which implies the maximum value of $|T_1|$ is $p + 1$. We conclude the proof.

B.5 Proof of Theorem B.27 (Gradient error)

B.5.1 Preparing Lemmas, Propositions and Notation

We start with some useful notations. We group the tuple $\boldsymbol{T}$ based on the cardinality of each T_k. We define the collection of tuples $\mathcal{S}_{\boldsymbol{m}}$ where $\boldsymbol{m} = (m_1, \ldots, m_N) \in \mathbb{N}^N$:

$$\mathcal{S}_{\boldsymbol{m}} = \left\{ \boldsymbol{T} = (T_1, \ldots, T_N) \mid T_k \in 2^{\mathcal{T}_k}, \ |T_k| = m_k \ \text{ for } \ k \in [N] \right\}. \tag{100}$$

Then $\phi_{\mathcal{S}_{\boldsymbol{m}}}$ is defined as

$$\phi_{\mathcal{S}_{\boldsymbol{m}}} := (\phi_{\boldsymbol{T}})_{\boldsymbol{T} \in \mathcal{S}_{\boldsymbol{m}}} \in \mathbb{R}^{\boldsymbol{s}^\top \boldsymbol{m}}. \tag{101}$$

Clearly, given $\boldsymbol{m}$, any $\phi_{\boldsymbol{T}}$ for $\boldsymbol{T} \in \mathcal{S}_{\boldsymbol{m}}$ has identical effective degree $\boldsymbol{\lambda}^\top \boldsymbol{m}$. Since there is a one-to-one correspondence between the tuple $\boldsymbol{T}$ and the set $\cup_{k=1}^N T_k$, we sometimes abuse the notation by treating $\boldsymbol{T}$ as $\cup_{k=1}^N T_k$ when it is clear from the context. We sometimes use $\boldsymbol{T}$ in subscripts, such as in $D_{\boldsymbol{T}, \boldsymbol{T}}$, to denote the submatrix of D corresponding to the rows and columns indexed by the elements in $\boldsymbol{T}$.

For simplicity of the notation, given the order of the weights in each T_k, i.e., $(\gamma_1, \ldots, \gamma_N)$, we denote a vector

$$\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_N), \quad \lambda_k = 1 - \gamma_k \text{ for } k \in [N].\tag{102}$$

Following Definition 5, we define the *adjusted effective degree* for $\boldsymbol{T}$, which is

$$p_{\text{aeff}}(\boldsymbol{T}) := \sum_{k=1}^{N} s_k \cdot |T_k|.\tag{103}$$

Clearly, we have $p_{\text{aeff}}(\boldsymbol{T}) \leq p_{\text{eff}}(\boldsymbol{T})$ since $s_k \leq \lambda_k$.

We define two constants that measure the difference of the (adjusted) effective degree of $\phi_{\boldsymbol{T}} \in \phi_{\mathcal{S}_{\boldsymbol{m}}}$ from $p + \delta$:

$$\zeta_1 := \min_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \left(p + \delta - \boldsymbol{s}^\top \boldsymbol{m}\right)\tag{104}$$

$$\zeta_2 := \min_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta} \left(\boldsymbol{\lambda}^\top \boldsymbol{m} - p - \delta\right).\tag{105}$$

Claim 9. *The constants ζ_1, ζ_2 are bounded from zero.*

Proof. By definition, $\zeta_2 > 0$ trivially holds. Now we show $\zeta_1 > 0$ also holds. By assumption, $s_k = \lambda_k$ holds iff $s_k = 1$. Therefore, we can see that if one of $m_k > 0$ and the corresponding $s_k < 1$, then $\boldsymbol{s}^\top \boldsymbol{m} < \boldsymbol{\lambda}^\top \boldsymbol{m} \leq p + \delta$ which implies $\zeta_1 > 0$. Otherwise, if $s_k = 0$ or 1 for all $m_k > 0$, then $\zeta_1 = p + \delta - \boldsymbol{s}^\top \boldsymbol{m} \geq \delta > 0$. $\square$

For simplicity, we denote

$$\zeta := \min(\zeta_1, \zeta_2) > 0.\tag{106}$$

Lastly, we denote the maximum order of w_r as

$$\gamma_{\max} := \max_{k \in [N]} \gamma_k < 1.\tag{107}$$

Taylor approximation of kernels. We similarly denote

$$g_{w,m}(t) := \sum_{k=0}^{m} \frac{g_w^{(k)}(0)}{k!} t^k \text{ and } K_{w,m}(x^{(i)}, x^{(j)}) := g_{w,m}\left(\frac{\langle x^{(i)}, x^{(j)}\rangle}{d}\right).$$

For simplicity of notation, we denote $z := \sqrt{w} \odot x$. Invoking Lemma A.16, K_w can be decomposed into

$$K_w(X, X) = K(Z, Z) = \Phi(Z)D\Phi(Z)^\top\tag{108}$$

where D is a diagonal matrix satisfying $\|D_{\mathcal{S}_k} - g^{(k)}(0)d^{-k}I_{|\mathcal{S}_k|}\|_{\text{op}} = O_d(d^{-k-1})$ for $k = 0, 1, \ldots, \infty$.

We present a similar result to Theorem A.14 for parameterized kernels. The proof appears in Appendix C.4.2.

Theorem B.28 (Taylor approximation of parameterized kernels). *Consider independent, mean zero, isotropic random vectors $x^{(1)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$ and suppose that the coordinates of each vector $x^{(i)}$ are independent and are uniformly drawn from $\mathbb{H}^d$. Suppose that we are in the regime $n = d^{p+\delta}$, where $\delta \in (0, 1)$ is a constant. Fix an arbitrary constant $C > 2$ and approximation order $\widetilde{p} \geq \frac{2(p+\delta)}{1-\gamma}$. Given $w \in \mathbb{R}_+^d$ that satisfies $\|w\|_1 = d$, and $\|w\|_\infty = O_d(d^\gamma)$ with $\gamma \in [0, 1)$, if $d^{1-\gamma}/\log(d) > c_\epsilon$ then the estimate*

$$\|K_w(X, X) - K_{w,\widetilde{p}}(X, X) - (g(1) - g_{\widetilde{p}}(1))I_n\|_{\text{op}} \leq c_{g,\widetilde{p}}\sqrt{\frac{(C\log(n)\|w\|_\infty)^{\widetilde{p}+1}}{d^{\widetilde{p}-2p+1-2\delta}}}\tag{109}$$

holds with probability at least $1 - \frac{4}{n^{C-2}}$, where $c_{g,\widetilde{p}}, c_\epsilon < \infty$ are constants that depend only on $\widetilde{p}$, $\sup_{k \geq \widetilde{p}+1} g^{(k)}(0)$, and ϵ.

Plugging

$$\left\lceil \frac{2p + 2\delta + 2}{1 - \gamma_{\max}} - 1 \right\rceil := \widetilde{p}_{\gamma_{\max}}$$

as $\widetilde{p}$ into Eq. (109), we immediately have

$$\left\| K_w(X, X) - K_{w, \widetilde{p}_{\gamma_{\max}}}(X, X) - (g(1) - g_{\widetilde{p}_{\gamma_{\max}}}(1))I_n \right\|_{\mathrm{op}} = O_{d, \mathbb{P}}(d^{-\frac{1}{2}}). \tag{110}$$

Now we develop the isometric properties of transformed features $\phi_{\mathcal{S}_m}(Z)$ corresponding to Theorem A.13, which stands with a scaling below. See the proof in Appendix C.4.1.

Lemma B.29 (Isometric properties of scaled orthogonal monomials)**.** *Consider the regime* $n = d^{p+\delta}$. *For any* $\boldsymbol{m}$ *that satisfies* $\|\boldsymbol{m}\|_1 \leq \widetilde{p}_\gamma$, *the following holds for any* $\epsilon > 0$:

$$\left\| \frac{\Phi_{\mathcal{S}_m}(Z)\left(\Phi_{\mathcal{S}_m}(Z)\right)^\top}{d^{(\boldsymbol{s}+\boldsymbol{1}-\boldsymbol{\lambda})^\top \boldsymbol{m}}} - \mu_{\mathcal{S}_m} I \right\|_{\mathrm{op}} = O_{d, \mathbb{P}}(d^{p+\delta-\boldsymbol{s}^\top \boldsymbol{m}+\epsilon} + d^{(p+\delta-\boldsymbol{s}^\top \boldsymbol{m})/2+\epsilon}), \tag{111}$$

where

$$\mu_{\mathcal{S}_m} := \frac{1}{d^{(\boldsymbol{s}+\boldsymbol{1}-\boldsymbol{\lambda})^\top \boldsymbol{m}}} \mathbb{E}_z \left[\|\phi_{\mathcal{S}_m}(z)\|_2^2 \right] = \Theta_d(1) \tag{112}$$

Corresponding to Eq. (82), we now write $\mathcal{A}$ in terms of the union of $\mathcal{S}_m$:

$$\mathcal{A}_{\leq p_\delta} := \bigcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \mathcal{S}_m,$$

$$\mathcal{A}_{p_\delta}^+(r) := \bigcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_{\pi_N(r)}]} \mathcal{S}_m,$$

and we additionally denote

$$\mathcal{A}_{> p_\delta} := \bigcup_{\substack{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta \\ \|\boldsymbol{m}\|_1 \leq \widetilde{p}_{\gamma_{\max}}}} \mathcal{S}_m.$$

Accordingly, we can further decompose $K(Z, Z)$ based on Eq. (108) as

$$K(Z, Z) = \Phi_{\mathcal{A}_{\leq p_\delta}}(Z)D_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}(Z)^\top + \Phi_{\mathcal{A}_{> p_\delta}}(Z)D_{\mathcal{A}_{> p_\delta}}\Phi_{\mathcal{A}_{> p_\delta}}(Z)^\top + g(1) - g_{\widetilde{p}_{\gamma_{\max}}}(1). \tag{113}$$

In parallel to Lemma A.17, we have the following estimate for $K(Z, Z)$. We defer the proof to Appendix C.4.3.

Lemma B.30. *Consider the regime* $n = d^{p+\delta}$. *Consider a weight vector* $w \in \mathbb{R}_+^d$ *satisfying* $\|w\|_1 = d$ *and let the coordinate set be an* $(\boldsymbol{s}, \boldsymbol{\lambda})$-*active set for* w. *The following holds for any* $\epsilon > 0$:

$$\left\| K(Z, Z) - \Phi_{\mathcal{A}_{\leq p_\delta}}(Z)D_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}(Z)^\top - \rho I_n \right\|_{\mathrm{op}} = O_{d, \mathbb{P}}(d^{-\zeta_2/2+\epsilon}), \tag{114}$$

$$\left\| K(Z, Z) - \Phi_{\mathcal{A}_{\leq p_\delta}}(Z)D_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}(Z)^\top - \sum_{\boldsymbol{m} \in J_1} \mathrm{off}\left(\Phi_{\mathcal{S}_m}(Z)D_{\mathcal{S}_m}\Phi_{\mathcal{S}_m}(Z)^\top\right) - \rho I_n \right\|_{\mathrm{op}} = O_{d, \mathbb{P}}(d^{-1+\epsilon}), \tag{115}$$

where

$$\rho = \sum_{\substack{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta, \\ \|\boldsymbol{m}\|_1 \leq \widetilde{p}_{\gamma_{\max}}}} d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top \boldsymbol{m}} g^{(\|\boldsymbol{m}\|_1)}(0)\mu_{\mathcal{S}_m} + g(1) - g_{\widetilde{p}_{\gamma_{\max}}}(1) = \Theta_d(1), \tag{116}$$

$$J_1 = \{\boldsymbol{m} \mid \boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta \ \text{and} \ \boldsymbol{s}^\top \boldsymbol{m} < p+2\} \tag{117}$$

52

Through the proof, we consistently treat functions with argument X or x and we will specify when arguments are z. Since $z = \sqrt{w} \odot x$, we can find a diagonal matrix W to transform from $\phi(X)$ to $\phi(Z)$: for any multi-index S, we define

$$W_{S,S} = w^S. \tag{118}$$

Then we can write $\phi(Z)$ as

$$\phi(Z) = \phi(X)W^{1/2}.$$

Lastly, for a diagonal matrix D, we use $D_{\mathcal{S}}$ where $\mathcal{S}$ is a set to denote a sub-matrix of D with entries $D_{S,S}$ with $S \in \mathcal{S}$.

B.5.2 Main proof

Given the parameterized kernel Eq. (77), for a new input x, the KRR prediction function is given by

$$\hat{f}_w(x) = K_w(x, X)(K_w(X, X) + \lambda I_n)^{-1} y \qquad \text{where} \qquad \beta := (K_w(X, X) + \lambda I_n)^{-1} y, \tag{119}$$

where $\lambda > 0$ is the ridge parameter.

We will estimate the following:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\partial_r \hat{f}_w(x^{(i)}) \right)^2 = \frac{w_r^2}{nd^2} \left\| K'_w X_r K_{w,\lambda}^{-1} y \right\|_2^2. \tag{120}$$

For simplicity of notation, we denote

$$\widehat{D} := WD.$$

In parallel to Eq. (65) and (66), applying Lemma B.30 with the two Kernels K_w and K'_w shows that for any $\epsilon > 0$ there exist matrices Δ, Δ_1 satisfying $\|\Delta\|_{\mathrm{op}}, \|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$ and

$$K_w = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + \rho I_n + \Delta_1, \tag{121}$$

$$K'_w = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + \rho' I_n + \Delta, \tag{122}$$

where ρ is defined in Eq. (116) and ρ' is defined as

$$\rho' := \sum_{\substack{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} > p + \delta, \\ \|\boldsymbol{m}\|_1 \leq \widetilde{p}_{\gamma_{\max}}}} d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top \boldsymbol{m}} g^{(\|\boldsymbol{m}\|_1)}(0) \mu_{\mathcal{S}_{\boldsymbol{m}}} + g'(1) - g_{\widetilde{p}_{\gamma_{\max}}+1}(1).$$

With the expression (122) in place of K'_w, equation (120) becomes

$$\frac{1}{n} \sum_{i=1}^{n} \left(\partial_r \hat{f}_w(x^{(i)}) \right)^2 = \frac{w_r^2}{nd^2} \left\| \left(\underbrace{\Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top X_r K_{w,\lambda}^{-1}}_{=:T_1} + \underbrace{(\rho' I_n + \Delta) X_r K_{w,\lambda}^{-1}}_{=:T_2} \right) y \right\|_2^2.$$

Expanding the square we have

$$\frac{1}{n} \sum_{i=1}^{n} \left(\partial_r \hat{f}_w(x^{(i)}) \right)^2 = \frac{w_r^2}{nd^2} \left(\|T_1 y\|_2^2 + \|T_2 y\|_2^2 + 2 \langle T_1 y, T_2 y \rangle \right). \tag{123}$$

We will analyze each addend separately. We suppose $\iota = \pi_N(r)$, i.e., $r \in \mathcal{T}_\iota$ for some $k \in [N]$. Specifically, we will show

Claim 10. *The following estimates hold for any $\epsilon > 0$ uniformly over all coordinates $r \in \mathcal{T}_\iota$ where $c_r, c_{\boldsymbol{T}} > 0$ are constants:*

(a) $\frac{w_r^2}{nd^2} \|T_2 y\|_2^2 = O_{d,\mathbb{P}} \left(d^{-2\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2);$

(b) $\left| \frac{w_r^2}{nd^2} \|T_1 y\|_2^2 - \left\| \partial_r f^*_{\mathcal{A}_{<p_\delta}} \right\|_{L_2}^2 - c_r \cdot \left\| \partial_r f^*_{\mathcal{A}_{=p_\delta}} \right\|_{L_2}^2 - \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} c_{\boldsymbol{T}} \cdot d^{-2\kappa(\boldsymbol{T})} \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2}^2 \right|$

$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \left(\left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2}^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} d^{-2\kappa(\boldsymbol{T})} \cdot \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2}^2 \right)$

$+ O_{d,\mathbb{P}} \left(d^{(-\lambda_\iota + \epsilon)/2} \right) \cdot \|f^*\|_{L_2} \left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2} + O_{d,\mathbb{P}} \left(d^{-\lambda_\iota/2} \right) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} d^{-\kappa(\boldsymbol{T})} \cdot \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2} \|f^*\|_{L_2};$

(c) $\frac{w_r^2}{nd^2} \langle T_1 y, T_2 y \rangle = O_d \left(d^{-\frac{3}{2}\lambda_\iota + \epsilon} \right) \cdot \|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota + \epsilon}) \|f^*\|_{L_2} \left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2}$

$+ \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} \cdot d^{-2\kappa(\boldsymbol{T}) - \lambda_\iota + \epsilon} \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2} \|f^*\|_{L_2}.$

Plugging Claim 10 (a,b,c) into Eq. (123) yields:

$$\left| \mathbb{E}_n (\partial_r \hat{f}_w)^2 - \mathbb{E} \left(\partial_r f^*_{\mathcal{A}_{<p}} \right)^2 - c_r \cdot \mathbb{E} \left(\partial_r f^*_{\mathcal{A}_{=p}} \right)^2 - \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} c_{\boldsymbol{T}} \cdot d^{-2\kappa(\boldsymbol{T})} \mathbb{E}(\partial_r f^*_{\boldsymbol{T}})^2 \right|$$

$$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \left(\left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2}^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} d^{-2\kappa(\boldsymbol{T})} \cdot \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2}^2 \right)$$

$$+ O_{d,\mathbb{P}} \left(d^{(-\lambda_\iota + \epsilon)/2} \right) \cdot \|f^*\|_{L_2} \left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2} + O_{d,\mathbb{P}} \left(d^{-\lambda_\iota/2} \right) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{p_\delta}^+(r)} d^{-\kappa(\boldsymbol{T})} \cdot \|\partial_r f^*_{\boldsymbol{T}}\|_{L_2} \|f^*\|_{L_2}.$$

$$(124)$$

Similarly, the rest of the section is devoted to proving Claim 10. Note that Item (c) follows immediately from Item (a) and (b) and the Cauchy-Schwartz inequality. Throughout the proof, all asymptotic terms $o_{d,\mathbb{P}}$ and $O_{d,\mathbb{P}}$ that appear will always be uniform over the coordinates $r = 1, \ldots, d$.

B.5.3 Proof of Item (a) of Claim 10

The proof is identical to the proof of Claim 1 (a). We refer the readers to Section B.1.1 for more details.
We can yield the following bound:

$$\frac{w_r^2}{nd^2} \|T_2 y\|_2^2 = O_{d,\mathbb{P}} \left(w_r^2 \cdot \frac{\log(d)}{d^2} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) = O_{d,\mathbb{P}} \left(d^{-2\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2),$$

where $\epsilon > 0$ is an arbitrarily small constant.

B.5.4 Proof of Item (b) of Claim 10

We denote $\beta = K_{w,\lambda}^{-1} y$. Similar to Eq. (60), we can write

$$\frac{w_r^2}{nd^2} \|T_1 y\|^2 = \frac{w_r^2}{nd^2} \beta^\top [X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top X_r] \beta$$

$$= \frac{w_r^2}{d^2} \beta^\top [X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} (I + \widetilde{\Delta}) \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top X_r] \beta$$

$$= \frac{1 + O_d(\|\widetilde{\Delta}\|_{\mathrm{op}})}{d^2} \cdot \|w_r \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}}\|_2^2, \qquad (125)$$

where $\left\| \widetilde{\Delta} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2 + \epsilon}).$

Notice that $\mathcal{A}_{\leq p_\delta} = \cup_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \mathcal{S}_{\boldsymbol{m}}$. Let e_ι be ι-th standard basis vector. We decompose $X_r \Phi_{\mathcal{A}_{\leq p_\delta}}$ into components:

$$\{\Phi_{\mathcal{S}_{\boldsymbol{m}}^0}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \quad \text{and} \quad \{\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\}_{\boldsymbol{\lambda}^\top (\boldsymbol{m}-e_\iota) \leq p+\delta}, \tag{126}$$

where

$$\mathcal{S}_{\boldsymbol{m}}^1 := \left\{ \boldsymbol{T} \mid r \in \cup_{k=1}^N T_k \text{ and } |T_k| = m_k, \quad k \in [N] \right\}, \tag{127}$$

$$\mathcal{S}_{\boldsymbol{m}}^0 := \left\{ \boldsymbol{T} \mid r \notin \cup_{k=1}^N T_k \text{ and } |T_k| = m_k, \quad k \in [N] \right\}. \tag{128}$$

We correspondingly decompose $\widehat{D}'_{\mathcal{A}_{\leq p_\delta}}$. We can show there exist diagonal matrices $\{R_{\boldsymbol{m}}^0\}_{\boldsymbol{m}}$ and $\{R_{\boldsymbol{m}}^1\}_{\boldsymbol{m}}$ such that the following holds:

$$\left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \left(\frac{w_r}{d} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \right) \right\|_2^2$$
$$= \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \| \beta^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^0} R_{\boldsymbol{m}}^0 \|_2^2 + \sum_{\boldsymbol{\lambda}^\top (\boldsymbol{m}-e_\iota) \leq p+\delta} \| \beta^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^1} R_{\boldsymbol{m}}^1 \|_2^2, \tag{129}$$

Claim 11. *The following estimate holds:*

$$\left\| R_{\boldsymbol{m}}^0 - g^{(\|\boldsymbol{m}\|_1 + 2)}(0) \cdot w_r^2 \cdot W_{\mathcal{S}_{\boldsymbol{m}}^0} d^{-\|\boldsymbol{m}\|_1 - 2} \right\|_{\text{op}} = w_r^2 \cdot O_d(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m} - 3}), \tag{130}$$

$$\left\| R_{\boldsymbol{m}}^1 - g^{(\|\boldsymbol{m}\|_1)}(0) \cdot W_{\mathcal{S}_{\boldsymbol{m}}^1} d^{-\|\boldsymbol{m}\|_1} \right\|_{\text{op}} = O_d(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m} - 1}), \tag{131}$$

Proof. Recall that D' is a diagonal matrix with entry $D'_{\mathcal{S}_{\boldsymbol{m}}} = g^{(\|\boldsymbol{m}\|_1 + 1)}(0) d^{-\|\boldsymbol{m}\|_1} + O_d(d^{-\|\boldsymbol{m}\|_1 - 1})$. Also each entry of $W_{\mathcal{S}_{\boldsymbol{m}}}$ is of the order $d^{\|\boldsymbol{m}\|_1 - \boldsymbol{\lambda}^\top \boldsymbol{m}}$ based on Eq. (118).

Based on Eq. (129), we can see that $\Phi_{\mathcal{S}_{\boldsymbol{m}}^0} R_{\boldsymbol{m}}^0$ corresponds to $x_r \cdot \Phi_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1} \frac{w_r}{d} D'_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1}$ since $\phi_{\mathcal{S}_{\boldsymbol{m}}^0} = x_r \phi_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1}$. Then we can deduce that

$$R_{\boldsymbol{m}}^0 = \frac{w_r}{d} W_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1} D'_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1} = \frac{w_r^2}{d} W_{\mathcal{S}_{\boldsymbol{m}}^0} D'_{\mathcal{S}_{\boldsymbol{m}+e_\iota}^1}.$$

Plugging the expression of W and D' into the above equation completes the proof for Eq. (130). A similar analysis applies to

$$R_{\boldsymbol{m}}^1 = \frac{w_r}{d} W_{\mathcal{S}_{\boldsymbol{m}-e_\iota}^0} D'_{\mathcal{S}_{\boldsymbol{m}-e_\iota}^0} = \frac{1}{d} W_{\mathcal{S}_{\boldsymbol{m}}^1} D'_{\mathcal{S}_{\boldsymbol{m}-e_\iota}^0}.$$

$\square$

Regarding Eq. (129), we group the features based on the inequalities (1) $\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p + \delta$ and (2) $\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta$. Then we can define a diagonal matrix R with each block $R_{\mathcal{S}_{\boldsymbol{m}}^1, \mathcal{S}_{\boldsymbol{m}}^1} = R_{\boldsymbol{m}}^1$ and $R_{\mathcal{S}_{\boldsymbol{m}}^0, \mathcal{S}_{\boldsymbol{m}}^0} = R_{\boldsymbol{m}}^0$ such that Eq. (129) can be written as

$$\left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \left(\frac{w_r}{d} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \right) \right\|_2^2 = \left\| \beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 + \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} \| \beta^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^1} R_{\boldsymbol{m}}^1 \|_2^2. \tag{132}$$

We start by analyzing the first term on the right side of (132). The proof of the proposition appears in Appendix C.4.4.

Proposition B.31. *For any $\epsilon > 0$, the following estimate holds uniformly for all coordinates $r \in \mathcal{T}_\iota$ where $c_r > 0$ is a constant:*

$$\left| \left\| \beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 - \left\| \partial_r f_{\mathcal{A}_{<p_\delta}}^* \right\|_{L_2}^2 - c_r \cdot \left\| \partial_r f_{\mathcal{A}_{=p_\delta}}^* \right\|_{L_2}^2 \right|$$

$$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota} \right) \cdot \|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{A}_{\leq p_\delta}}^* \right\|_{L_2}^2 + O_{d,\mathbb{P}} \left(d^{-\lambda_\iota/2} \right) \cdot \|f^*\|_{L_2} \left\| \partial_r f_{\mathcal{A}_{\leq p_\delta}}^* \right\|_{L_2}, \tag{133}$$

55

We proceed to analyze the second term on the right side of (132). The proof appears in Appendix C.4.5.

Proposition B.32. *For any $\epsilon > 0$, the estimate holds uniformly for all coordinates $r \in \mathcal{T}_\iota$ and $\boldsymbol{m} \in \mathbb{R}^N$ that satisfies $\boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]$:*

$$\left| \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 \right\|_2^2 - \left(g^{(\|\boldsymbol{m}\|_1)}(0) \right)^2 \left\| (\rho+\lambda)^{-1} d^{p+\delta - \|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2^2 \right|$$

$$= O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + d^{2(p+\delta - \boldsymbol{\lambda}^\top \boldsymbol{m})} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}(d^{p+\delta - \boldsymbol{\lambda}^\top \boldsymbol{m} - \lambda_\iota/2}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{134}$$

Following the triangle inequality, the second term in Eq. (132) can be approximated as

$$\left| \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 \right\|_2^2 - \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} \left(g^{(\|\boldsymbol{m}\|_1)}(0) \right)^2 \left\| (\rho+\lambda)^{-1} d^{p+\delta - \|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2^2 \right|$$

$$= O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} d^{2(p+\delta - \boldsymbol{\lambda}^\top \boldsymbol{m})} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}^2$$

$$+ \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} O_{d,\mathbb{P}}(d^{p+\delta - \boldsymbol{\lambda}^\top \boldsymbol{m} - \lambda_\iota/2}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{135}$$

We can further simplify the above equation. Note that $|W_{\boldsymbol{T},\boldsymbol{T}}| = \Theta_d(d^{\|\boldsymbol{m}\|_1 - \boldsymbol{\lambda}^\top \boldsymbol{m}})$ for any $\boldsymbol{T} \in \mathcal{S}_m^1$. We can deduce that there exists a constant $c_{\boldsymbol{T}} > 0$ such that

$$\left(g^{(\|\boldsymbol{m}\|_1)}(0) \right)^2 \left\| (\rho+\lambda)^{-1} d^{p+\delta - \|\boldsymbol{m}\|_1} b_{\boldsymbol{T}}^\top W_{\boldsymbol{T},\boldsymbol{T}} \right\|^2 = c_{\boldsymbol{T}} \cdot d^{2(p+\delta - \boldsymbol{\lambda}^\top \boldsymbol{m})} \cdot \|b_{\boldsymbol{T}}\|_2^2 = c_{\boldsymbol{T}} d^{-2\kappa(\boldsymbol{T})} \cdot \|b_{\boldsymbol{T}}\|_2^2 \tag{136}$$

where $\kappa(\boldsymbol{T}) := \boldsymbol{\lambda}^\top \boldsymbol{m} - p + \delta$ is defined in Eq. (85).

Finally with the notation $\zeta := \min(\zeta_1, \zeta_2)$, Eq. (135) can be reduced to

$$\left| \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta, p+\delta+\lambda_\iota]} \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 \right\|_2^2 - \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} c_{\boldsymbol{T}} \cdot d^{-2\kappa(\boldsymbol{T})} \|b_{\boldsymbol{T}}\|_2^2 \right|$$

$$= O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} d^{-2\kappa(\boldsymbol{T})} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \|b_{\boldsymbol{T}}\|_2^2$$

$$+ \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} d^{-\kappa(\boldsymbol{T})} \cdot O_{d,\mathbb{P}}(d^{-\lambda_\iota/2}) \cdot \|b_{\boldsymbol{T}}\|_2 (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{137}$$

Eq. (133) and (137) together give

$$\left| \left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p\delta}} \left(\frac{w_r}{d} \widehat{D}'_{\mathcal{A}_{\leq p\delta}} \right) \right\|_2^2 - \left\| \partial_r f_{\mathcal{A}_{\leq p\delta}}^* \right\|_{L_2}^2 - c_r \cdot \left\| \partial_r f_{\mathcal{A}_{=p\delta}}^* \right\|_{L_2}^2 - \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} c_{\boldsymbol{T}} \cdot d^{-2\kappa(\boldsymbol{T})} \|b_{\boldsymbol{T}}\|_2^2 \right|$$

$$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{A}_{\leq p\delta}}^* \right\|_{L_2}^2 + \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} d^{-2\kappa(\boldsymbol{T})} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}} \left(d^{-\lambda_\iota/2} \right) \cdot (\|f^*\|_{L_2} + \sigma_\varepsilon) \left\| \partial_r f_{\mathcal{A}_{\leq p\delta}}^* \right\|_{L_2} + \sum_{\boldsymbol{T} \in \mathcal{A}_{p\delta}^+(r)} d^{-\kappa(\boldsymbol{T})} \cdot O_{d,\mathbb{P}}(d^{-\lambda_\iota/2}) \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon).$$

$$\tag{138}$$

Combining this expression with (125) completes the proof.

B.6 Derivative norm estimator

The proof is analogous to the proof for equation (87) hence we adopt the notation used in Appendix B.5. We will estimate the following:

$$w_r \cdot \mathcal{D}_r(w) := \frac{w_r}{nd} y^\top K_{w,\lambda}^{-1} X_r K'_w X_r K_{w,\lambda}^{-1} y \tag{139}$$

With the expression (122) in place of K'_w, equation (139) becomes

$$w_r \cdot \mathcal{D}_r(w) = \frac{w_r}{nd} \left\| \underbrace{(\widehat{D}'_{\mathcal{A}_{\leq p_\delta}})^{\frac{1}{2}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top X_r K_{w,\lambda}^{-1} y}_{:=T_1} \right\|_2^2 + \frac{w_r}{nd} \left\| \underbrace{(\rho' I_n + \Delta)^{\frac{1}{2}} X_r K_{w,\lambda}^{-1} y}_{:=T_2} \right\|_2^2 . \tag{140}$$

We will analyze each addend separately. We suppose $\iota = \pi_N(r)$, i.e., $r \in \mathcal{T}_\iota$ for some $\iota \in [N]$. Specifically, we will show

Claim 12. *The following estimates hold for any $\epsilon > 0$ uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

(a) $\frac{w_r}{nd} \|T_2 y\|_2^2 = O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2);$

(b) $\left| \frac{w_r}{nd} \|T_1 y\|_2^2 - \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} \Theta_d(d^{-|\kappa(\boldsymbol{T})|}) \cdot \|\partial_r f_{\boldsymbol{T}}\|_{L_2}^2 \right|$
$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|} \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2}^2$
$+ O_{d,\mathbb{P}}(d^{-\lambda_\iota/2 + \epsilon}) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|/2} \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon).$

Plugging Claim 12 (a,b) into Eq. (140) yields:

$$\left| w_r \mathcal{D}_r(w) - \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} \Theta_d(d^{-|\kappa(\boldsymbol{T})|}) \cdot \|\partial_r f_{\boldsymbol{T}}\|_{L_2}^2 \right|$$
$$= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|} \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2}^2$$
$$+ O_{d,\mathbb{P}}(d^{-\lambda_\iota/2 + \epsilon}) \cdot \sum_{\boldsymbol{T} \in \mathcal{A}_{\leq p_\delta} \sqcup \mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|/2} \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{141}$$

Similarly, the rest of the section is devoted to proving Claim 12. Throughout the proof, all asymptotic terms $o_{d,\mathbb{P}}$ and $O_{d,\mathbb{P}}$ that appear will always be uniform over the coordinates $r = 1, \ldots, d$.

B.6.1 Proof of Item (a) of Claim 12

The proof is almost identical to the proof of Claim 1 (a). We refer the readers to Section B.1.1 for more details. We can similarly obtain the following bound:

$$\frac{w_r}{nd} \|T_2 y\|_2^2 = O_{d,\mathbb{P}} \left(w_r \cdot \frac{\log(d)}{d} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) = O_{d,\mathbb{P}} \left(d^{-\lambda_\iota + \epsilon} \right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2),$$

where $\epsilon > 0$ is an arbitrarily small constant.

B.6.2 Proof of Item (b) of Claim 12

We denote $\beta = K_{w,\lambda}^{-1} y$. Following Eq. (126), we decompose $X_r \Phi_{\mathcal{A}_{\leq p_\delta}}$ into components:

$$\{\Phi_{\mathcal{S}_m^0}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \text{ and } \{\Phi_{\mathcal{S}_m^1}\}_{\boldsymbol{\lambda}^\top (\boldsymbol{m}-e_\iota) \leq p+\delta}. \tag{142}$$

57

We correspondingly decompose $\widehat{D}'_{\mathcal{A}_{\leq p_\delta}}$. We can show there exist diagonal matrices $\{R^0_{\boldsymbol{m}}\}_{\boldsymbol{m}}$ and $\{R^1_{\boldsymbol{m}}\}_{\boldsymbol{m}}$ such that the following holds:

$$
\left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \left(\frac{w_r}{nd} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \right)^{\frac{1}{2}} \right\|_2^2
$$
$$
= \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \| \beta^\top \Phi_{\mathcal{S}^0_{\boldsymbol{m}}} R^0_{\boldsymbol{m}} \|_2^2 + \sum_{\boldsymbol{\lambda}^\top (\boldsymbol{m}-e_\iota) \leq p+\delta} \| \beta^\top \Phi_{\mathcal{S}^1_{\boldsymbol{m}}} R^1_{\boldsymbol{m}} \|_2^2, \tag{143}
$$

Claim 13. *The following estimate holds:*

$$
(R^0_{\boldsymbol{m}})_{\boldsymbol{T},\boldsymbol{T}} = \left(g^{(\|\boldsymbol{m}\|_1+2)}(0) \cdot \frac{w_r^2}{nd} W_{\boldsymbol{T},\boldsymbol{T}} d^{-\|\boldsymbol{m}\|_1-1} (1 + O_d(d^{-1})) \right)^{\frac{1}{2}}, \quad \forall\, \boldsymbol{T} \in \mathcal{S}^0_{\boldsymbol{m}};
$$
$$
(R^1_{\boldsymbol{m}})_{\boldsymbol{T},\boldsymbol{T}} = \left(g^{(\|\boldsymbol{m}\|_1)}(0) \cdot \frac{1}{nd} W_{\boldsymbol{T},\boldsymbol{T}} d^{-\|\boldsymbol{m}\|_1+1} (1 + O_d(d^{-1})) \right)^{\frac{1}{2}}, \quad \forall\, \boldsymbol{T} \in \mathcal{S}^1_{\boldsymbol{m}}. \tag{144}
$$

Proof. Recall that D' is a diagonal matrix with entry $D'_{\mathcal{S}_{\boldsymbol{m}}} = g^{(\|\boldsymbol{m}\|_1+1)}(0) d^{-\|\boldsymbol{m}\|_1} + O_d(d^{-\|\boldsymbol{m}\|_1-1})$. Based on Eq. (143), we can see that $\Phi_{\mathcal{S}^0_{\boldsymbol{m}}} R^0_{\boldsymbol{m}}$ corresponds to $x_r \cdot \Phi_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}} (\frac{w_r^2}{nd} \widehat{D}'_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}})^{\frac{1}{2}}$ since $\phi_{\mathcal{S}^0_{\boldsymbol{m}}} = x_r \phi_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}}$. Then we can deduce that

$$
R^0_{\boldsymbol{m}} = \left(\frac{w_r}{nd} W_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}} D'_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}} \right)^{\frac{1}{2}} = \left(\frac{w_r^2}{nd} W_{\mathcal{S}^0_{\boldsymbol{m}}} D'_{\mathcal{S}^1_{\boldsymbol{m}+e_\iota}} \right)^{\frac{1}{2}}.
$$

Plugging the expression of D' into the above equation completes the proof for Eq. (130). A similar analysis applies to

$$
R^1_{\boldsymbol{m}} = \left(\frac{w_r}{nd} W_{\mathcal{S}^0_{\boldsymbol{m}-e_\iota}} D'_{\mathcal{S}^0_{\boldsymbol{m}-e_\iota}} \right)^{\frac{1}{2}} = \left(\frac{1}{nd} W_{\mathcal{S}^1_{\boldsymbol{m}}} D'_{\mathcal{S}^0_{\boldsymbol{m}-e_\iota}} \right)^{\frac{1}{2}}.
$$

$\square$

Regarding Eq. (143), we group the features based on the inequalities (1) $\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p + \delta$ and (2) $\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta$. Then we can define a diagonal matrix R with each block $R_{\mathcal{S}^1_{\boldsymbol{m}},\mathcal{S}^1_{\boldsymbol{m}}} = R^1_{\boldsymbol{m}}$ and $R_{\mathcal{S}^0_{\boldsymbol{m}},\mathcal{S}^0_{\boldsymbol{m}}} = R^0_{\boldsymbol{m}}$ such that Eq. (143) can be written as

$$
\left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \left(\frac{w_r}{nd} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \right)^{\frac{1}{2}} \right\|_2^2 = \left\| \beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 + \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \in (p+\delta,p+\delta+\lambda_\iota]} \left\| \beta^\top \Phi_{\mathcal{S}^1_{\boldsymbol{m}}} R^1_{\boldsymbol{m}} \right\|_2^2. \tag{145}
$$

We start by analyzing the first term on the right side of (145). The proof of the proposition appears in Appendix C.5.1.

Proposition B.33. *For any $\epsilon > 0$, the following estimate holds uniformly for all coordinates $r \in \mathcal{T}_\iota$:*

$$
\left| \left\| \beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 - \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \Theta_d(d^{\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta}) \cdot \left\| \partial_r f^*_{\mathcal{S}^1_{\boldsymbol{m}}} \right\|_{L_2}^2 \right|
$$
$$
= O_{d,\mathbb{P}} \left(d^{-\lambda_\iota} \right) \cdot \| f^* \|_{L_2}^2 + \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta)}) \cdot \left\| \partial_r f^*_{\mathcal{S}^1_{\boldsymbol{m}}} \right\|_{L_2}^2
$$
$$
+ O_{d,\mathbb{P}} \left(d^{-\lambda_\iota/2} \right) \cdot \| f^* \|_{L_2} \left\| \partial_r f^*_{\mathcal{A}_{\leq p_\delta}} \right\|_{L_2}, \tag{146}
$$

We proceed to analyze the second term on the right side of (145). The proof appears in Appendix C.5.2.

Proposition B.34. *For any $\epsilon > 0$, the estimate holds uniformly for all coordinates $r \in \mathcal{T}_\iota$ and $\boldsymbol{m} \in \mathbb{R}^N$ that satisfies $\boldsymbol{\lambda}^\top \boldsymbol{m} \in (p + \delta, p + \delta + \lambda_\iota]$:*

$$\left| \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 \right\|_2^2 - \Theta_d(d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2^2 \right|$$

$$= O_{d,\mathbb{P}}(d^{-\lambda_\iota+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}(d^{(p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m})/2-\lambda_\iota/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{147}$$

Following the triangle inequality, the second term in Eq. (145) can be approximated as

$$\left| \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m}\in(p+\delta,p+\delta+\lambda_\iota]} \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 \right\|_2^2 - \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m}\in(p+\delta,p+\delta+\lambda_\iota]} \Theta_d(d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2^2 \right|$$

$$= O_{d,\mathbb{P}}(d^{-\lambda_\iota+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m}\in(p+\delta,p+\delta+\lambda_\iota]} d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}^2$$

$$+ \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m}\in(p+\delta,p+\delta+\lambda_\iota]} O_{d,\mathbb{P}}(d^{(p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m})/2-\lambda_\iota/2+\epsilon}) \cdot \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{148}$$

Lastly, Eq. (146) and (148) together give

$$\left| \left\| \beta^\top X_r \Phi_{\mathcal{A}_{\leq p_\delta}} \left(\frac{w_r}{nd} \widehat{D}'_{\mathcal{A}_{\leq p_\delta}} \right)^{\frac{1}{2}} \right\|_2^2 - \sum_{\boldsymbol{T}\in\mathcal{A}_{\leq p_\delta}\sqcup\mathcal{A}_{p_\delta}^+} \Theta_d(d^{-|\kappa(\boldsymbol{T})|}) \cdot \|\partial_r f_{\boldsymbol{T}}\|_{L_2}^2 \right|$$

$$= O_{d,\mathbb{P}}\left(d^{-\lambda_\iota+\epsilon}\right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{T}\in\mathcal{A}_{\leq p_\delta}\sqcup\mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2}^2$$

$$+ \sum_{\boldsymbol{T}\in\mathcal{A}_{\leq p_\delta}\sqcup\mathcal{A}_{p_\delta}^+} d^{-|\kappa(\boldsymbol{T})|/2} \cdot O_{d,\mathbb{P}}(d^{-\lambda_\iota/2+\epsilon}) \cdot \|\partial_r f_{\boldsymbol{T}}^*\|_{L_2} (\|f^*\|_{L_2} + \sigma_\varepsilon). \tag{149}$$

Then we complete the proof.

B.7 Proof of Theorem B.27 (Generalization bound)

We denote $z^{(i)} = \sqrt{w} \odot x^{(i)}$. Plugging the expression

$$\hat{f}_w(x) = K_w(x, X) K_{w,\lambda}^{-1} y = K(\sqrt{w} \odot x, Z) K_{w,\lambda}^{-1} y$$

into the left-hand side of Eq. (86) gives

$$\mathbb{E}_x \left[\left(\hat{f}_w(x) - f_{\mathcal{A}_{<p_\delta}}^*(x) - \sqrt{c} \cdot f_{\mathcal{A}_{=p_\delta}}^*(x) \right)^2 \right]$$

$$= \underbrace{y^\top K_{w,\lambda}^{-1} \mathbb{E}_x[K(Z, \sqrt{w} \odot x) K(\sqrt{w} \odot x, Z)] K_{w,\lambda}^{-1} y}_{:=T_1} + \left(\mathbb{E}_x \left[f_{\mathcal{A}_{<p_\delta}}^*(x)^2 \right] + c \cdot \mathbb{E}_x \left[f_{\mathcal{A}_{=p_\delta}}^*(x)^2 \right] \right)$$

$$- 2 \underbrace{\mathbb{E}_x \left[(f_{\mathcal{A}_{<p_\delta}}^*(x) + \sqrt{c} \cdot f_{\mathcal{A}_{=p_\delta}}^*(x)) K(\sqrt{w} \odot x, Z) \right] K_{w,\lambda}^{-1} y}_{:=T_2},$$

where we denote $K_{w,\lambda} = K_\lambda(Z, Z) = K(Z, Z) + \lambda I_n$ and $c > 0$ is a constant.

We will analyze each summand separately. Specifically, we will verify the following claim, from which the above equation follows immediately.

Claim 14. *There exists a constant $c > 0$ such that the following estimates hold*

(a) $\left| T_1 - \mathbb{E}_x \left[f^*_{\mathcal{A}_{<p_\delta}}(x)^2 \right] - c \cdot \mathbb{E}_x \left[f^*_{\mathcal{A}_{=p_\delta}}(x)^2 \right] \right| = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|^2_{L_2} + \sigma^2_\varepsilon),$

(b) $\left| T_2 - \mathbb{E}_x \left[f^*_{\mathcal{A}_{<p_\delta}}(x)^2 \right] - c \cdot \mathbb{E}_x \left[f^*_{\mathcal{A}_{=p_\delta}}(x)^2 \right] \right| = o_{d,\mathbb{P}}(1) \cdot (\|f^*\|^2_{L_2} + \sigma^2_\varepsilon).$

Before presenting the proof of the claims, we give a useful orthogonal decomposition of the kernel. The proof appears in Appendix C.6.1.

Lemma B.35. *Let $g : \mathbb{R} \to \mathbb{R}$ satisfy assumption 1 and fix $w \in \mathbb{R}^d_+$ with $\|w\|_\infty = O_d(d^\gamma)$ for some $\gamma < 1$. For every $x, y \in \mathbb{H}^d := \{\pm 1\}^d$, endowed with the uniform measure τ_d, the kernel $K_w(x, y) := g(x^\top \mathsf{Diag}(w)y/d)$ admits an orthogonal expansion in $L_2(\mathbb{H}^d \times \mathbb{H}^d, \tau_d \otimes \tau_d)$:*

$$K_w(x, y) = \sum_{r=0}^{\infty} \sum_{S:|S|=r} \lambda_S \phi_S(x)^\top \phi_S(y), \tag{150}$$

where $\phi_S(x) := \prod_{i \in S} x_i$ and $\lambda_S = (1 + O_d(d^{-2(1-\gamma)})) \cdot g^{(|S|)}(0)d^{-|S|}w^S$.

B.7.1 Proof of Claim (a)

We define diagonal matrices $\Lambda_{\mathcal{A}_{\leq p_\delta}}$ and V with each entry

$$\Lambda_{S,S} = \lambda_S, \quad V_{S,S} = \frac{n\lambda_s}{n\lambda_s + (\rho + \lambda)}, \tag{151}$$

for $S \in \mathcal{A}_{\leq p_\delta}$, where λ_S is defined in Proposition B.35. We proceed to define the constant c as

$$c := \frac{\sum_{S \in \mathcal{A}_{=p_\delta}} V^2_{S,S}b^2_S}{\sum_{S \in \mathcal{A}_{=p_\delta}} b^2_S}. \tag{152}$$

For simplicity of notation, we denote

$$M := \mathbb{E}_x[K(Z, \sqrt{w} \odot x)K(\sqrt{w} \odot x, Z)].$$

Similar to Lemma B.30, we show M can be approximated using Fourier features. The proof appears in Appendix C.6.2.

Proposition B.36. *The following estimate holds*

$$\mathbb{E}_x[K(Z, \sqrt{w} \odot x)K(\sqrt{w} \odot x, Z)] = \underbrace{\Phi_{\mathcal{A}_{\leq p_\delta}}(X)\Lambda^2_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}(X)^\top}_{:=M_{\mathcal{A}_{\leq p_\delta}}} + \Delta_M, \tag{153}$$

where $\|\Delta_M\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-p-\delta-\zeta_2}).$

Consequently, we decompose T_1 into

$$T_1 = \underbrace{y^\top K^{-1}_{w,\lambda}M_{\mathcal{A}_{\leq p_\delta}}K^{-1}_{w,\lambda}y}_{T_{11}} + \underbrace{y^\top K^{-1}_{w,\lambda}\Delta_M K^{-1}_{w,\lambda}y}_{T_{12}}.$$

A simple calculation shows $|T_{12}| = o_{d,\mathbb{P}}(1)$:

$$\begin{aligned} |T_{12}| &\leq \|y\|^2_2 \left\| K^{-1}_{w,\lambda} \right\|^2_{\mathrm{op}} \|\Delta_M\|_{\mathrm{op}} \\ &= O_d(n\log(d)) \cdot (\|f^*\|^2_{L_2} + \sigma^2_\varepsilon) \cdot O_{d,\mathbb{P}}(1) \cdot O_{d,\mathbb{P}}(d^{-p-\delta-\zeta_2}) \\ &= O_{d,\mathbb{P}}(d^{-\zeta_2}\log(d)) \cdot (\|f^*\|^2_{L_2} + \sigma^2_\varepsilon). \end{aligned} \tag{154}$$

60

where the bound $\left\|K_{w,\lambda}^{-1}\right\|_{\mathrm{op}} \leq 1/(\rho + \lambda - \|\Delta_1\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(1)$ is implied by the expression (121) and $\|y\|_2^2 = O_{d,\mathbb{P}}(n \log(d)) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$ follows from Proposition B.21.

Next we proceed to analyze T_{11}, which can be estimated by the proposition below. The proof can be found in Appendix C.6.3.

Proposition B.37. *The following estimate holds for any $\epsilon > 0$:*

$$\left| T_{11} - \mathbb{E}_x \left[f^*_{\mathcal{A}_{<p_\delta}}(x)^2 \right] - c \cdot \mathbb{E}_x \left[f^*_{\mathcal{A}_{=p_\delta}}(x)^2 \right] \right|$$
$$= O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + d^{-2+2\gamma_{\max}+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2), \tag{155}$$

where $\gamma_{\max} := \max_{k \in [N]} \gamma_k$.

Combining Eq. (154) and (155) completes the proof.

B.7.2 Proof of Claim (b)

The claim is directly implied by the following proposition. See the proof in Appendix C.6.4.

Proposition B.38. *The following estimate holds for any $\epsilon > 0$:*

$$\left| T_2 - \mathbb{E}_x \left[f^*_{\mathcal{A}_{<p_\delta}}(x)^2 \right] - c \cdot \mathbb{E}_x \left[f^*_{\mathcal{A}_{=p_\delta}}(x)^2 \right] \right| = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2). \tag{156}$$

We have finished the proof for Claim (a) and (b) thus completing the proof.

C Proof of propositions for main results

C.1 Missing proofs for Section A.1 and A.2

C.1.1 Proof of Lemma A.11 and Theorem A.12.

We will prove Lemmas A.11 and A.12 simultaneously. To this end, following the setting of Lemma A.11 consider the regime $n = d^{p+\delta}$ where $\delta \in (0,1)$ is a constant. Fix a set $\mathcal{S} \subseteq \bar{\mathcal{S}}$ indexing polynomials of degree at most l. Note that the rows of $\Phi_\mathcal{S}$ are independent isotropic random vectors in $\mathbb{R}^n$. Therefore, applying [70, Theorem 5.45] yields the estimate

$$\mathbb{E}\left[\left\| \tfrac{1}{n} \Phi_\mathcal{S}^\top \Phi_\mathcal{S} - I_{|\mathcal{S}|} \right\|_{\mathrm{op}} \right] \leq \frac{C_1 \sqrt{m \log \min(|\mathcal{S}|, n)}}{\sqrt{n}}, \tag{157}$$

where we set $m := \mathbb{E}\left[\max_{i \leq n} \left\| \phi_\mathcal{S}(x^{(i)}) \right\|^2 \right]$ and $C_1 > 0$ is an absolute constant.

We next upper bound m. To this end, to simplify notation we let $\phi_{i,\mathcal{S}}$ be shorthand for $\phi_\mathcal{S}(x^{(i)})$. Now, for any $q > 1$, we successively estimate

$$\frac{m}{|S|} = \frac{1}{|S|} \mathbb{E}\left[\max_{i \leq n} \|\phi_{i,\mathcal{S}}\|^2 \right]$$

$$\leq \mathbb{E}\left[\max_{i \leq n} \max_{j \in \mathcal{S}} (\phi_{i,j})^2 \right] \tag{158}$$

$$\leq \left(\mathbb{E}\left[\max_{i \leq n} \max_{j \in \mathcal{S}} (\phi_{i,j})^{2q} \right] \right)^{1/q} \tag{159}$$

$$\leq \left(\sum_{i \leq n,\ j \in \mathcal{S}} \mathbb{E}(\phi_{i,j})^{2q} \right)^{1/q} \tag{160}$$

$$\leq (n|S|)^{1/q} \max_{i \leq n,\ j \in \mathcal{S}} \mathbb{E}\left[(\phi_{i,j})^{2q} \right]^{1/q}, \tag{161}$$

where (158) and (160) follow from bounding the maximum by a sum and (159) follows from Lyapunov's Inequality. The hypercontractivity Assumption 4 implies

$$\mathbb{E}\left[(\phi_{i,j})^{2q}\right]^{1/q} \leq C_{2q,l}^2 \mathbb{E}\left[(\phi_{i,j})^2\right] = C_{2q,l}^2. \tag{162}$$

Consequently, plugging Eq. (162) into Eq. (161) yields

$$m \leq n^{1/q}|\mathcal{S}|^{1+1/q}C_{2q,l}^2. \tag{163}$$

Plugging Eq. (163) into Eq. (157), we obtain

$$\mathbb{E}\left[\left\|\tfrac{1}{n}\Phi_{\mathcal{S}}^{\top}\Phi_{\mathcal{S}} - I_{|\mathcal{S}|}\right\|_{\mathrm{op}}\right] \leq \frac{C_1 C_{2q,l}\sqrt{n^{1/q}|\mathcal{S}|^{1+1/q}\log\min(|S|,n)}}{\sqrt{n}}. \tag{164}$$

In particular, we deduce the bound on the operator norm

$$\mathbb{E}\|\Phi_{\mathcal{S}}\|_{\mathrm{op}}^2 \leq n + C_1 C_{2q,l}\sqrt{(n|\mathcal{S}|)^{1+1/q}\log\min(|S|,n)}$$

Let us look at two cases.

Case 1: large $|\mathcal{S}|$. Suppose that $|\mathcal{S}| \geq Cd^{p+\delta_0}$ with $\delta_0 > \delta$. Then simple arithmetic shows that there exists $\epsilon > 0$ such that the inequality $(n|\mathcal{S}|)^{1+1/q} \leq C^{-1}|\mathcal{S}|^{2-2\epsilon}$ holds for all large q. Therefore, we deduce

$$\mathbb{E}\|\Phi_{\mathcal{S}}\|_{\mathrm{op}}^2 \lesssim n + |\mathcal{S}|^{1-\epsilon}\sqrt{\log(n)} \lesssim |\mathcal{S}|d^{-(\delta_0-\delta)} + |\mathcal{S}|^{1-\epsilon}\sqrt{\log(n)}.$$

Markov's inequality therefore ensures $\|\Phi_{\mathcal{S}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{|\mathcal{S}|})$ as claimed in item 2 of Lemma A.11.

Case 2: small $|\mathcal{S}|$. Suppose that $|\mathcal{S}| \leq Cd^{p+\delta_0}$ with $\delta_0 < \delta$. Then from (164), we obtain

$$\mathbb{E}\left[\left\|\tfrac{1}{n}\Phi_{\mathcal{S}}^{\top}\Phi_{\mathcal{S}} - I_{|\mathcal{S}|}\right\|_{\mathrm{op}}\right] = C_1 C_{2q,l}C^{\frac{1}{2}+\frac{1}{2q}}d^{-(\frac{\delta-\delta_0}{2}-\frac{2p+\delta+\delta_0}{2q})}\sqrt{\log\min(Cd^p,n)}. \tag{165}$$

Therefore for any $\epsilon > 0$, by choosing δ sufficiently large, we may ensure the estimate:

$$\mathbb{E}\left[\left\|\Phi_{\mathcal{S}}^{\top}\Phi_{\mathcal{S}}/n - I_{|\mathcal{S}|}\right\|_{\mathrm{op}}\right] \lesssim d^{-(\delta-\delta_0-\epsilon)/2}\sqrt{\log(n)}.$$

Item 1 of Lemma A.11 now follows directly from Markov's inequality and the triangle inequality. Similarly, setting $\delta' = \delta - \delta_0 - \epsilon$ and applying Markov's inequality, then absorbing the logarithmic factors into ϵ concludes the proof of Lemma A.12.

C.1.2 Proof of Lemma A.13

In the proof, we omit the subscript of Φ to simplify notation and set $\phi_i = \phi(x^{(i)})$. Define the off-diagonal entries of $\frac{1}{|\mathcal{S}|}\Phi\Phi^{\top}$ by

$$\Delta_{i,j} := \left\{ \begin{array}{ll} \frac{1}{|\mathcal{S}|}\phi_i^{\top}\phi_j, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{array} \right.$$

We now split $\frac{1}{|\mathcal{S}|}\Phi\Phi^{\top}$ into the diagonal and off-diagonal parts. Namely, the triangle inequality yields

$$\mathbb{E}\left[\left\|\tfrac{1}{|\mathcal{S}|}\Phi\Phi^{\top} - I_n\right\|_{\mathrm{op}}\right] \leq \mathbb{E}\left[\left\|\tfrac{1}{|\mathcal{S}|}\mathsf{Diag}(\Phi\Phi^{\top}) - I_n\right\|_{\mathrm{op}}\right] + \mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]. \tag{166}$$

We will show that both terms on the RHS of Eq. (166) are small.

Bounding $\mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}|}\mathsf{Diag}(\Phi\Phi^\top) - I_n\right\|_{\mathrm{op}}\right]$. Using Lyapunov's inequality for any $q \geq 1$, we deduce

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tfrac{1}{|\mathcal{S}|}\mathsf{Diag}(\Phi\Phi^\top) - I_n\right\|_{\mathrm{op}}\right] &= \mathbb{E}\left[\max_i \left|\|\phi_i\|^2/|\mathcal{S}| - 1\right|\right] \\
&\leq \left(\mathbb{E}\left[\max_i \left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^{2q}\right]\right)^{1/(2q)} \\
&\leq \left(\sum_{i=1}^n \mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^{2q}\right]\right)^{1/(2q)} \\
&= n^{1/(2q)}\left(\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^{2q}\right]\right)^{1/(2q)}.
\end{aligned}
\tag{167}
$$

The hypercontractivity Assumption 4 then implies

$$
\left(\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^{2q}\right]\right)^{1/(2q)} \leq C_{2l,2q}\left(\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^2\right]\right)^{1/2}.
\tag{168}
$$

Taking into account the equality $\mathbb{E}\left[(\phi_i)_s(\phi_i)_t\right] = \delta_{st}$, we deduce

$$
\begin{aligned}
\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^2\right] &= \mathbb{E}\left[\|\phi_i\|^4\right]/|\mathcal{S}|^2 - 1 \\
&= \frac{1}{|\mathcal{S}|^2}\left(|\mathcal{S}|\cdot\mathbb{E}\left[(\phi_i)_s^4\right] + |\mathcal{S}|(|\mathcal{S}| - 1)\right) - 1 \\
&= \left(\mathbb{E}\left[(\phi_i)_s^4\right] - 1\right)/|\mathcal{S}|.
\end{aligned}
$$

The hypercontractivity Assumption 4 again implies

$$
\mathbb{E}\left[(\phi_i)_s^4\right] \leq C_{l,4}^4 \cdot \mathbb{E}[(\phi_i)_s^2]^2 = C_{l,4}^4.
$$

Therefore, we conclude

$$
\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^2\right] \leq \left(C_{l,4}^4 - 1\right)/|\mathcal{S}|.
\tag{169}
$$

Plugging Eq. (169) into Eq. (168) we obtain

$$
\left(\mathbb{E}\left[\left(\|\phi_i\|^2/|\mathcal{S}| - 1\right)^{2q}\right]\right)^{(1/2q)} \leq C_{2l,2q}\left(\left(C_{l,4}^4 - 1\right)/|\mathcal{S}|\right)^{1/2}.
\tag{170}
$$

Plugging Eq. (170) into Eq. (167) yields the estimates

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tfrac{1}{|\mathcal{S}|}\mathsf{Diag}(\Phi\Phi^\top) - I_n\right\|_{\mathrm{op}}\right] &\leq n^{1/(2q)}C_{2l,2q}\left(\left(C_{l,4}^4 - 1\right)/|\mathcal{S}|\right)^{1/2} \\
&\leq C_{2l,2q}(C_{l,4}^4 - 1)^{1/2}C^{-1/2}d^{\frac{p+\delta}{2q} - \frac{p+\delta}{2}},
\end{aligned}
\tag{171}
$$

where we use the fact that $|\mathcal{S}| \geq Cd^{p+\delta_0}$.

Bound $\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]$. Now we consider the second term in Eq. (166). We first apply the matrix decoupling from [70] in Lemma 5.60. For $T_1, T_1 \subseteq [n]$, we define $\Delta_{T_1,T_2} = (\Delta_{i,j})_{i\in T_1, j\in T_2}$. By Lemma D.68, we have

$$
\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] \leq 4\sup_{T\subseteq[n]}\mathbb{E}\left[\|\Delta_{T,T^c}\|_{\mathrm{op}}\right],
$$

where T^c denotes the complement of T.

For any set $J \subset [n]$, we let $\mathbb{E}_J$ denote the expectation with respect to $\{x^{(i)}\}_{i \in J}$ and conditionally on $\{x^{(i)}\}_{i \in J^c}$. We fix $T \in [n]$ and apply Lemma D.69 conditionally on $\{x^{(i)}\}_{i \in T^c}$, thereby obtaining

$$\mathbb{E}_T \left[\|\Delta_{T,T^c}\|_{\mathrm{op}} \right] \leq (\Sigma(T) \cdot n)^{1/2} + c \cdot (\Gamma(T) \cdot \log(n))^{1/2}, \tag{172}$$

where $\Sigma(T) := \left\| \mathbb{E}_{x^{(j)}} \left[\Delta_{T^c,j} \Delta_{T^c,j}^\top \right] \right\|_{\mathrm{op}}$ for some $j \in T$ and $\Gamma(T) := \mathbb{E}_T \left[\max_{j \in T} \|\Delta_{j,T^c}\|^2 \right]$.

Using the tower rule, We split $\mathbb{E}$ into $\mathbb{E}_{T^c} \mathbb{E}_T$ and deduce

$$\begin{aligned}
\mathbb{E} \left[\|\Delta\|_{\mathrm{op}} \right] &\leq 4 \sup_{T \subseteq [n]} \mathbb{E} \left[\|\Delta_{T,T^c}\|_{\mathrm{op}} \right] \\
&= 4 \sup_{T \subseteq [n]} \mathbb{E}_{T^c} \mathbb{E}_T \left[\|\Delta_{T,T^c}\|_{\mathrm{op}} \right] \\
&\leq 4 \sup_{T \subseteq [n]} \mathbb{E}_{T^c} \left[(\Sigma(T) \cdot n)^{1/2} \right] + c \cdot 4 \sup_{T \subseteq [n]} \mathbb{E}_{T^c} \left[(\Gamma(T) \cdot \log(n))^{1/2} \right] \\
&\leq 4 n^{1/2} \sup_{T \subseteq [n]} (\mathbb{E}_{T^c} [\Sigma(T)])^{1/2} + 4c \log^{1/2}(n) \cdot \sup_{T \subseteq [n]} (\mathbb{E}_{T^c} [\Gamma(T)])^{1/2}.
\end{aligned}$$
$$\tag{173}$$
$$\tag{174}$$

where (173) follows from (172) and (174) follows from Jensen's inequality.

We next bound separately the two terms on the right side of (174).

Step 1: Bound $\mathbb{E}_{T^c} [\Sigma(T)]$. Note that each entry of $\Delta_{T^c,j} \Delta_{T^c,j}^\top$ takes the form $\phi_i^\top \phi_j \phi_j^\top \phi_l / |\mathcal{S}|^2$ for $i, l \in T^c, j \in T$. Since $\mathbb{E}_{x^{(j)}} \left[\phi_j \phi_j^\top \right] = I_{|\mathcal{S}|}$, we have each entry of $\mathbb{E}_{x^{(j)}} \left[\Delta_{T^c,j} \Delta_{T^c,j}^\top \right]$ takes the form $\phi_i^\top \phi_l / |\mathcal{S}|^2$. Therefore, we deduce

$$\Sigma(T) = \left\| \mathbb{E}_{x^{(j)}} \left[\Delta_{T^c,j} \Delta_{T^c,j}^\top \right] \right\|_{\mathrm{op}} = \left\| (\Phi\Phi^\top)_{T^c,T^c} \right\|_{\mathrm{op}} / |\mathcal{S}|^2 \leq \mathbb{E}_T \left[\|\Phi\Phi^\top\|_{\mathrm{op}} \right] / |\mathcal{S}|^2,$$

where in the last inequality, we use Cauchy's interlacing theorem [41]. Specifically, the theorem shows that adding a row and a column to a Hermitian matrix will only increase its operator norm.

Note that splitting $\Phi\Phi^\top$ into its diagonal and off-diagonal parts yields the estimate

$$\mathbb{E}_T \left[\|\Phi\Phi^\top\|_{\mathrm{op}} \right] / |\mathcal{S}|^2 \leq \mathbb{E}_T \left[\|\Delta\|_{\mathrm{op}} \right] / |\mathcal{S}| + \mathbb{E}_T \left[\max_{j \in [n]} \|\phi_j\|^2 \right] / |\mathcal{S}|^2.$$

Taking the expectation and we have

$$\mathbb{E}_{T^c} [\Sigma(T)] \leq \mathbb{E} \left[\|\Delta\|_{\mathrm{op}} \right] / |\mathcal{S}| + \mathbb{E} \left[\max_{j \in [n]} \|\phi_j\|^2 \right] / |\mathcal{S}|^2. \tag{175}$$

We control the second term using Lyapunov's inequality: for any $q \geq 1$, we have

$$\begin{aligned}
\mathbb{E} \left[\max_{j \in [n]} \|\phi_j\|^2 \right] &\leq |\mathcal{S}| \cdot \mathbb{E} \left[\max_{i \in \mathcal{S}} \max_{j \in [n]} (\phi_j)_i^2 \right] \\
&\leq |\mathcal{S}| \cdot \left(\mathbb{E} \left[\max_{i \in \mathcal{S}, j \in [n]} (\phi_j)_i^{2q} \right] \right)^{1/q} \\
&\leq |\mathcal{S}| \cdot \left(\sum_{i \in \mathcal{S}} \sum_{j \in [n]} \mathbb{E} \left[(\phi_j)_i^{2q} \right] \right)^{1/q} \\
&= n^{1/q} |\mathcal{S}|^{1+1/q} \left(\mathbb{E} \left[(\phi_j)_i^{2q} \right] \right)^{1/q}.
\end{aligned}$$
$$\tag{176}$$

Using the hypercontractivity Assumption 4, we obtain

$$\left(\mathbb{E} \left[(\phi_j)_i^{2q} \right] \right)^{1/q} \leq C_{l,2q}^2 \left(\mathbb{E} \left[(\phi_j)_i^2 \right] \right) = C_{l,2q}^2.$$

Consequently, we deduce

$$\mathbb{E}\left[\max_{j\in[n]}\|\phi_j\|^2\right] \leq C_{l,2q}^2 |\mathcal{S}|^{1+1/q} n^{1/q}. \tag{177}$$

Plugging this estimate into Eq. (175) yields

$$\mathbb{E}_{T^c}\left[\Sigma(T)\right] \leq \mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]/|\mathcal{S}| + C_{l,2q}^2 |\mathcal{S}|^{(1/q)-1} n^{1/q}. \tag{178}$$

Step 2: Bound $\mathbb{E}_{T^c}\left[\Gamma(T)\right]$. Using Lyapunov's inequality, for any $q \geq 1$, we have

$$\begin{aligned}
\mathbb{E}_{T^c}\mathbb{E}_T\left[\max_{i\in T}\|\Delta_{i,T^c}\|^2\right] &= \mathbb{E}\left[\max_{i\in T}\|\Delta_{i,T^c}\|^2\right] \\
&\leq n\mathbb{E}\left[\max_{i\in T}\max_{j\in T^c}\Delta_{i,j}^2\right] \\
&\leq n\mathbb{E}\left[\max_{i\in T}\max_{j\in T^c}\Delta_{i,j}^{2q}\right]^{1/q} \\
&\leq n\left(\sum_{i\in T}\sum_{j\in T^c}\mathbb{E}\left[\Delta_{i,j}^{2q}\right]\right)^{1/q} \\
&\leq n^{1+2/q}\mathbb{E}\left[\Delta_{i,j}^{2q}\right]^{1/q}. \tag{179}
\end{aligned}$$

Since $\Delta_{i,j}$ is a degree $2l$ polynomial, the hypercontractivity Assumption 4 implies

$$\mathbb{E}\left[\Delta_{i,j}^{2q}\right]^{1/q} \leq C_{2l,2q}^2 \mathbb{E}\left[\Delta_{i,j}^2\right] = \frac{C_{2l,2q}^2}{|\mathcal{S}|}. \tag{180}$$

Consequently, plugging the above inequality into Eq. (179) we have

$$\mathbb{E}_{T^c}\left[\Gamma(T)\right] \leq \frac{n^{1+2/q}C_{2l,2q}^2}{|\mathcal{S}|}. \tag{181}$$

Combining Eq. (178) and Eq. (181) with Eq. (174) we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] &\leq 4\left(\frac{n}{|\mathcal{S}|}\right)^{1/2}\left(\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] + C_{l,2q}^2|\mathcal{S}|^{1/q}n^{1/q}\right)^{1/2} + 4cC_{2l,2q}^2\frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{1/2}(n)}{\sqrt{|\mathcal{S}|}} \\
&\leq 4\left(\frac{n}{|\mathcal{S}|}\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]\right)^{\frac{1}{2}} + 4C_{l,2q}n^{\frac{1}{2}+\frac{1}{2q}}|\mathcal{S}|^{\frac{1}{2q}-\frac{1}{2}} + 4cC_{2l,2q}^2\frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{\frac{1}{2}}(n)}{\sqrt{|\mathcal{S}|}},
\end{aligned}$$

where the last inequality is obtained from the elementary estimate $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a,b \geq 0$. Solving the quadratic inequality with respect to $\sqrt{\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]}$ we obtain:

$$\sqrt{\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]} \leq 2\left(\frac{n}{|\mathcal{S}|}\right)^{1/2} + 2\sqrt{\frac{n}{|\mathcal{S}|} + C_{l,2q}n^{\frac{1}{2}+\frac{1}{2q}}|\mathcal{S}|^{\frac{1}{2q}-\frac{1}{2}} + cC_{2l,2q}^2\frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{1/2}(n)}{\sqrt{|\mathcal{S}|}}}.$$

Squaring both sides and using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we conclude

$$\frac{1}{8}\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] \leq \frac{2n}{|\mathcal{S}|} + C_{l,2q}n^{\frac{1}{2}+\frac{1}{2q}}|\mathcal{S}|^{\frac{1}{2q}-\frac{1}{2}} + cC_{2l,2q}^2\frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{1/2}(n)}{\sqrt{|\mathcal{S}|}}$$

Taking into account that $|\mathcal{S}| \gtrsim d^{p+\delta_0}$ and $n = d^{p+\delta}$, we therefore conclude

$$\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] \lesssim d^{\delta-\delta_0} + C_{l,2q}d^{\frac{\delta-\delta_0}{2}+\frac{2p+\delta+\delta_0}{2q}} + C_{l,2q}^2\sqrt{\log(n)}d^{\frac{\delta-\delta_0}{2}+\frac{p+\delta}{q}}.$$

Combining this estimate with (166) and (171) we conclude

$$\mathbb{E}\left[\left\|\Phi\Phi^\top/|\mathcal{S}| - I_n\right\|_{\mathrm{op}}\right] \lesssim C_{2l,2q}(C_{l,4}^4-1)^{1/2}d^{\frac{p+\delta}{2q}-\frac{p+\delta}{2}} + d^{\delta-\delta_0} + C_{2l,2q}d^{\frac{\delta-\delta_0}{2}+\frac{2p+\delta+\delta_0}{2q}}$$
$$+ C_{2l,2q}^2\sqrt{\log(n)}d^{\frac{\delta-\delta_0}{2}+\frac{p+\delta}{q}}.$$

Clearly, for any $\epsilon > 0$ we may choose q sufficiently large so that the right side is bounded by a constant multiple of

$$d^{-\frac{p+\delta}{2}+2\epsilon} + \sqrt{\log(n)} \cdot d^{-\frac{\delta_0-\delta}{2}+\epsilon}, \tag{182}$$

thereby completing the proof of (41). In the case, $p \geq 1$, $\delta_0 = 1$, and $\epsilon \in (0, \delta)$ the first term in (182) is negligible compared to the second; applying Markov's inequality and absorbing logarithmic factors into ϵ completes the proof.

C.1.3 Proof of Theorem A.14

First, applying Lemma D.65 we deduce that for any $s \in (0, 1)$, the estimate

$$\max_{i,j}\left|\frac{1}{d}\left\langle x^{(i)}, x^{(j)}\right\rangle - \delta_{ij}\right| \leq s\sigma^2, \tag{183}$$

holds with probability at least $1 - 2n^2\exp\left(-c_1 s^2 d\right)$, where $c_1 > 0$ is some constant. Similarly, using Lemma D.67 and taking a union bound over $i = 1, \ldots, n$ we deduce that for any $t \in (0, \sigma^{-2})$ the estimate

$$\left|\frac{1}{d^q}\|x^{(i)}\|_2^{2q} - 1\right| \leq q2^{2q}t\sigma^2 \qquad \forall q \geq 1, \tag{184}$$

holds with probability at least $1 - 2n\exp\left(-c_2 dt^2\right)$. We assume that the events (183) and (184) occur for the remainder of the proof.

To simplify notation, let $\bar{c} := g_m(1) = \sum_{k=0}^m \frac{g^{(k)}(0)}{k!}$ and define the matrix $E(X) := K(X, X) - K_m(X, X)$. Then splitting $E(X)$ into its diagonal and off-diagonal parts, we may estimate

$$\|E(X) - (g(1) - \bar{c})I_n\|_{\mathrm{op}} = \|E(X) - \mathsf{Diag}(E(X)) + \mathsf{Diag}(E(X)) - (g(1) - \bar{c})I_n\|_{\mathrm{op}}$$
$$\leq \|E(X) - \mathsf{Diag}(E(X))\|_{\mathrm{op}} + \|\mathsf{Diag}(E(X)) - (g(1) - \bar{c})I_n\|_{\mathrm{op}}$$
$$\leq \|E(X) - \mathsf{Diag}(E(X))\|_F + \max_{i=1,\ldots,n}|E(X)_{ii} - (g(1) - \bar{c})I_n|, \tag{185}$$

where the first inequality follows from the triangle inequality and the second follows the norm comparison $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_F$.

We consider the off-diagonal entries first. By Assumption 1, we can write

$$g(t) = \sum_{m=0}^\infty \frac{g^{(m)}(0)}{m!}t^m, \quad t \in (-\epsilon, \epsilon). \tag{186}$$

Setting $s < \frac{\min(1/2,\epsilon)}{\sigma^2}$, we successively estimate

$$\|E(X) - \mathsf{Diag}(E(X))\|_F^2 = \sum_{i \neq j} \left(g(x^{(i)}, x^{(j)}) - g_m(x^{(i)}, x^{(j)}) \right)^2$$

$$\leq n^2 \sum_{k=m+1}^{\infty} \left(\frac{g^{(k)}(0)}{(k)!} \right)^2 \max_{i \neq j} \left| \left\langle x^{(i)}, x^{(j)} \right\rangle / d \right|^{2k} \tag{187}$$

$$\leq \frac{n^2 \sup_{k \geq m+1} g^{(k)}(0)}{(m+1)!} \sum_{k=m+1}^{\infty} (s\sigma^2)^{2k} \tag{188}$$

$$\leq \frac{n^2 \sup_{k \geq m+1} g^{(k)}(0)}{(m+1)!} \cdot \frac{(s\sigma^2)^{2m+2}}{1 - 1/4}$$

$$\leq \frac{4n^2 \sup_{k \geq m+1} g^{(k)}(0)(s\sigma)^{2m+2}}{3(m+1)!}. \tag{189}$$

where(187) follows from (183) and (188) follows from the geometric series formula.

Moving on to the diagonal entries, we compute:

$$|E(X)_{ii} - (g(1) - \bar{c})I_n| = \left| g\left(\frac{1}{d} \left\| x^{(i)} \right\|^2 \right) - g(1) + \left(\bar{c} - g_m \left(\frac{1}{d} \left\| x^{(i)} \right\|^2 \right) \right) \right|$$

$$\leq \left| g\left(\frac{1}{d} \left\| x^{(i)} \right\|^2 \right) - g(1) \right| + \left| \bar{c} - g_m \left(\frac{1}{d} \left\| x^{(i)} \right\|^2 \right) \right|$$

$$\leq L_0 \left| \frac{1}{d} \left\| x^{(i)} \right\|^2 - 1 \right| + \sum_{k=1}^{m} \frac{g^{(k)}(0)}{k!} \left| \frac{1}{d^k} \left\| x^{(i)} \right\|^{2k} - 1 \right| \tag{190}$$

$$\leq L_0 s\sigma^2 + t\sigma^2 \sum_{k=1}^{m} \frac{g^{(k)}(0)2^{2k}}{(k-1)!} k, \tag{191}$$

where (190) follows from Lipschitz continuity of g with $L_0 := \sup_{t \in (1-\epsilon, 1+\epsilon)} |g'(t)|$ and (191) follows from (184).

Therefore combining (185), (189), and (191) we obtain the estimate

$$\|E(X) - (g(1) - \bar{c})I_n\|_{\mathrm{op}} \leq \sqrt{\frac{4n^2 \sup_{k \geq m+1} g^{(k)}(0)}{3(m+1)!}} (s\sigma)^{m+1} + L_0 s\sigma^2 + \sum_{k=1}^{m} \frac{g^{(k)}(0)2^{2k}}{(k-1)!} t\sigma^2. \tag{192}$$

Defining the constant $C_{g,m} := \max \left(\sqrt{\frac{4 \sup_{k \geq m+1} g^{(k)}(0)}{3(m+1)!}}, \sum_{k=1}^{m} \frac{g^{(k)}(0)2^{2k}}{(k-1)!}, 1 \right) < \infty$ and we therefore have the bound

$$\|E(X) - (g(1) - \bar{c})I_n\|_{\mathrm{op}} \leq C_{g,m} \cdot (n(s\sigma^2)^{m+1} + L_0 s\sigma^2 + t\sigma^2). \tag{193}$$

Then for any $C > 2$, we may set $s := \sqrt{C \log(n)/c_1 d}$ and $t := \sqrt{(C-1)\log(n)/c_2 d}$ thereby ensuring that the events (183) and (184) occur with probability at least $1 - \frac{4}{n^{C-2}}$. Plugging these values into (193) yields the estimate

$$\|E(X) - (g(1) - \bar{c})I_n\|_{\mathrm{op}} \leq c \cdot C_{g,m} \left(C^{\frac{m+1}{2}} d^{\frac{2p-m-1}{2}+\delta} \log^{\frac{m+1}{2}}(n)\sigma^{2m+2} + (L_0 + 1)\sigma^2 \sqrt{\frac{C \log(n)}{d}} \right),$$

where c is a numerical constant.

To meet the assumption $s < \frac{\min(1/2,\epsilon)}{\sigma^2}$, d needs to satisfy $d/\log(d) > \frac{\min(1/2,\epsilon)^2 c_1}{C(p+\delta)\sigma^4} := c_\epsilon$. The proof is complete.

C.1.4 Proof of Lemma A.16

Let $C_d(k)$ consists of all vectors $\lambda \in \mathbb{N}^d$ satisfying $\lambda \geq 0$ and $\sum_{i=1}^d \lambda_i = k$. Then the multinomial expansion of $(x^\top z)^k$ for any $x, y \in \mathbb{R}^d$ yields the expression

$$\frac{1}{k!} \left(\frac{x^\top z}{d} \right)^k = \frac{1}{d^k} \sum_{\lambda \in C_d(k)} \frac{1}{\lambda!} x^\lambda z^\lambda. \tag{194}$$

Now if x and y lie in the hypercube $\mathbb{H}^d$, equality $x^\lambda = x^{\lambda \bmod 2}$ holds. Thus we may now rewrite (194) as

$$\sum_{\lambda \in C_d(k)} \frac{1}{\lambda!} x^\lambda z^\lambda = \sum_{\lambda \in \{0,1\}^d} D_\lambda x^\lambda y^\lambda,$$

where we define the coefficients

$$D_\lambda = \sum_{\substack{\alpha \in C_d(k): \\ \lambda = \alpha \bmod 2}} \frac{1}{\alpha!}. \tag{195}$$

In particular, it is clear from (195) that if $k - |\lambda|$ is odd or strictly negative then equality $D_\lambda = 0$ holds. Moreover, for any $\lambda \in \{0,1\}^d$ with $|\lambda| = k$, the only summand on the right-side of (195) corresponds to $\alpha = \lambda$ and therefore equality $D_\lambda = \frac{1}{\lambda!} = 1$ holds. More generally, for any $\lambda \in \{0,1\}^d$ such that $k - |\lambda|$ is even and nonnegative, the number of terms in the sum in (195) is exactly the cardinality of the set $C_d(\frac{k-|\lambda|}{2})$, which is $\binom{\frac{k-|\lambda|}{2}+d-1}{d-1} = \Theta(d^{(k-|\lambda|)/2})$. Dividing D_λ by d^k completes the proof.

C.1.5 Proof of Lemma A.17

We first present a proof for Eq. (45). To this end, Theorem A.15 with $m = 2p + 1$ yields the estimate:

$$\|K - K_{2p+1} - (g(1) - g_{2p+1}(1))I_n\|_{\mathrm{op}} = O_{d,\mathbb{P}} \left(\frac{\log^{p+1}(n)}{d^{1-\delta}} \right). \tag{196}$$

Invoking Lemma A.16 for each k, we may write:

$$\frac{1}{k!} \left(\frac{XX^\top}{d} \right)^{\odot k} = d^{-k} \cdot \Phi_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top + \sum_{\substack{j:\, 0 \leq j < k, \\ k-j \text{ is even}}} \Phi_{\mathcal{S}_j} \widetilde{D}_{\mathcal{S}_j} \Phi_{\mathcal{S}_j}^\top,$$

where each $\widetilde{D}_{\mathcal{S}_j}$ is a diagonal matrix with all entries on the order of $\Theta(d^{-(j+k)/2}) = O(d^{-j-1})$. Since $k - j$ is even and positive we deduce $k \geq j + 2$. Therefore every entry of $\widetilde{D}_{\mathcal{S}_j}$ is bounded by $O(d^{-j-1})$. Thus, merging $\Phi_{\mathcal{S}_j} \widetilde{D}_{\mathcal{S}_j} \Phi_{\mathcal{S}_j}^\top$ with the lower-order bases, we can write

$$K_{2p+1} = \Phi_{\leq 2p+1} D_{\leq 2p+1} \Phi_{\leq 2p+1}^\top, \tag{197}$$

where D is a diagonal matrix satisfying

$$\|D_{\mathcal{S}_k} - g^{(k)}(0)d^{-k}I_{\mathcal{S}_k}\|_{\mathrm{op}} = O_d(d^{-k-1}), \tag{198}$$

for each $k = 0, \ldots, 2p+1$. We now claim that the terms in the product corresponding to $k = p+1, \ldots, 2p+1$ are close to a multiple of the identity. To see this, using the triangle inequality we estimate

$$\left\| \sum_{k=p+1}^{2p+1} \Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \sum_{k=p+1}^{2p+1} \frac{g^{(k)}(0)}{k!} I_n \right\|_{\mathrm{op}} \leq \sum_{k=p+1}^{2p+1} \left\| \Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \frac{g^{(k)}(0)}{k!} I_n \right\|_{\mathrm{op}}.$$

Using the triangle inequality with the expression (198), we subsequently deduce

$$\left\|\Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \frac{g^{(k)}(0)}{k!} \cdot I_n\right\|_{\text{op}} \leq \frac{g^{(k)}(0)}{k!} \left\|\frac{k!}{d^k}\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top - I_n\right\|_{\text{op}} + O_d(d^{-k-1}) \cdot \|\Phi_{\mathcal{S}_k}\|_{\text{op}}^2. \tag{199}$$

We now bound the two terms separately. To this end, note the estimate $|\mathcal{S}_k| = \frac{d^k}{k!} + O_d(d^{k-1})$. Therefore Lemma A.11 implies $\|\Phi_{\mathcal{S}_k}\|_{\text{op}}^2 = O_{d,\mathbb{P}}(d^k)$ and hence the second term on the right side of (199) is $O_{d,\mathbb{P}}(d^{-1})$. In order to control the first term, we invoke Theorem A.13 yielding

$$\left\|\frac{1}{|\mathcal{S}_k|}\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top - I_n\right\|_{\text{op}} \leq O_{d,\mathbb{P}}\left(d^{-\frac{k-p-\delta}{2}+\epsilon}\log(n)\right), \tag{200}$$

for any $\epsilon \in (0, \delta)$. By the triangle inequality, replacing $|\mathcal{S}_k|$ by $d^k/k!$ in (200) incurs an error of size

$$\left|\frac{k!}{d^k} - \frac{1}{|\mathcal{S}_k|}\right| \cdot \|\Phi_{\mathcal{S}_k}\|_{\text{op}}^2 = \frac{||\mathcal{S}_k|/d^k - \frac{1}{k!}|}{|S_k|/k!}\|\Phi_{\mathcal{S}_k}\|_{\text{op}}^2 = O_d(d^{-1}). \tag{201}$$

Therefore, we conclude that the first term on the right side of (199) is on the order of

$$O_{d,\mathbb{P}}\left(d^{-\frac{k-p-\delta}{2}+\epsilon}\log(n) + d^{-1}\right). \tag{202}$$

Combining with the bound for the second term on the right side of (199), we obtain

$$\left\|\Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \frac{g^{(k)}(0)}{k!} \cdot I_n\right\|_{\text{op}} = O_{d,\mathbb{P}}\left(d^{-\frac{k-p-\delta}{2}+\epsilon}\log(n) + d^{-1}\right). \tag{203}$$

Note that the term with $k = p+1$ will dominate the summation, thus, we deduce

$$\sum_{k=p+1}^{2p+1} \Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top = \underbrace{\sum_{k=p+1}^{2p+1} \frac{g^{(k)}(0)}{k!}I_n}_{=g_{2p+1}(1)-g_p(1)} + O_{d,\mathbb{P}}\left(d^{-\frac{1-\delta}{2}+\epsilon}\log(n)\right).$$

Combining this expression with (196) completes the proof for Eq. (45).

Now we proceed to prove Eq. (46). To this end, applying Theorem A.15 with $m = 2p+3$ yields the estimate:

$$\|K - K_{2p+3} - (g(1) - g_{2p+3}(1))I_n\|_{\text{op}} = O_{d,\mathbb{P}}\left(\frac{\log^{p+2}(n)}{d^{2-\delta}}\right) = O_{d,\mathbb{P}}(d^{-1}). \tag{204}$$

By the same reasoning that led to Eq. (197), we can write K_{2p+3} as

$$K_{2p+3} = \Phi_{\leq 2p+3} D_{\leq 2p+3} \Phi_{\leq 2p+3}^\top. \tag{205}$$

We can bound the term with $k \geq p+2$ by letting $k = p+2$ in Eq. (203), which yields

$$\left\|\Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \frac{g^{(k)}(0)}{k!}I_n\right\|_{\text{op}} = O_{d,\mathbb{P}}\left(d^{-\frac{2-\delta}{2}+\epsilon}\log(n)\right). \tag{206}$$

For $k = p+1$, we provide the following approximation:

$$\left\|\Phi_{\mathcal{S}_{p+1}} D_{\mathcal{S}_{p+1}} \Phi_{\mathcal{S}_{p+1}}^\top - \frac{g^{(p+1)}(0)}{(p+1)!} \cdot I_n - \frac{g^{(p+1)}(0)}{d^{p+1}}\text{off}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right)\right\|_{\text{op}}$$

$$\overset{(a)}{\leq} \left\|\frac{g^{(p+1)}(0)}{d^{p+1}}\text{diag}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right) - \frac{g^{(p+1)}(0)}{(p+1)!} \cdot I_n\right\|_{\text{op}} + O_d(d^{-p-2}) \cdot \left\|\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right\|_{\text{op}}$$

$$\overset{(b)}{=} \left\|\frac{g^{(p+1)}(0)}{(p+1)!|\mathcal{S}_{p+1}|}\text{diag}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right) - \frac{g^{(p+1)}(0)}{(p+1)!} \cdot I_n\right\|_{\text{op}} + O_d(d^{-1}) + O_d(d^{-p-2}) \cdot \left\|\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right\|_{\text{op}}$$

$$\overset{(c)}{=} O_{d,\mathbb{P}}(d^{-1}),$$

where (a) follows from Eq. (198), (b) follows from Eq. (201) and (c) follows from Lemma A.11 and the fact $\mathsf{diag}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right)/|\mathcal{S}_{p+1}| = I_n$. Combining with Eq. (206) we deduce

$$\left\| \sum_{k=p+1}^{2p+3} \Phi_{\mathcal{S}_k} D_{\mathcal{S}_k} \Phi_{\mathcal{S}_k}^\top - \underbrace{\sum_{k=p+1}^{2p+3} \frac{g^{(k)}(0)}{k!}}_{=g_{2p+3}(1)-g_p(1)} I_n - \frac{g^{(p+1)}(0)}{d^{p+1}}\mathsf{off}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right) \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{-\frac{2-\delta}{2}+\epsilon}\log(n)\right).$$

Combining this expression with Eq. (204) completes the proof.

C.1.6   Proof of Lemma A.18

Let us partition the second-order Hermite polynomials $\mathcal{S}_2$ into the pure and mixed parts as

$$\phi_{\mathcal{S}_2} = \left(\underbrace{\{h_{2e_i}\}_{i=1}^d}_{=:\phi_{\mathcal{S}_2}^s(x)}, \underbrace{\{h_{e_i+e_j}\}_{1\leq i<j\leq d}}_{=:\phi_{\mathcal{S}_2}^m(x)} \right).$$

For any $x, y \in \mathbb{R}^d$, expanding the square yields the expression

$$\begin{aligned}
\frac{1}{2}\left(x^\top y\right)^2 &= \frac{1}{2}\left(\sum_{i=1}^d x_i y_i\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^d x_i^2 y_i^2 + \sum_{1\leq i<j\leq d}(x_i x_j)\cdot(y_i y_j) \\
&= \frac{1}{2}\sum_{i=1}^d (\sqrt{2}h_2(x_i)+1)(\sqrt{2}h_2(y_i)+1) + \sum_{1\leq i<j\leq d}(h_1(x_i)h_1(x_j))\cdot(h_1(y_i)h_1(y_j)) \\
&= \frac{1}{2}\langle \sqrt{2}\phi_{\mathcal{S}_2}^s(x)+h_0(x)\cdot\mathbf{1}_d, \sqrt{2}\phi_{\mathcal{S}_2}^s(y)+h_0(y)\cdot\mathbf{1}_d\rangle + \langle \phi_{\mathcal{S}_2}^m(x), \phi_{\mathcal{S}_2}^m(y)\rangle \\
&= \langle \phi_{\mathcal{S}_2}(x), \phi_{\mathcal{S}_2}(y)\rangle + \frac{d}{2}\langle \phi_{\mathcal{S}_0}(x), \phi_{\mathcal{S}_0}(y)\rangle + \frac{1}{\sqrt{2}}\langle \mathbf{1}_d, \phi_{\mathcal{S}_2}^s(x)+\phi_{\mathcal{S}_2}^s(y)\rangle.
\end{aligned}$$

The equality (47) follows immediately.

C.1.7   Proof of Lemma A.19

For now, let $k \in \mathbb{N}$ be arbitrary. Observe that since $\left(x^\top y\right)^k$ is a degree k polynomial in x and in y, it can be expressed in terms of Hermite polynomials:

$$\left(x^\top y\right)^k = \sum_{\substack{\alpha,\beta\in\mathbb{N}^d \\ \|\alpha\|_1\leq k \\ \|\beta\|_1\leq k}} C[\alpha,\beta]h_\alpha(x)h_\beta(y), \tag{207}$$

for some constants $C[\alpha,\beta]$. In particular, We may regard $C[\cdot,\cdot]$ as a matrix satisfying

$$\frac{1}{k!}\left(\frac{XX^\top}{d}\right)^{\odot k} = \frac{1}{k!d^k}\Phi C\Phi^\top.$$

70

Our immediate goal is therefore to estimate the values $C[\alpha, \beta]$. We will do so by relating $C[\alpha, \beta]$ to the derivatives of the function $\phi(x, y) := (x^\top y)^k$. Namely, multiplying both sides of (207) by $h_\alpha(x)$ and $h_\beta(y)$ and taking expectations, we deduce

$$
\begin{aligned}
C[\alpha, \beta] &= \mathbb{E}_{x,y}[\phi(x, y) h_\alpha(x) h_\beta(y)] \\
&= \frac{1}{\sqrt{\alpha!}} \mathbb{E}_y[[\mathbb{E}_x \partial_x^{(\alpha)} \phi(x, y)] h_\beta(y)] \\
&= \frac{1}{\sqrt{\alpha! \beta!}} \mathbb{E}_{x,y} \partial_{x,y}^{(\alpha, \beta)} \phi(x, y),
\end{aligned}
$$

where the second and third equalities follow from (11). In order, to compute the derivatives of ϕ we apply the multinomial expansion to obtain

$$
\phi(x, y) = \sum_{\lambda \in C_d(k)} \frac{k!}{\lambda!} x^\lambda y^\lambda, \tag{208}
$$

where we let $C_d(k)$ be the set consisting of all vectors $\lambda \in \mathbb{N}^d$ satisfying $\lambda \geq 0$ and $\|\lambda\|_1 = k$. Now for any $\alpha, \beta \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k$ and $\|\beta\|_1 \leq k$, differentiating the equation (208) yields the expression

$$
\mathbb{E}[\partial_{x,y}^{(\alpha, \beta)} \phi(x, y)] = \sum_{\lambda \in C_d(k)} \frac{k!}{\lambda!} \cdot (\mathbb{E}\partial_x^{(\alpha)} x^\lambda) \cdot (\mathbb{E}\partial_y^{(\beta)} y^\lambda) = \sum_{\lambda \in \Lambda_{(\alpha, \beta)}} \frac{k! \lambda!}{(\lambda - \alpha)!(\lambda - \beta)!} \cdot \mathbb{E}[x^{\lambda - \alpha}] \mathbb{E}[y^{\lambda - \beta}],
$$

where we define the set

$$
\Lambda_{\alpha, \beta} := \{\lambda \in C_d(k) : \lambda - \alpha \text{ and } \lambda - \beta \text{ are nonnegative and even}\}.
$$

Summarizing, we have learned that $C[\alpha, \beta]$ can be computed as:

$$
C[\alpha, \beta] = \frac{1}{\sqrt{\alpha! \beta!}} \sum_{\lambda \in \Lambda_{\alpha, \beta}} \frac{k! \lambda!}{(\lambda - \alpha)!(\lambda - \beta)!} \cdot \mathbb{E}[x^{\lambda - \alpha}] \mathbb{E}[\boldsymbol{y}^{\lambda - \beta}]. \tag{209}
$$

Note that since each summand in (209) has constant order, equality $C[\alpha, \beta] = \Theta(|\Lambda_{\alpha, \beta}|)$ holds. Using the formula (209), one may easily compute the diagonal elements of C exactly for the two cases of interest $k = 3$ and $k = 4$. Moreover, for any principle submatrix $C[\mathcal{I}, \mathcal{I}]$, we may estimate the operator norm of the diagonal block as

$$
\|\Phi_\mathcal{I} \cdot \mathsf{diag}(C[\mathcal{I}, \mathcal{I}]) \cdot \Phi_\mathcal{I}^\top\|_{\mathrm{op}} \leq \|\mathsf{diag}(C[\mathcal{I}, \mathcal{I}])\|_{\mathrm{op}} \cdot \|\Phi_\mathcal{I}\|_{\mathrm{op}}^2.
$$

Using Lemma A.11 we may then further bound $\|\Phi_\mathcal{I}\|_{\mathrm{op}}^2$ as $O_{d,\mathbb{P}}(\sqrt{n})$ or as $O_{d,\mathbb{P}}(\sqrt{|\mathcal{I}|})$ depending on the ratio $|\mathcal{I}|/n$. We summarize in Tables 3 and Table 4 the diagonal entries of C, decomposed as a sum of diagonal blocks along with their operator norm.

$\mathcal{I}$	$\mathcal{J}$	$C[\mathcal{I}, \mathcal{J}]$	$d^{-k}\left\|\Phi_\mathcal{I} C \Phi_\mathcal{J}^\top\right\|_{\mathrm{op}}$
e_i	e_i	$3d$	$O_{d,\mathbb{P}}(d^\delta)$
e_i	e_i	6	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$3e_i$	$3e_i$	6	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$2e_i + e_j \ (i \neq j)$	$2e_i + e_j$	6	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$e_i + e_j + e_k \ (i, j, k)$ distinct	$e_i + e_j + e_k$	6	$O_{d,\mathbb{P}}(1)$

Table 3: $(p = 3)$ Diagonal elements of C arranged in blocks along with their operator norm.

$\mathcal{I}$	$\mathcal{J}$	$C[\mathcal{I},\mathcal{J}]$	$d^{-k}\left\|\Phi_\mathcal{I}C\Phi_\mathcal{J}^\top\right\|_{\mathrm{op}}$
0	0	$6d^2$	$O_{d,\mathbb{P}}(d^\delta)$
0	0	$1-6d$	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$e_i+e_j\ (i\neq j)$	e_i+e_j	$24d+48$	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$2e_i$	$2e_i$	$12d+60$	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$4e_i$	$4e_i$	24	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$3e_i+e_j\ (i\neq j)$	$3e_i+e_j$	24	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$2e_i+e_j+e_k$ $(i,j,k\text{ distinct})$	$2e_i+e_j+e_k$	24	$O_{d,\mathbb{P}}(d^{-1})$
$2e_i+2e_j\ (i\neq j)$	$2e_i+2e_j$	24	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$e_i+e_j+e_k+e_\ell$ $(i,j,k,\ell)\text{ distinct}$	$e_i+e_j+e_k+e_\ell$	24	$O_{d,\mathbb{P}}(1)$

Table 4: ($p=4$) Diagonal elements of C arranged in blocks along with their operator norm.

In particular, we deduce that the two displayed equations (48) and (49) hold with the matrix $XX^\top$ replaced by its diagonal $\mathsf{diag}(XX^\top)$. It remains therefore to argue that matrix $d^{-k}\Phi E\Phi^\top$ corresponding to off-diagonal terms $E := C - \mathsf{diag}(C)$ has a small operator norm. We do so for $k=3,4$ by identifying submatrices $E[\mathcal{I},\mathcal{J}]$, for finitely many sets $\mathcal{I}$ and $\mathcal{J}$, which cover all entries of E and so that the operator norm $\frac{1}{k!d^k}\|\Phi_\mathcal{I}E[\mathcal{I},\mathcal{J}]\Phi_\mathcal{J}^\top\|_{\mathrm{op}}$ tends to zero as d increases. In each case, we bound the norm as

$$\|\Phi_\mathcal{I}E[\mathcal{I},\mathcal{J}]\Phi_\mathcal{J}^\top\|_{\mathrm{op}} \leq \|E[\mathcal{I},\mathcal{J}]\|_{\mathrm{op}}\|\Phi_\mathcal{I}\|_{\mathrm{op}}\cdot\|\Phi_\mathcal{J}\|_{\mathrm{op}}.$$

The operator norms $\|\Phi_\mathcal{I}\|_{\mathrm{op}}$ and $\|\Phi_\mathcal{J}\|_{\mathrm{op}}$ can be bounded using Lemma A.11. In order to effectively bound $\|E[\mathcal{I},\mathcal{J}]\|_{\mathrm{op}}$, we choose $\mathcal{I}$ and $\mathcal{J}$ in such a way that $E[\mathcal{I},\mathcal{J}]$, up to column/row permutations and taking the transpose, is a matrix of two possible types:

$$\textbf{Type I}:\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_m \end{pmatrix}, \quad \textbf{Type II}:\begin{pmatrix} \mathbf{c_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{c_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{c_m} \end{pmatrix},$$

where $c_1,\ldots,c_m$ are real numbers and $\mathbf{c_1},\mathbf{c_1},\ldots,\mathbf{c_m}$ are row vectors. The operator norms of the two types of matrices are straightforward to compute. The results are summarized in Tables 5 and 6. Since the decay rate of $d^{-k}\|\Phi_\mathcal{I}E\Phi_\mathcal{J}^\top\|_{\mathrm{op}}$ is at least on the order of $O_{d,\mathbb{P}}(d^{\delta-1/2})$, the result follows.

$\mathcal{I}$	$\mathcal{J}$	$\Lambda_{\mathcal{I},\mathcal{J}}$	$E[\mathcal{I},\mathcal{J}]$	Type of E	$d^{-k}\left\|\Phi_\mathcal{I}E\Phi_\mathcal{J}^\top\right\|_{\mathrm{op}}$
e_i	$e_i+2e_j\ (j\neq i)$	e_i+2e_j	$\Theta(1)$	II	$O_{d,\mathbb{P}}(d^{-1/2+\delta})$
e_i	$3e_i$	$3e_i$	$\Theta(1)$	I	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$3e_i$	e_i	$3e_i$	$\Theta(1)$	I	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$2e_i+e_j\ (i\neq j)$	e_j	$2e_i+e_j$	$\Theta(1)$	II	$O_{d,\mathbb{P}}(d^{-1/2+\delta})$

Table 5: All off-diagonal terms in Eq. (207) when $p=3$.

$\mathcal{I}$	$\mathcal{J}$	$\Lambda_{\mathcal{I},\mathcal{J}}$	$E[\mathcal{I},\mathcal{J}]$	Type of E	$d^{-k}\left\|\Phi_{\mathcal{I}}E\Phi_{\mathcal{J}}^\top\right\|_{\mathrm{op}}$
0	$2e_i$	$\{2e_i+2e_j\}_j$	$\Theta_d(d)$	II	$O_{d,\mathbb{P}}(d^{-1/2+\delta})$
0	$4e_i$	$4e_i$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-3/2+\delta})$
$e_i+e_j\ (i\neq j)$	$e_i+e_j+2e_k$ (distinct)	$e_i+e_j+2e_k$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-1+\delta/2})$
$e_i+e_j\ (i\neq j)$	$3e_i+e_j$	$3e_i+e_j$	$\Theta_d(1)$	I	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$2e_i$	0	$\{2e_i+2e_j\}_j$	$\Theta_d(d)$	II	$O_{d,\mathbb{P}}(d^{-1/2+\delta})$
$2e_i$	$2e_i+2e_j\ (i\neq j)$	$2e_i+2e_j$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-3/2+\delta})$
$2e_i$	$4e_i$	$4e_i$	$\Theta_d(1)$	I	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$4e_i$	0	$4e_i$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-3/2+\delta})$
$4e_i$	$2e_i$	$4e_i$	$\Theta_d(1)$	I	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$3e_i+e_j\ \ (i\neq j)$	e_i+e_j	$3e_i+e_j$	$\Theta_d(1)$	I	$O_{d,\mathbb{P}}(d^{-2+\delta})$
$2e_i+e_j+e_k$ $(i,j,k$ distinct$)$	e_j+e_k	$2e_i+e_j+e_k$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-1+\delta/2})$
$2e_i+2e_j\ \ (i\neq j)$	0	$2e_i+2e_j$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-1+\delta})$
$2e_i+2e_j\ \ (i\neq j)$	$2e_i$	$2e_i+2e_j$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-3/2+\delta})$
$2e_i+2e_j\ \ (i\neq j)$	$2e_j$	$2e_i+2e_j$	$\Theta_d(1)$	II	$O_{d,\mathbb{P}}(d^{-3/2+\delta})$

Table 6: All off-diagonal terms in Eq. (207) when $p=4$.

C.1.8 Proof of Lemma A.20

Applying Theorem A.14 in the regime $n=d^{2+\delta}$ with $m=4$, we deduce

$$\|K-K_4-(g(1)-g_4(1))I_n\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2}\log^3(n)\right). \tag{210}$$

Recall that K_4 can be equivalently written as

$$K_4 = \sum_{k=0}^{4}\frac{g^{(k)}(0)}{k!}\left(\frac{XX^\top}{d}\right)^{\odot k}.$$

Combining Lemma A.18 and Lemma A.19 directly yields the expression:

$$K_4 = \Phi_{\leq 2}D\Phi_{\leq 2}^\top - \frac{g^{(3)}(0)}{d^3}\Phi_{\mathcal{S}_3}\Phi_{\mathcal{S}_3}^\top - \frac{g^{(4)}(0)}{d^4}\Phi_{\mathcal{S}_4}\Phi_{\mathcal{S}_4}^\top.$$

It remains to argue the second two terms are close to the identity. To this end, applying Lemma A.13 with $k=3,4$ yields the expression

$$\left\|\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top - |\mathcal{S}_k|I_n\right\|_{\mathrm{op}} = |\mathcal{S}_k|\cdot O_{d,\mathbb{P}}(d^{(\delta'-1)/2}\log(d)),$$

for any $\delta'\in(\delta,1)$. Therefore, we deduce

$$\frac{g^{(k)}(0)}{d^k}\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top = \frac{g^{(k)}(0)|\mathcal{S}_k|}{d^k}I_n + O_{d,\mathbb{P}}(d^{(\delta'-1)/2}\log(d))$$

$$= \frac{g^{(k)}(0)}{k!}I_n + O_{d,\mathbb{P}}(d^{(\delta'-1)/2}\log(d)),$$

where the last inequality uses the elementary observation $d^{-k}|\mathcal{S}_k| = \frac{1}{k!}+O(d^{-1})$. Therefore we deduce

$$\frac{g^{(3)}(0)}{d^3}\Phi_{\mathcal{S}_3}\Phi_{\mathcal{S}_3}^\top + \frac{g^{(4)}(0)}{d^4}\Phi_{\mathcal{S}_4}\Phi_{\mathcal{S}_4}^\top = \underbrace{\sum_{k=3,4}\frac{g^{(k)}(0)}{k!}}_{=g_4(1)-g_2(1)} I_n + O_{d,\mathbb{P}}(d^{(\delta'-1)/2}\log(d)).$$

Combining this expression with (210) and the triangle inequality completes the proof.

C.2 Proof of Propositions for Theorem 4.5

C.2.1 Proof of Proposition B.22

Invoking the definition of D' (Eq. (51)) and using the triangle inequality we have

$$\left\|D'_{\mathcal{S}_2}\right\|_{\mathrm{op}} \leq \left\|D'_{13}\right\|_{\mathrm{op}} + \left\|D'_{\mathcal{S}_2,\mathcal{S}_2}\right\|_{\mathrm{op}} = \frac{g^{(3)}(0)}{\sqrt{2}d^{3/2}} + \frac{g^{(3)}(0)}{d^2} = O_d(d^{-3/2}).$$

Taking into account Eq. (57) and submultiplicativity of the spectral norm, we deduce

$$\left\|\frac{1}{d}\beta^\top X_r \Phi_{\mathcal{S}_2} D'_{\mathcal{S}_2}\right\|_2 \leq \frac{1}{d}\left\|\beta\right\|_2 \left\|X_r\right\|_{\mathrm{op}} \left\|\Phi_{\mathcal{S}_2}\right\|_{\mathrm{op}} \left\|D'_{\mathcal{S}_2}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta-1/2}\log(d)) \cdot \sqrt{\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2},$$

where the last equality follows from the estimates $\|\beta\|_2 \leq \sqrt{n\log(n)(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)}$ (Eq. (58)), $\sup_r \|X_r\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{\log(d)})$ (Eqn. (59)), and $\|\Phi_{\mathcal{S}_2}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11). Taking the square of both sides and absorbing logarithmic factors into ϵ completes the proof.

C.2.2 Proof of Proposition B.23

Invoking the definition of D' (Eq. (51)), we have

$$\left\|D'_{\leq 1,\mathcal{S}_2}\right\|_2 = \|D'_{13}\|_2 = \frac{|g^{(3)}(0)|}{\sqrt{2}}d^{-3/2}.$$

Taking into account Eq. (57) and submultiplicativity of the spectral norm, we deduce

$$\left\|\frac{1}{d}\beta^\top X_r \Phi_{\leq 1} D'_{\leq 1,\mathcal{S}_2}\right\|_2 \leq \frac{1}{d}\left\|\beta\right\|_2 \left\|X_r\right\|_{\mathrm{op}} \left\|\Phi_{\leq 1}\right\|_{\mathrm{op}} \left\|D'_{\leq 1,\mathcal{S}_2}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta-1/2}\log(d)),$$

where the last equality follows from the estimates $\|\beta\|_2 \leq \sqrt{n\log(n)(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)}$ (Eq. (58)), $\sup_r \|X_r\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{\log(d)})$ (Eqn. (59)), and $\|\Phi_{\leq 1}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11). Taking the square of both sides and absorbing logarithmic factors into ϵ completes the proof.

C.2.3 Proof of Proposition B.24

In all arguments in this section, the asymptotic terms $o_{d,\mathbb{P}}$ and $O_{d,\mathbb{P}}$ will always be uniform overall the coordinates $r = 1,\ldots,d$. Recall that all Hermite polynomials of order at most two are products of polynomials of the form

$$h_0(x_i) = 1, \qquad h_1(x_i) = x_i, \qquad h_2(x_i) = \frac{x_i^2 - 1}{\sqrt{2}},$$

evaluated along the coordinates x_i. Then taking into account the definition $f^*_{\leq 2}(x) = \sum_{\lambda \in \mathbb{N}^d:\ \|\lambda\|_1 \leq 2} b_\lambda h_\lambda(x)$, we may write

$$\begin{aligned}
\mathbb{E}[\partial_{x_r} f^*_{\leq 2}(x)]^2 &= \mathbb{E}\left[\sum_{\lambda \in \mathcal{A}(r)} b_\lambda \partial_{x_r} h_\lambda(x)\right]^2 \\
&= \sum_{\lambda,\hat{\lambda} \in \mathcal{A}(r)} b_\lambda b_{\hat{\lambda}} \mathbb{E}[\partial_{x_r} h_\lambda(x) \partial_{x_r} h_{\hat{\lambda}}(x)] \\
&= b_{e_r}^2 + \sum_{i \in [d]\setminus\{r\}} b_{e_r+e_i}^2 + \sqrt{2} \cdot b_{2e_r}^2,
\end{aligned}$$

where the last equality follows from computing the expectation $\mathbb{E}[\partial_r h_\lambda(x)\partial_r h_{\hat\lambda}(x)]$ in all cases. In particular, we may write

$$\mathbb{E}[\partial_r f^*_{\leq 2}(x)]^2 = \|b^\top_{\leq 2}E_r\|_2^2,$$

where $E_r \in \mathbb{R}^{L(2)\times(d+1)}$ is the matrix defined in the obvious way.

Invoking the Hermite expansion of f^* in (53) we may write

$$y = \Phi b + \varepsilon = \underbrace{\Phi_{\leq 2}b_{\leq 2}}_{=:f^*_{\leq 2}(X)} + \underbrace{\Phi_{>2}b_{>2}}_{=:f^*_{>2}(x)} + \varepsilon,$$

and similarly

$$\beta = K_\lambda^{-1}y = \underbrace{K_\lambda^{-1}f^*_{\leq 2}(X)}_{=:\beta_{\leq 2}} + \underbrace{K_\lambda^{-1}f^*_{>2}(X)}_{=:\beta_{>2}} + \underbrace{K_\lambda^{-1}\varepsilon}_{\beta_\varepsilon},$$

where $\varepsilon := (\varepsilon_1,\ldots,\varepsilon_n) \in \mathbb{R}^n$.

With this decomposition of β, we accordingly bound the norm by the triangle inequality:

$$\left\|\frac{1}{d}\beta^\top X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} - b^\top_{\leq 2}E_r\right\|_2$$
$$\leq \left\|\frac{1}{d}\beta^\top_{\leq 2}X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} - b^\top_{\leq 2}E_r\right\|_2 + \left\|\frac{1}{d}\beta^\top_{>2}X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1}\right\|_2 + \left\|\frac{1}{d}\varepsilon^\top K_\lambda^{-1}X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1}\right\|. \quad (211)$$

We will show all the terms on the right-hand side of the above equation tend to zero, from which the theorem will follow quickly. We start with the first term in the following proposition. The proof of the proposition appears in Appendix C.2.4.

Proposition C.39. *For any $\epsilon > 0$, the estimate holds:*

$$\left\|\frac{1}{d}\beta^\top_{\leq 2}X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} - b^\top_{\leq 2}E_r\right\|_2 = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right) \cdot \left\|\partial_r f^*_{\leq 2}\right\|_{L_2}.$$

The next proposition shows that the second term on the right side of (211) is small. The proof of the proposition appears in Appendix C.2.5.

Proposition C.40. *For any $\epsilon > 0$, the estimate holds:*

$$\left\|\frac{1}{d}\beta^\top_{>2}X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1}\right\|_2 = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right) \cdot \|f^*_{>2}\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right) \cdot \|f^*_{>2}\|_{L_{2+\eta}}. \quad (212)$$

Now we show that the third term on the right side of (211) is small. The proof of the proposition appears in Appendix C.2.6.

Proposition C.41. *For any $\epsilon > 0$, the following estimate holds*

$$\left\|\frac{1}{d}\beta_\varepsilon X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1}\right\|_2 = O_{d,\mathbb{P}}(d^{-(\delta+1)/2+\epsilon}) \cdot \sigma_\varepsilon.$$

Consequently, with Proposition C.39, C.40 and C.41 we finish the bound for Eq. (211):

$$\left\|\frac{1}{d}\beta^\top X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} - b^\top_{\leq 2}E_r\right\|_2 = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right) \cdot \|f^*\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right) \cdot \|f^*_{>2}\|_{L_{2+\eta}}$$
$$+ O_{d,\mathbb{P}}(d^{-(\delta+1)/2+\epsilon}) \cdot \sigma_\varepsilon. \quad (213)$$

Now recall the elementary estimate

$$|\|v\|_2^2 - \|w\|_2^2| = |\langle v - w, v + w\rangle| \leq \|v - w\|_2(\|v\| + \|w\|_2),$$

75

that is valid for any vectors v, w. We therefore compute

$$\left| \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} \right\|_2^2 - \left\| b_{\leq 2}^\top E_r \right\|_2^2 \right| \leq \left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} - b_{\leq 2}^\top E_r \right\|_2 \tag{214}$$

$$\cdot \left(\left\| \frac{1}{d} \beta^\top X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} \right\|_2 + \left\| b_{\leq 2}^\top E_r \right\|_2 \right). \tag{215}$$

The term (214) is bounded by $O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right) \cdot \|f^*\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right) \cdot \|f^*_{>2}\|_{L_{2+\eta}} + O_{d,\mathbb{P}}(d^{-(\delta+1)/2 + \epsilon}) \cdot \sigma_\varepsilon$ due to (213) while the term (215) is clearly $O_{d,\mathbb{P}}(\|\partial_r f^*_{\leq 2}\|_{L_2})$ due to Proposition C.39. Then plugging these bounds into the above inequality yields an upper bound

$$O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right) \cdot \|f^*\|_{L_2} \|\partial_r f^*_{\leq 2}\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right) \cdot \|f^*_{>2}\|_{L_{2+\eta}} \|\partial_r f^*_{\leq 2}\|_{L_2}$$
$$+ O_{d,\mathbb{P}}(d^{-(\delta+1)/2 + \epsilon}) \cdot \sigma_\varepsilon \|\partial_r f^*_{\leq 2}\|_{L_2}.$$

Using the fact $\|\partial_r f^*_{\leq 2}\|_{L_2}^2 = O_d(1) \cdot \|f^*_{\leq 2}\|_{L_2}^2 \leq O_d(1) \cdot \|f^*\|_{L_2}^2$ and applying AM-GM inequality to the above bound yields

$$O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right) \cdot \|f^*\|_{L_2}^2 + O_{d,\mathbb{P}}\left(d^{-\delta/2}\right) \cdot \|f^*_{>2}\|_{L_{2+\eta}}^2 + O_{d,\mathbb{P}}(d^{-(\delta+1)/2 + \epsilon}) \cdot \sigma_\varepsilon^2.$$

Finally, combining with the fact $\|b_{\leq 2}^\top E_r\|_2^2 = \mathbb{E}[\partial_r f^*_{\leq 2}(x)]^2$ completes the proof.

C.2.4 Proof of Proposition C.39

Without loss of generality we focus on the first coordinate $r = 1$. We first establish some notation. We define the column vector e_k to be 1 in k'th coordinate and zero everywhere else. The length of e_k will be clear from context. Next, let us specify an ordering on the Hermite basis of degree at most two as

$$\phi_{\leq 2}(x) = [\underbrace{1}_{\phi_{S_0}},$$
$$\underbrace{h_1(x_1), \ldots, h_1(x_d)}_{\phi_{S_1}},$$
$$\underbrace{h_2(x_1), \ldots, h_2(x_d), h_1(x_1)h_1(x_2), \ldots, h_1(x_1)h_1(x_d), \ldots, h_1(x_{d-1})h_1(x_d)}_{\phi_{S_2}}].$$

In other words, the terms in ϕ_{S_2} are ordered by concatenating the diagonal of the matrix $[h_1(x_i)h_1(x_j)]_{i,j}$ with its rows above the diagonal collapsed into a single vector. With this notation, the matrix E_1 can be simply written as

$$E_1 = \left(e_2, \sqrt{2} e_{L(1)+1}, e_{L(1)+d+1}, e_{L(1)+d+2} \ldots, e_{L(1)+2d-1} \right)$$

We now write $X_1 \Phi_{\leq 1} D'_{\leq 1, \leq 1}$ in a Hermite basis. More precisely, a quick computation shows the equality

$$X_1 \Phi_{\leq 1} = \Phi_{\leq 2} G,$$

where $G := [e_2, \sqrt{2}(e_{L(1)+1} + e_1), e_{L(1)+d+1}, e_{L(1)+d+2} \ldots, e_{L(1)+2d-1}]$. Thus multiplying both sides on the right by $D'_{\leq 1, \leq 1}$ yields the equality

$$X_1 \Phi_{\leq 1} D'_{\leq 1, \leq 1} = \Phi_{\leq 2} R, \qquad \text{where} \qquad R = G D'_{\leq 1, \leq 1}. \tag{216}$$

Thus it remains to estimate the error

$$\left\| \frac{1}{d} \beta_{\leq 2}^\top \Phi_{\leq 2} R - b_{\leq 2}^\top E_1 \right\|_2 = \left\| b_{\leq 2}^\top \left(\frac{1}{d} \Phi_{\leq 2}^\top K_\lambda^{-1} \Phi_{\leq 2} R - E_1 \right) \right\|_2$$

$$\leq \sqrt{\mathbb{E}|f^*_{\leq 2}|^2} \cdot \left\| \frac{1}{d} \Phi_{\leq 2}^\top K_\lambda^{-1} \Phi_{\leq 2} R - E_1 \right\|_{op}. \tag{217}$$

We now argue in the following proposition that $d^{-1}D^{-1}R$ is close to E_1. The proof appears in Appendix C.2.7.

Proposition C.42. *The estimate holds:*

$$\left\|\frac{1}{d}D^{-1}R - E_1\right\|_{\mathrm{op}} = O_d(d^{-1}). \tag{218}$$

Thus the right side of (217) can be further bounded by

$$\sqrt{\mathbb{E}|f_{\leq 2}^*|^2} \cdot \left\|\frac{1}{d}\Phi_{\leq 2}^\top K_\lambda^{-1}\Phi_{\leq 2}R - \frac{1}{d}D^{-1}R\right\|_{\mathrm{op}} + O_d(d^{-1}). \tag{219}$$

Next, we further upper bound the right side of (219) with the estimate (220) of the following proposition. The proof appears in Section C.2.8.

Proposition C.43. *The two estimates hold:*

$$\left\|\frac{1}{d}\Phi_{\leq 2}^\top K_\lambda^{-1}\Phi_{\leq 2}R - \frac{1}{d}D^{-1}R\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right), \tag{220}$$

$$\left\|\frac{\sqrt{n}}{d}K_\lambda^{-1}\Phi_{\leq 2}R - \frac{1}{d\sqrt{n}}\Phi_{\leq 2}D^{-1}R\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right). \tag{221}$$

Combining the estimates (220) and (219) completes the proof. We will invoke the second estimate (221) as part of the proof of Proposition C.40.

C.2.5 Proof of Proposition C.40

Proof. Following Eq. (216), we write $X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} = \Phi_{\leq 2}R_r$ for a matrix R_r. Equation (221) from Proposition C.43 shows that for any $\epsilon > 0$, the following holds:

$$\left\|\sqrt{n}K_\lambda^{-1}\Phi_{\leq 2}\left(\frac{1}{d}R_r\right) - \frac{1}{\sqrt{n}}\Phi_{\leq 2}\left(\frac{1}{d}D^{-1}R_r\right)\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right).$$

By Lemma A.11, we have $\|\Phi_{\leq 2}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$, and by Proposition C.42, we have $\left\|\frac{1}{d}D^{-1}R_r - E_r\right\|_{\mathrm{op}} = O_d(d^{-1})$. Therefore, by triangle inequality and the submultiplicativity of the spectral norm, we have

$$\left\|\sqrt{n}K_\lambda^{-1}\Phi_{\leq 2}\left(\frac{1}{d}R_r\right) - \frac{1}{\sqrt{n}}\Phi_{\leq 2}E_r\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right). \tag{222}$$

Next, observe $\|\Phi_{>2}b_{>2}\|_2 = O_{d,\mathbb{P}}(\sqrt{n\log(n)\cdot\mathbb{E}|f_{>2}^*|^2})$ by Proposition B.21. Consequently, we compute

$$\left\|\frac{1}{d}\beta_{>2}^\top X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} - \frac{1}{n}b_{>2}^\top\Phi_{>2}^\top\Phi_{\leq 2}E_r\right\|_2$$

$$= \left\|b_{>2}^\top\Phi_{>2}^\top\left(\frac{1}{d}K_\lambda^{-1}\Phi_{\leq 2}R_r - \frac{1}{n}\Phi_{\leq 2}E_r\right)\right\|_2$$

$$\leq \frac{1}{\sqrt{n}}\|\Phi_{>2}b_{>2}\|_2\left\|\frac{\sqrt{n}}{d}K_\lambda^{-1}\Phi_{\leq 2}R_r - \frac{1}{\sqrt{n}}\Phi_{\leq 2}E_r\right\|_{\mathrm{op}}$$

$$= O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon}\log^{1/2}(d) + d^{-\delta/2+\epsilon}\log^{1/2}(d)\right)\cdot\sqrt{\mathbb{E}|f_{>2}^*|^2}, \tag{223}$$

where the first line follows from the defining property $X_r\Phi_{\leq 1}D'_{\leq 1,\leq 1} = \Phi_{\leq 2}R_r$, the second line follows from the triangle inequality, and the third line follows from (222).

77

We next show that $\max_{r \in [d]} \left\| \frac{1}{n} b_{>2}^\top \Phi_{>2}^\top \Phi_{\leq 2} E_r \right\|_2^2$ is small, from which the conclusion of the lemma follows immediately. To this end, define the vector

$$F_{>2} = \begin{bmatrix} f_{>2}^*(x^{(1)}) & \cdots & f_{>2}^*(x^{(n)}) \end{bmatrix} \in \mathbb{R}^{1 \times n},$$

and note the equality $F_{>2} = b_{>2}^\top \Phi_{>2}^\top$. Taking the expectation over $x^{(1)}, \cdots, x^{(n)}$ we have:

$$\mathbb{E}\left[\max_{r \in [d]} \frac{1}{n^2} \|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right] \leq \frac{1}{n^2} \mathbb{E}\left[\sum_{r=1}^d \|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right] \leq \frac{d}{n^2} \mathbb{E}\left[\|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right], \qquad (224)$$

where the first inequality from bounding the maximum by a sum. We now bound $\mathbb{E}\left[\|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right]$ as follows. Define the set

$$\mathcal{A}_r := \{ \lambda \in \mathbb{N}^d : \sum_i \lambda_i \leq 2 \text{ and } \lambda_r \geq 1 \},$$

of multi-indices of order at most two that depend on coordinate r. We then compute

$$\mathbb{E}\left[\|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right] = \mathbb{E}\langle F_{>2}^\top F_{>2}, \Phi_{\leq 2} E_r E_r^\top \Phi_{\leq 2}^\top \rangle$$

$$= \mathbb{E}\left\langle F_{>2}^\top F_{>2}, \sum_{\lambda \in \mathcal{A}_r} (1 + \mathbb{1}_{\lambda = 2e_r}) \Phi_\lambda \Phi_\lambda^\top \right\rangle \qquad (225)$$

$$= \sum_{\lambda \in \mathcal{A}_r} (1 + \mathbb{1}_{\lambda = 2e_r}) \mathbb{E}(F_{>2} \Phi_\lambda)^2, \qquad (226)$$

where (225) follows from the definition of E_r. Continuing, for any $\lambda \in \mathcal{A}_r$, expanding the square we compute

$$\mathbb{E}(F_{>2} \Phi_\lambda)^2 = \sum_{i,j=1}^n \mathbb{E}[f_{>2}^*(x^{(i)}) \phi_\lambda(x^{(i)}) f_{>2}^*(x^{(j)}) \phi_\lambda(x^{(j)})]$$

$$= \sum_{i \neq j} \mathbb{E}[f_{>2}^*(x^{(i)}) \phi_\lambda(x^{(i)})] \cdot \mathbb{E}[f_{>2}^*(x^{(j)}) \phi_\lambda(x^{(j)})] + \sum_{i=1}^n \mathbb{E}[f_{>2}^*(x^{(i)})^2 \phi_\lambda(x^{(i)})^2]$$

$$= n \cdot \mathbb{E}[f_{>2}^*(x)^2 \phi_\lambda(x)^2]$$

where the last equality follows from the fact that ϕ_λ and $f_{>2}^*$ are orthogonal for any $\lambda \in \mathbb{N}^d$ with $\lambda \leq 2$. Thus combining with (226) and (224) we conclude

$$\mathbb{E}\left[\max_{r \in [d]} \frac{1}{n^2} \|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right] \leq \frac{2d}{n} \cdot |\mathcal{A}_r| \cdot \max_{\lambda \in \mathcal{A}_r} \mathbb{E}[f_{>2}^*(x)^2 \phi_\lambda(x)^2]$$

$$= O_d(d^{-\delta}) \cdot \max_{\lambda \in \mathcal{A}_r} \mathbb{E}[f_{>2}^*(x)^2 \phi_\lambda(x)^2],$$

where the last equality uses the fact that $\mathcal{A}_r$ has cardinality of order d.

It remains to show that $\mathbb{E}[f_{>2}^*(x)^2 \phi_\lambda(x)^2]$ has constant order $O_d(1)$ for any $\lambda \in \mathcal{A}_r$. To see this, for any constant $\eta > 0$, applying Holder's inequality yields

$$\mathbb{E}_x \left[f_{>2}^*(x)^2 \phi_\lambda(x)^2 \right] \leq \left\| f_{>2}^*(x)^2 \right\|_{L_{1+\frac{\eta}{2}}} \left\| \phi_\lambda^2 \right\|_{L_{1+\frac{2}{\eta}}} \qquad (227)$$

$$= \|f_{>2}^*(x)\|_{L_{2+\eta}}^2 \cdot \|\phi_\lambda\|_{L_{2+\frac{4}{\eta}}}^2. \qquad (228)$$

By hypercontractivity of Guassian random variables (Eq. (12)) we have

$$\|\phi_\lambda\|_{L_{2+\frac{4}{\eta}}} \leq \left(1 + \frac{4}{\eta} \right) \sqrt{\mathbb{E}_x \left[\phi_\lambda(x)^2 \right]} = 1 + \frac{4}{\eta}.$$

Then we conclude $\mathbb{E}_x \left[f_{>2}^*(x)^2 \phi_\lambda(x)^2 \right] = O_d(\|f_{>2}^*(x)\|_{L_{2+\eta}}^2)$, which implies $\mathbb{E}\left[\max_{r \in [d]} \frac{1}{n^2} \|F_{>2} \Phi_{\leq 2} E_r\|_2^2 \right] = O_d(d^{-\delta} \cdot \|f_{>2}^*(x)\|_{L_{2+\eta}}^2)$. Plugging the bound into Eq. (223) and absorbing logarithmic factors into ϵ we complete the proof.

C.2.6 Proof of Proposition C.41

Following Eq. (216), we write $X_r \Phi_{\leq 1} D'_{\leq 1, \leq 1} = \Phi_{\leq 2} R_r$ for a matrix R_r. Since $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, we have

$$
\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \frac{1}{d} \boldsymbol{\varepsilon}^\top K_\lambda^{-1} \Phi_{\leq 2} R_r \right\|_2^2 \right] = \frac{\sigma_\varepsilon^2}{d^2} \operatorname{Tr} \left(K_\lambda^{-1} \Phi_{\leq 2} R_r R_r^\top \Phi_{\leq 2}^\top K_\lambda^{-1} \right)
$$

$$
= \frac{\sigma_\varepsilon^2}{d^2} \operatorname{Tr} \left(K_\lambda^{-1} \Phi_{\leq 2} D D^{-1} R_r R_r^\top D^{-1} D \Phi_{\leq 2}^\top K_\lambda^{-1} \right)
$$

$$
\leq \sigma_\varepsilon^2 \cdot \left\| K_\lambda^{-1} \Phi_{\leq 2} D \right\|_{\text{op}}^2 \cdot \operatorname{Tr} \left(\frac{1}{d^2} D^{-1} R_r R_r^\top D^{-1} \right), \tag{229}
$$

where the last inequality follows from the fact that $|\operatorname{Tr}(AB)| \leq \operatorname{Tr}(A) \|B\|_{\text{op}}$ holds for Hermitian positive semidefinite matrix A and arbitrary matrix B.

For the term $\left\| K_\lambda^{-1} \Phi_{\leq 2} D \right\|_{\text{op}}^2$, we plug $U = D$ into Eq. (234) and yield

$$
K_\lambda^{-1} \Phi_{\leq 2} D = \frac{1}{n} E^{-1} \Phi_{\leq 2} V.
$$

With $\left\| E^{-1} \right\|_{\text{op}}, \|V\|_{\text{op}} = O_{d, \mathbb{P}}(1)$ established in Eq. (239), and $\|\Phi_{\leq 2}\|_{\text{op}} = O_{d, \mathbb{P}}(\sqrt{n})$ (Lemma A.11), we conclude

$$
\left\| K_\lambda^{-1} \Phi_{\leq 2} D \right\|_{\text{op}}^2 = \left\| \frac{1}{n} E^{-1} \Phi_{\leq 2} V \right\|_{\text{op}}^2 = O_{d, \mathbb{P}}(n^{-1}). \tag{230}
$$

We proceed to bound $\operatorname{Tr} \left(\frac{1}{d^2} D^{-1} R_r R_r^\top D^{-1} \right)$ in Eq. (229). We denote $\Delta := \frac{1}{d} D^{-1} R_r - E_r$ and Proposition C.42 shows $\|\Delta\|_{\text{op}} = O_d(d^{-1})$. Then using Weyl's inequality yields

$$
\operatorname{Tr} \left((E_r + \Delta)(E_r + \Delta)^\top \right) = \sum_i \sigma_i (E_r + \Delta)^2
$$

$$
\leq \sum_i (\sigma_i(E_r) + \|\Delta\|_{\text{op}})^2
$$

$$
\leq \|E_r\|_F^2 + 2 \|\Delta\|_{\text{op}} \cdot \left(\sum_i \sigma_i(E_r) \right) + \|\Delta\|_{\text{op}}^2 \cdot O_d(d),
$$

where σ_i denotes the singular values and we use the fact that the rank of E_r is of order $O_d(d)$ in the last inequality.

Following the definition of E_r, we have $\|E_r\|_F^2 = O_d(d)$ and $\left(\sum_i \sigma_i(E_r) \right) = O_d(d)$ which concludes the bound

$$
\operatorname{Tr} \left(\frac{1}{d^2} D^{-1} R_r R_r^\top D^{-1} \right) = O_d(d).
$$

Together with Eq. (230), we conclude the bound for Eq. (229). Then applying Markov's inequality, we obtain with probability at least $1 - 1/\log(d)$, the following holds:

$$
\left\| \frac{1}{d} \boldsymbol{\varepsilon}^\top K_\lambda^{-1} \Phi_{\leq 2} R_r \right\|_2^2 = O_d(d^{-1-\delta}) \log(d).
$$

Using an arbitrarily small constant $\epsilon > 0$ to absorb the logarithmic factor completes the proof.

C.2.7 Proof of Proposition C.42

Recall that we have $R = G D_{\leq 1, \leq 1}$ with

$$
G = [e_2, \sqrt{2}(e_{L(1)+1} + e_1), \widetilde{I}],
$$

and $\widetilde{I} := [e_{L(1)+d+1}, e_{L(1)+d+2} \ldots, e_{L(1)+2d-1}]$ is a zero-padded identity matrix. Let us organize $D'_{\leq 1, \leq 1}$ and D in a block form as follows:

$$D'_{\leq 1, \leq 1} = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 I_{|\mathcal{S}_1|} \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} a & v^\top \\ v & Q \end{bmatrix} \quad \text{with} \quad Q := \begin{bmatrix} q_2 I_{|\mathcal{S}_1|} & 0 \\ 0 & q_3 I_{|\mathcal{S}_2|} \end{bmatrix}.$$

Here $a, q_1, q_2, \gamma_1, \gamma_2 \in \mathbb{R}$ are real values and $v \in \mathbb{R}^{L(2)-1}$ is a column vector. The standard inversion formula for arrow-head matrices shows that D^{-1} is a rank one perturbation of a diagonal matrix

$$D^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & Q^{-1} \end{bmatrix} + \rho w w^\top, \tag{231}$$

where we define

$$\rho := \frac{1}{a - v^\top Q^{-1} v} \quad \text{and} \quad w := \begin{bmatrix} -1 \\ Q^{-1} v \end{bmatrix}.$$

In particular, plugging in the specific values of a, v, Q in our case yields the explicit expressions:

$$\rho = \frac{1}{g(0) + \frac{g^{(4)}(0)}{4d^2}} = O(1) \quad \text{and} \quad w := \begin{bmatrix} -1 \\ \mathbf{0}_d \\ \frac{1}{\sqrt{2}} \mathbf{1}_d \\ \mathbf{0}_{|\mathcal{S}_2|-d} \end{bmatrix}.$$

Multiplying the right side of (231) by R we deduce:

$$\frac{1}{d} D^{-1} R = \frac{1}{d} D^{-1} G D'_{\leq 1, \leq 1} = \begin{bmatrix} 0 & 0 \\ 0 & Q^{-1} \end{bmatrix} G D'_{\leq 1, \leq 1} + \frac{\rho}{d} w w^\top G D'_{\leq 1, \leq 1}. \tag{232}$$

We will show that the first term on the right side of (232) is close to E_1 while the second-one is negligible. We begin by explicitly computing the first term:

$$\frac{1}{d} \begin{bmatrix} 0 & 0 \\ 0 & Q^{-1} \end{bmatrix} G D'_{\leq 1, \leq 1} = \begin{bmatrix} \frac{\gamma_1}{dq_2} e_2 & \frac{\gamma_2 \sqrt{2}}{dq_3} e_{L(1)+1} & \frac{\gamma_2}{dq_3} \widetilde{I} \end{bmatrix}.$$

Now we compute

$$\frac{\gamma_1}{dq_2} = \frac{\frac{g^{(1)}(0)}{d} + \frac{g^{(3)}(0)}{2d^2} + \frac{g^{(5)}(0)}{4d^3}}{\frac{g^{(1)}(0)}{d} + \frac{g^{(3)}(0)}{2d^2}} = 1 + O(d^{-1}),$$

and similarly

$$\frac{\gamma_2}{dq_3} = \frac{\frac{g^{(2)}(0)}{d^2} + \frac{g^{(4)}(0)}{2d^3}}{\frac{g^{(2)}(0)}{d^2}} = 1 + O(d^{-1}).$$

The triangle inequality therefore directly implies

$$\left\| \frac{1}{d} \begin{bmatrix} 0 & 0 \\ 0 & Q^{-1} \end{bmatrix} G D'_{\leq 1, \leq 1} - E_1 \right\|_{\text{op}} = O(d^{-1}).$$

Moving on to the second term on the right side of (232), a quick computation shows

$$(w^\top G) D'_{\leq 1, \leq 1} = \begin{bmatrix} 0 & -\sqrt{2}+1 & \mathbf{0}_{d-1} \end{bmatrix} D'_{\leq 1, \leq 1} = \begin{bmatrix} 0 & (-\sqrt{2}+1)\gamma_2 & \mathbf{0}_{d-1} \end{bmatrix}$$

Therefore we deduce

$$\left\| \frac{\rho}{d} w w^\top G D'_{\leq 1, \leq 1} \right\|_{\text{op}} = \underbrace{\rho}_{=O(1)} d^{-1} \underbrace{\|w\|_2}_{=O(\sqrt{d})} \| \cdot \underbrace{\|w^\top G D'_{\leq 1, \leq 1}\|_{\text{op}}}_{=O(d^{-1})} = O(d^{-3/2}).$$

This completes the proof.

80

C.2.8 Proof of Proposition C.43

Throughout the proof, set $U := R/d$ and $S_{\leq 2} := \left(I_{\leq 2} + (\rho + \lambda)(nD)^{-1}\right)^{-1}$. By Lemma A.20, there exists a matrix Δ with $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon})$ such that

$$K_\lambda = \Phi_{\leq 2} D \Phi_{\leq 2}^\top + (\rho + \lambda)(I_n + \Delta),$$

where $\rho = g(1) - g_2(1)$. Setting $E := I_n + \Delta$, we use the Sherman-Morrison-Woodbury formula to compute K_λ^{-1} as follows:

$$
\begin{aligned}
K_\lambda^{-1}\Phi_{\leq 2} &= \frac{1}{\rho + \lambda}\left((\rho + \lambda)^{-1}\Phi_{\leq 2}D\Phi_{\leq 2}^\top + E\right)^{-1}\Phi_{\leq 2} \\
&= \frac{1}{\rho + \lambda}\left(E^{-1} - E^{-1}\Phi_{\leq 2}\left((\rho + \lambda)D^{-1} + \Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right)^{-1}\Phi_{\leq 2}^\top E^{-1}\right)\Phi_{\leq 2} \\
&= \frac{1}{\rho + \lambda}E^{-1}\Phi_{\leq 2}\left(I - \left((\rho + \lambda)D^{-1} + \Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right)^{-1}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right).
\end{aligned}
\tag{233}
$$

Using the trivial identity $I - (B + A)^{-1}A = (B + A)^{-1}B$ for any A, B such that $A + B$ is invertible, we deduce

$$I - \left((\rho + \lambda)D^{-1} + \Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right)^{-1}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2} = (\rho + \lambda)\left((\rho + \lambda)D^{-1} + \Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right)^{-1}D^{-1}.$$

Plugging this expression back into Eq. (233) and multiplying through by U, we obtain

$$K_\lambda^{-1}\Phi_{\leq 2}U = \frac{1}{n}E^{-1}\Phi_{\leq 2}\underbrace{\left((\rho + \lambda)(nD)^{-1} + \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right)^{-1}}_{=:V}D^{-1}U. \tag{234}$$

We will next show that V is close to $S_{\leq 2}$ and $\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}/n$ is be close to I_n. We first show that E^{-1} is close to a multiple of identity I_n by computing:

$$
\begin{aligned}
\left\|E^{-1} - I_n\right\|_{\mathrm{op}} &= \left\|(I_n + \Delta)^{-1} - I_n\right\|_{\mathrm{op}} \\
&\leq \frac{\|\Delta\|_{\mathrm{op}}}{1 - \|\Delta\|_{\mathrm{op}}} \\
&= O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon}).
\end{aligned}
\tag{235}
$$

Recall from Lemma A.12 that we have $\left\|\Phi_{\leq 2}^\top\Phi_{\leq 2}/n - I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2 + \epsilon})$. Consequently, the triangle inequality directly implies

$$\left\|\frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2} - I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right). \tag{236}$$

We are now ready to show $\|V - S_{\leq 2}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. First, Eq. (236) implies the bound

$$\left\|V^{-1} - S_{\leq 2}^{-1}\right\|_{\mathrm{op}} = \left\|\frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2} - I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right).$$

Therefore, the norm of V can be bounded as

$$
\begin{aligned}
\|V\|_{\mathrm{op}} &= \frac{1}{\|V^{-1}\|_{\mathrm{op}}} = \frac{1}{\left\|(\rho + \lambda)(nD)^{-1} + \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right\|_{\mathrm{op}}} \\
&\leq \frac{1}{\left\|\frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right\|_{\mathrm{op}} - \left\|(\rho + \lambda)(nD)^{-1}\right\|_{\mathrm{op}}} \\
&\leq \frac{1}{1 - \left\|\frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2} - I_n\right\|_{\mathrm{op}} - \left\|(\rho + \lambda)(nD)^{-1}\right\|_{\mathrm{op}}}.
\end{aligned}
$$

81

Using (236) and (231), when d is sufficiently large, we have $\|V\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. Since for two invertible matrices A, B, we have $\|A - B\|_{\mathrm{op}} = \|A(A^{-1} - B^{-1})B\|_{\mathrm{op}}$, we conclude

$$\|V - S_{\leq 2}\|_{\mathrm{op}} \leq \|V\|_{\mathrm{op}} \|S_{\leq 2}\|_{\mathrm{op}} \left\|V^{-1} - S_{\leq 2}^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right). \tag{237}$$

Consequently, using the triangle inequality together with Eq. (236) and (237), we get

$$\left\|S_{\leq 2} - \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}V\right\|_{\mathrm{op}} \leq \|S_{\leq 2} - V\|_{\mathrm{op}} + \left\|V - \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}V\right\|_{\mathrm{op}}$$

$$\leq \|S_{\leq 2} - V\|_{\mathrm{op}} + \|V\|_{\mathrm{op}}\left\|I - \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}\right\|_{\mathrm{op}}$$

$$= O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right).$$

We are now ready to establish the claimed estimate (220). To this end, using Eq. (234) we compute

$$\left\|S_{\leq L(2)}D^{-1}U - \Phi_{\leq 2}K_\lambda^{-1}\Phi_{\leq 2}U\right\|_{\mathrm{op}} = \left\|S_{\leq 2}D^{-1}U - \frac{1}{n}\Phi_{\leq 2}^\top E^{-1}\Phi_{\leq 2}VD^{-1}U\right\|_{\mathrm{op}}$$

$$\leq \left\|S_{\leq L(2)} - \frac{1}{n}\Phi_{\leq L(2)}^\top E^{-1}\Phi_{\leq 2}V\right\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}}$$

$$= O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right). \tag{238}$$

where in the last inequality we use the fact that $\left\|D^{-1}U\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ by Proposition C.42. It remains to argue that $S_{\leq L(2)}$ is close to the identity. To see this, recall that the inverse of D was computed in (231). In particular, the triangle inequality shows

$$\|(nD)^{-1}\|_{\mathrm{op}} \leq O(d^{-\delta}).$$

Consequently, setting $\hat{\Delta} = (\rho + \lambda)(nD)^{-1}$ we estimate

$$\|I - S_{\leq 2}\|_{\mathrm{op}} = \|I - (I + \hat{\Delta})^{-1}\|_{\mathrm{op}} \leq \frac{\|\hat{\Delta}\|_{\mathrm{op}}}{1 - \|\hat{\Delta}\|_{\mathrm{op}}} = O(d^{-\delta}).$$

Combining this estimate with (238) and using the triangle inequality completes the proof of (220).

Next, we establish the estimate (221). To this end, using the triangle inequality and the submultiplicativity of the spectral norm, we obtain:

$$\left\|\sqrt{n}K_\lambda^{-1}\Phi_{\leq 2}U - \frac{1}{\sqrt{n}}\Phi_{\leq 2}S_{\leq 2}D^{-1}U\right\|_{\mathrm{op}}$$

$$= \left\|\frac{1}{\sqrt{n}}E^{-1}\Phi_{\leq 2}VD^{-1}U - \frac{1}{\sqrt{n}}\Phi_{\leq 2}S_{\leq 2}D^{-1}U\right\|_{\mathrm{op}}$$

$$\leq \left\|\frac{1}{\sqrt{n}}E^{-1}\Phi_{\leq 2}V - \frac{1}{\sqrt{n}}\Phi_{\leq 2}S_{\leq 2}\right\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}}$$

$$\leq \frac{1}{\sqrt{n}}\left\|(E^{-1} - I_n)\Phi_{\leq 2}V\right\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}} + \frac{1}{\sqrt{n}}\left\|\Phi_{\leq 2}(V - S_{\leq 2})\right\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}}$$

$$\leq \frac{1}{\sqrt{n}}\left\|E^{-1} - I_n\right\|_{\mathrm{op}}\|\Phi_{\leq 2}\|_{\mathrm{op}}\|V\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}} + \frac{1}{\sqrt{n}}\|\Phi_{\leq 2}\|_{\mathrm{op}}\|V - S_{\leq 2}\|_{\mathrm{op}}\left\|D^{-1}U\right\|_{\mathrm{op}}.$$

Recall that we have already shown the estimates

$$\|V\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \quad \left\|E^{-1} - I_n\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{\delta - 1/2 + \epsilon}),$$

$$\|V - S_{\leq 2}\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta - 1/2 + \epsilon} + d^{-\delta/2 + \epsilon}\right). \tag{239}$$

Additionally, by Lemma A.11, we have $\|\Phi_{\leq 2}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ and by Proposition C.42 we have $\|D^{-1}U\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. With these bounds, we deduce that

$$\left\|\sqrt{n}K_\lambda^{-1}\Phi_{\leq 2}U - \frac{1}{\sqrt{n}}\Phi_{\leq 2}S_{\leq 2}D^{-1}U\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right).$$

Moreover, recall that $\|I - S_{\leq 2}\|_{\mathrm{op}} = O(d^{-\delta})$. Consequently, using the triangle inequality, we deduce

$$\left\|\sqrt{n}K_\lambda^{-1}\Phi_{\leq 2}U - \frac{1}{\sqrt{n}}\Phi_{\leq 2}D^{-1}U\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}\left(d^{\delta-1/2+\epsilon} + d^{-\delta/2+\epsilon}\right),$$

which completes the proof of (221).

C.3   Proof of Propositions for Theorem 4.6

C.3.1   Proof of Proposition B.25

The proof is analogous to that of Proposition B.24. Let $\mathcal{E}(r)$ denote the union of all sets $S \subset [d]$ containing r with $|S| \leq p$, i.e., $\mathcal{E}(r) = \cup_{j=1}^p \mathcal{S}_j^1$. Similarly, define the complement $\mathcal{E}^c(r) = \left(\cup_{j=0}^\ell \mathcal{S}_j\right) \setminus \mathcal{E}(r)$.

Recall that from (18) we have

$$\mathbb{E}[\partial_r f_{\leq p}^*(x)]^2 = \sum_{S \in \mathcal{E}(r)} b_S^2 = \|b_{\leq p}^\top E_r\|_2^2,$$

where E_r is the partial identity matrix of size $L(p) \times L(p)$ defined as $(E_r)_{\mathcal{I},\mathcal{I}} = \mathbb{1}\{\mathcal{I} \in \mathcal{E}(r)\}$. We now write $\beta = \beta_{\mathcal{E}(r)} + \beta_{\mathcal{E}^c(r)} + \beta_\varepsilon$ where we define:

$$\beta_{\mathcal{E}(r)} := K_\lambda^{-1}(\Phi_{\mathcal{E}(r)} b_{\mathcal{E}(r)}), \quad \beta_{\mathcal{E}^c(r)} := K_\lambda^{-1}(\Phi_{\mathcal{E}^c(r)} b_{\mathcal{E}^c(r)}), \quad \beta_\epsilon := K_\lambda^{-1}\varepsilon.$$

With this decomposition of β, we accordingly bound the norm by triangle inequality:

$$\begin{aligned}
&\left\|\beta^\top \Phi_{\leq p} R - b_{\leq p}^\top E_r\right\|_2 \\
&\leq \left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\leq p} R - b_{\leq p}^\top E_r\right\|_2 + \left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\leq p} R\right\|_2 + \left\|\beta_\varepsilon^\top \Phi_{\leq p} R\right\|_2.
\end{aligned} \tag{240}$$

We will show that all the terms on the right-hand-side of the above equation are small.

The following two Propositions control the first and second terms in Eq. (240) respectively. We defer the proofs to Appendix C.3.3 and C.3.5 respectively.

Proposition C.44. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$\left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\leq p} R - b_{\leq p}^\top E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}.$$

Proposition C.45. *The estimate holds uniformly over all coordinates r:*

$$\left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\leq p} R\right\|_2 = O_{d,\mathbb{P}}\left(d^{-1/2}\right) \cdot \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}. \tag{241}$$

Next we show that the third term on the right side of (240) is small. The proof of the proposition appears in Appendix C.3.6.

Proposition C.46. *For any $\epsilon > 0$, the following estimate holds*

$$\left\|\beta_\varepsilon^\top \Phi_{\leq p} R\right\|_2 = O_{d,\mathbb{P}}(d^{-(1+\delta)/2+\epsilon}) \cdot \sigma_\varepsilon.$$

Combining the results of three propositions with Eq. (240), using the AM-GM inequality to bound the cross-term directly yields the estimate

$$\left\|\beta^\top \Phi_{\leq p} R - b_{\leq p}^\top E_r\right\|_2^2 = O_{d,\mathbb{P}}\left(d^{-1}\right) \cdot \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\delta+2\epsilon} + d^{\delta-1+2\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}^2$$
$$+ O_{d,\mathbb{P}}(d^{-1-\delta+2\epsilon}) \cdot \sigma_\varepsilon^2.$$

Using the elementary inequality $|\|a\|_2^2 - \|b\|_2^2| \leq \|a - b\|_2^2 + 2\|b\|_2 \|a - b\|_2$ for any vectors a, b yields

$$\left|\left\|\beta^\top \Phi_{\leq p} R\right\|_2^2 - \left\|b_{\leq p}^\top E_r\right\|_2^2\right|$$
$$= O_{d,\mathbb{P}}\left(d^{-1}\right) \cdot \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\delta+2\epsilon} + d^{\delta-1+2\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-1-\delta+2\epsilon}) \cdot \sigma_\varepsilon^2$$
$$+ \left\|\partial_r f_{\leq p}^*\right\|_{L_2} \cdot \sqrt{O_{d,\mathbb{P}}\left(d^{-1}\right) \cdot \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\delta+2\epsilon} + d^{\delta-1+2\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-1-\delta+2\epsilon}) \cdot \sigma_\varepsilon^2}.$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we arrive at the bound

$$O_{d,\mathbb{P}}\left(d^{-1}\right) \cdot \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}^2 + O_{d,\mathbb{P}}(d^{-1-\delta+2\epsilon}) \cdot \sigma_\varepsilon^2$$
$$+ O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2} \left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2} + O_{d,\mathbb{P}}(d^{-(1+\delta)/2+\epsilon})\sigma_\varepsilon \left\|\partial_r f_{\leq p}^*\right\|_{L_2}.$$

Invoking the inequality $\left\|f_{\mathcal{E}^c(r)}^*\right\|_{L_2}^2 \leq \|f^*\|_{L_2}^2$ and applying the inequality

$$O_{d,\mathbb{P}}(d^{-(1+\delta)/2+\epsilon}) \cdot \sigma_\varepsilon \left\|\partial_r f_{\leq p}^*\right\|_{L_2} \leq O_{d,\mathbb{P}}(d^{-1}) \cdot \sigma_\varepsilon^2 + O_{d,\mathbb{P}}(d^{-\delta+2\epsilon}) \cdot \left\|\partial_r f_{\leq p}^*\right\|_{L_2}^2$$

completes the proof.

C.3.2   Proof of Proposition B.26

We first simplify the expression. By (72), the diagonal matrix R_{p+1}^1 has entries $g^{(p+1)}(0)d^{-p-1}+O_d(d^{-p-2})$. Consequently, we compute

$$\left\|\beta^\top \Phi_{\mathcal{S}_{p+1}^1} R_{p+1}^1 - \frac{g^{(p+1)}(0)}{d^{p+1}}\beta^\top \Phi_{\mathcal{S}_{p+1}^1}\right\|_2$$
$$\leq \|\beta^\top \Phi_{\mathcal{S}_{p+1}^1}\|_2 \cdot O_d(d^{-p-2})$$
$$\leq \|y\|_2 \left\|K_\lambda^{-1}\right\|_{\text{op}} \left\|\Phi_{\mathcal{S}_{p+1}^1}\right\|_{\text{op}} \cdot O_d(d^{-p-2})$$
$$= O_{d,\mathbb{P}}(d^{-2+\delta} \log^{1/2}(d))(\|f^*\|_{L_2} + \sigma_\varepsilon), \tag{242}$$

where we used the fact that $\left\|\Phi_{\mathcal{S}_{p+1}^1}\right\|_{\text{op}} = \|X_r\|_{\text{op}} \cdot \left\|\Phi_{\mathcal{S}_p^0}\right\|_{\text{op}} \leq \|\Phi_{\leq p}\|_{\text{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ from Lemma A.11 and the basic estimate $\|y\|_2 = O_{d,\mathbb{P}}(\sqrt{n}\log^{1/2}(d))(\|f^*\|_{L_2} + \sigma_\varepsilon)$. The bound $\left\|K_\lambda^{-1}\right\|_{\text{op}} \leq 1/(\rho + \lambda - \|\Delta_1\|_{\text{op}}) = O_{d,\mathbb{P}}(1)$ is implied by the expression (65).

We now split the whole index set $\cup_{j=0}^\ell \mathcal{S}_j$ into the disjoint union $\mathcal{Q} \sqcup \mathcal{S}_{p+1}^1$ for the family of sets $\mathcal{Q} = \cup_{j=0}^\ell \mathcal{S}_j \setminus \mathcal{S}_{p+1}^1$. Then we correspondingly decompose y and β as

$$y = \Phi_\mathcal{Q} b_\mathcal{Q} + \Phi_{\mathcal{S}_{p+1}^1} b_{\mathcal{S}_{p+1}^1} + \varepsilon,$$
$$\beta = \underbrace{K_\lambda^{-1} \Phi_\mathcal{Q} b_\mathcal{Q}}_{=: \beta_\mathcal{Q}} + \underbrace{K_\lambda^{-1} \Phi_{\mathcal{S}_{p+1}^1} b_{\mathcal{S}_{p+1}^1}}_{=: \beta_c} + \underbrace{K_\lambda^{-1} \varepsilon}_{=: \beta_\epsilon}.$$

Note the equality $\left\|b_{\mathcal{S}^1_{p+1}}\right\|_2 = \left\|\partial_r f^*_{p+1}\right\|_{L_2}$. Our goal now is to show the following estimate

$$\left\|d^{-p-1}\beta^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}\right\|_2 = o_{d,\mathbb{P}}(1),$$

which combined with Eq. (242) will quickly complete the proof. To this end, using the decomposition of β, we apply triangle inequality to the norm and obtain

$$\left\|d^{-p-1}\beta^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}\right\|_2$$
$$\leq \left\|d^{-p-1}\beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}^1_{p+1}}\right\|_2 + \left\|d^{-p-1}\beta_c^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}^\top\right\|_2 + \left\|d^{-p-1}\beta_\varepsilon^\top \Phi_{\mathcal{S}^1_{p+1}}\right\|_2. \tag{243}$$

Next, we show that all three terms on the right side of the above equation are of the order $o_{d,\mathbb{P}}(1)$. For the first term, we apply the following proposition. The proof is deferred to Appendix C.3.7.

Proposition C.47. *The following estimate holds uniformly for all $r \in [d]$:*

$$\left\|\frac{1}{d^{p+1}}\beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}^1_{p+1}}\right\|_2^2 = O_d(d^{-1})\left\|b_{\mathcal{Q}}\right\|_2^2.$$

Now we proceed to the second term. The following proposition gives an estimate for the second term. We defer the proof to Appendix C.3.8.

Proposition C.48. *For any $\epsilon > 0$, the following estimate holds uniformly for all $r \in [d]$*

$$\left\|\frac{1}{n}\beta_c^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}b_{\mathcal{S}^1_{p+1}}^\top\right\|_2 = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2. \tag{244}$$

Multiplying (244) through by $d^{\delta-1}$ we obtain

$$\left\|d^{-p-1}\beta_c^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}^\top\right\|_2 = O_{d,\mathbb{P}}(d^{\delta/2-1+\epsilon} + d^{3(\delta-1)/2+\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2. \tag{245}$$

The following proposition gives an estimate for the third term. We defer the proof to Appendix C.3.9.

Proposition C.49. *For any $\epsilon > 0$, the following estimate holds uniformly for all $r \in [d]$:*

$$\left\|d^{-p-1}\beta_\varepsilon^\top \Phi_{\mathcal{S}^1_{p+1}}\right\|_2 = O_{d,\mathbb{P}}(d^{-1+\delta/2+\epsilon}) \cdot \sigma_\varepsilon \tag{246}$$

Now, plugging Eq. (245), (245) and (246) into Eq. (243) yields:

$$\left\|d^{-p-1}\beta^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}^\top\right\|_2$$
$$= O_{d,\mathbb{P}}(d^{\delta/2-1+\epsilon} + d^{3(\delta-1)/2+\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2 + O_{d,\mathbb{P}}(d^{-1/2}) \cdot \left\|f^*\right\|_{L_2} + O_{d,\mathbb{P}}(d^{-1+\delta/2+\epsilon}) \cdot \sigma_\varepsilon,$$

where we use the fact $\|b_{\mathcal{Q}}\|_2 = \left\|f^*_{\mathcal{Q}}\right\|_{L_2} \leq \|f^*\|_{L_2}$.

Squaring both sides yields:

$$\left\|d^{-p-1}\beta^\top \Phi_{\mathcal{S}^1_{p+1}} - (\rho+\lambda)^{-1}d^{\delta-1}b_{\mathcal{S}^1_{p+1}}^\top\right\|_2^2 = O_{d,\mathbb{P}}(d^{\delta-2+2\epsilon} + d^{3(\delta-1)+2\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2^2 + O_{d,\mathbb{P}}(d^{-1}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2).$$

Combining this bound with Eq.(242) and using the elementary inequality $\left|\|a\|_2^2 - \|b\|_2^2\right| \leq \|a-b\|_2^2 + 2\|b\|_2\|a-b\|_2$ for any vectors a, b gives

$$\left|\left\|d^{-p-1}\beta^\top \Phi_{\mathcal{S}^1_{p+1}}\right\|_2^2 - (\rho+\lambda)^{-2}\left\|d^{\delta-1}b_{\mathcal{S}^1_{p+1}}\right\|_2^2\right|$$

$$= O_{d,\mathbb{P}}(d^{\delta-2+2\epsilon} + d^{3(\delta-1)+2\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2^2 + O_{d,\mathbb{P}}(d^{-1}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)$$

$$+ d^{\delta-1}\left\|b_{\mathcal{S}^1_{p+1}}\right\|_2 \cdot \sqrt{O_{d,\mathbb{P}}(d^{\delta-2+2\epsilon} + d^{3(\delta-1)+2\epsilon}) \cdot \left\|b_{\mathcal{S}^1_{p+1}}\right\|_2^2 + O_{d,\mathbb{P}}(d^{-1}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2)}.$$

Invoking the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b > 0$ the bound reduces to

$$O_{d,\mathbb{P}}(d^{-1}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{\frac{3}{2}\delta - 2 + \epsilon} + d^{\frac{5}{2}(\delta-1)+\epsilon}) \cdot \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2^2 + O_{d,\mathbb{P}}(d^{\delta - 3/2}) \cdot \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2 (\|f^*\|_{L_2} + \sigma_\varepsilon).$$

Combining with the fact $\left\| b_{\mathcal{S}_{p+1}^1} \right\|_2 = \left\| \partial_r f_{p+1}^* \right\|_{L_2}$ completes the proof.

C.3.3   Proof of Proposition C.44

The proof is nearly identical to that of Proposition C.43. In the proof, we will frequently use the following estimate for any Δ with $\|\Delta\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$:

$$\left\| (I + \Delta)^{-1} - I \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}}), \tag{247}$$

which can be deduced using Neumann series. Recall that Eq. (65) shows

$$K = \Phi_{\le p} D \Phi_{\le p}^\top + \rho I_n + \Delta_1,$$

where $\rho = g(1) - g_p(1)$ and Δ_1 satisfies $\|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon})$. We now write

$$K_\lambda = \Phi_{\le p} D \Phi_{\le p}^\top + (\rho + \lambda) H, \tag{248}$$

where $H = I_n + (\rho + \lambda)^{-1} \Delta_1$. Invoking Sherman-Morrison-Woodbury formula and following exactly the same argument as in Proposition C.2.8 (equation (234)) directly implies

$$\beta_{\mathcal{E}(r)}^\top \Phi_{\le p} R = b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top K_\lambda^{-1} \Phi_{\le p} R$$

$$= \frac{1}{n} b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top H^{-1} \Phi_{\le p} \left(\underbrace{(\rho + \lambda)(nD)^{-1} + \frac{\Phi_{\le p}^\top H^{-1} \Phi_{\le p}}{n}}_{G} \right)^{-1} \left(D^{-1} R \right). \tag{249}$$

Now we simplify G^{-1}. First using triangle inequality we have

$$\|G - I\|_{\mathrm{op}} \le (\rho + \lambda) \left\| (nD)^{-1} \right\|_{\mathrm{op}} + \left\| \frac{\Phi_{\le p}^\top H^{-1} \Phi_{\le p}}{n} - I_n \right\|_{\mathrm{op}}. \tag{250}$$

Based on the definition of D, we immediately have $\left\| (nD)^{-1} \right\|_{\mathrm{op}} = O_d(d^{-\delta})$. Notice that $\left\| \frac{\Phi_{\le p}^\top \Phi_{\le p}}{n} - I_n \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon})$ (Theorem A.12) and

$$\left\| H^{-1} - I_n \right\|_{\mathrm{op}} = O_d(\|\Delta_1\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon}), \tag{251}$$

which follows from Eq. (247). Thus we deduce

$$\left\| \frac{\Phi_{\le p}^\top H^{-1} \Phi_{\le p}}{n} - I \right\|_{\mathrm{op}} = \left\| \frac{\Phi_{\le p}^\top \Phi_{\le p}}{n} + \frac{\Phi_{\le p}^\top (H^{-1} - I) \Phi_{\le p}}{n} - I \right\|_{\mathrm{op}}$$

$$\le \left\| \frac{\Phi_{\le p}^\top \Phi_{\le p}}{n} - I \right\|_{\mathrm{op}} + \left\| \frac{\Phi_{\le p}}{\sqrt{n}} \right\|_{\mathrm{op}}^2 \|H^{-1} - I\|_{\mathrm{op}}$$

$$= O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}). \tag{252}$$

Plugging the above bound into Eq. (250) we obtain

$$\|G - I\|_{\mathrm{op}} = O_d(d^{-\delta}) + O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}).$$

Invoking Eq. (247) again we obtain

$$\left\|G^{-1} - I\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}). \tag{253}$$

The following claim gives an estimate for $D^{-1}R$. The proof is deferred to Appendix C.3.4.

Claim 15. *The following estimate holds*

$$\left\|D^{-1}R - E_r\right\|_{\text{op}} = O_d(d^{-1}). \tag{254}$$

We apply this estimate for $\frac{1}{d}D^{-1}R$ to Eq. (249) and obtain

$$\left\|\beta_{\mathcal{E}(r)}^{\top}\Phi_{\leq p}R - \frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}H^{-1}\Phi_{\leq p}G^{-1}E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-1}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2, \tag{255}$$

where we use the fact that $\left\|H^{-1}\right\|_{\text{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (251)), $\left\|G^{-1}\right\|_{\text{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (253)), and $\left\|\Phi_{\mathcal{E}(r)}\right\|_{\text{op}} \leq \left\|\Phi_{\leq p}\right\|_{\text{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11).

Next, we will show

$$\left\|\frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}H^{-1}\Phi_{\leq p}G^{-1}E_r - b_{\mathcal{E}(r)}\right\|_2 = o_{d,\mathbb{P}}(1), \tag{256}$$

then plugging it back into Eq. (255) complete the proof. To this end, applying the estimates (251) and (253) along with the triangle inequality yields

$$\left\|\frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}H^{-1}\Phi_{\leq p}G^{-1}E_r - \frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}\Phi_{\leq p}E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2, \tag{257}$$

where we used the fact that $\|E_r\|_{\text{op}} = O_d(1)$.

Next, invoking the bound $\left\|\Phi_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}/n - I\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-(1+\delta)/2+\epsilon})$ (Theorem A.12) and the definition of E_r, we have

$$\begin{aligned}
\left\|\frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}\Phi_{\leq p}E_r - b_{\leq p}^{\top}E_r\right\|_2 &= \left\|\frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)} - b_{\mathcal{E}(r)}^{\top}\right\|_2 \\
&\leq \left\|b_{\mathcal{E}(r)}^{\top}\left(\Phi_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}/n - I\right)\right\|_2 \\
&\leq \left\|b_{\mathcal{E}(r)}\right\|_2 \cdot O_{d,\mathbb{P}}(d^{-(\delta+1)/2+\epsilon}).
\end{aligned}$$

Plugging this estimate into Eq. (257) yields

$$\left\|\frac{1}{n}b_{\mathcal{E}(r)}^{\top}\Phi_{\mathcal{E}(r)}^{\top}H^{-1}\Phi_{\leq p}G^{-1}E_r - b_{\leq p}^{\top}E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2,$$

which completes the proof.

C.3.4 Proof of Claim 15

Recall that D is a diagonal matrix with entry $D_{\mathcal{I},\mathcal{I}} = g^{(|\mathcal{I}|)}(0)d^{-|\mathcal{I}|} + O_d(d^{-|\mathcal{I}|-1})$. Using the expression of each block of R, namely Eq. (71) and (72), we can compute each diagonal entry of $D^{-1}R$:

$$D_{\mathcal{I},\mathcal{I}}^{-1}R_{\mathcal{I},\mathcal{I}} = \begin{cases} 1 + O_d(d^{-1}), & \text{if } \mathcal{I} \in \cup_{j=1}^{p}\mathcal{S}_j^1; \\ O_d(d^{-2}), & \text{if } \mathcal{I} \in \cup_{j=0}^{p-1}\mathcal{S}_j^0. \end{cases}$$

Recalling the definition of E_r as $(E_r)_{\mathcal{I},\mathcal{I}} = \mathbb{1}\{\mathcal{I} \in \mathcal{E}(r)\}$, the claimed bound follows immediately.

C.3.5 Proof of Proposition C.45

In the proof, we will frequently use the following estimate for any Δ with $\|\Delta\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, which can be deduced from the Neumann series:

$$\left\|(I + \Delta)^{-1} - I\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}}), \tag{258}$$

therefore we have $(I + \Delta)^{-1} = I + \Delta'$ with some Δ' satisfying $\|\Delta'\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}})$. Now we are ready to present the proof. For simplicity of notation, we omit (r) in $\mathcal{E}(r)$ and $\mathcal{E}^c(r)$.

First, we expand K_λ^{-1} by using Eq. (46) in Lemma A.17. Consequently, we have

$$K_\lambda = K_p + \underbrace{(\rho + \lambda)I_n + g^{(p+1)}(0)d^{-p-1}\mathsf{off}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right) + \Delta}_{:=(\rho+\lambda)H}, \tag{259}$$

with $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\delta-2)/2+\epsilon}) = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$, and we denote

$$H := I_n + (\rho + \lambda)^{-1}g^{(p+1)}(0)d^{-p-1}\mathsf{off}\left(\Phi_{\mathcal{S}_k}\Phi_{\mathcal{S}_k}^\top\right) + (\rho + \lambda)^{-1}\Delta. \tag{260}$$

Analogously to Eq. (248), we write $K_\lambda = K_p + (\rho+\lambda)H$ and use the Sherman-Morrison-Woodbury formula to express $K_\lambda^{-1} = (K_p + (\rho + \lambda)H)^{-1}$ yielding:

$$\beta_{\mathcal{E}^c}^\top \Phi_{\leq p} R = \frac{1}{n}b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1}\Phi_{\leq p}\left(\underbrace{(\rho + \lambda)(nD)^{-1} + \frac{\Phi_{\leq p}^\top H^{-1}\Phi_{\leq p}}{n}}_{G}\right)^{-1}\left(D^{-1}R\right). \tag{261}$$

We further simplify Eq. (261). By the triangle inequality, the following holds

$$\left\|\beta_{\mathcal{E}^c}^\top \Phi_{\leq p} R - \frac{1}{n}b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1}\Phi_{\leq p}G^{-1}E_r\right\|_2$$

$$= \left\|\frac{1}{n}b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1}\Phi_{\leq p}G^{-1}\left(D^{-1}R - E_r\right)\right\|_2$$

$$\leq \left\|\frac{1}{\sqrt{n}}\Phi_{\mathcal{E}^c}b_{\mathcal{E}^c}\right\|_2 \left\|H^{-1}\right\|_{\mathrm{op}} \left\|\frac{\Phi_{\leq p}}{\sqrt{n}}\right\|_{\mathrm{op}} \left\|G^{-1}\right\|_{\mathrm{op}} \left\|D^{-1}R - E_r\right\|_{\mathrm{op}}$$

$$= O_{d,\mathbb{P}}(d^{-1}\log(d)) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}, \tag{262}$$

where $\left\|\frac{1}{\sqrt{n}}\Phi_{\mathcal{E}^c}b_{\mathcal{E}^c}\right\|_2 = O_{d,\mathbb{P}}(\log(d))\|f_{\mathcal{E}^c}^*\|_{L_2}$ (Markov's inequality), $\left\|H^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (251)), $\left\|\frac{\Phi_{\leq p}}{\sqrt{n}}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Lemma A.11), $\left\|G^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (253)) and $\left\|D^{-1}R - E_r\right\|_{\mathrm{op}} = O_d(d^{-1})$ follows from Proposition 15. Consequently, there exists Δ_1 that satisfies $\|\Delta_1\|_2 = O_{d,\mathbb{P}}(d^{-1}\log(d)) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}$ such that the following holds:

$$\beta_{\mathcal{E}^c}^\top \Phi_{\leq p} R = \frac{1}{n}b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1}\Phi_{\leq p}G^{-1}E_r + \Delta_1. \tag{263}$$

Next, we express H^{-1} and G^{-1} in terms of feature matrices. Note that for any $\Delta = o_{d,\mathbb{P}}(1)$, we can express $(I + \Delta)^{-1}$ using the Neumann series:

$$(I + \Delta)^{-1} = I + \sum_{k=1}^{\infty}(-\Delta)^k.$$

We compute the inverse of H in Eq. (260) using Neumann series yielding

$$H^{-1} = I_n + \sum_{q=1}^{\infty}\left(\underbrace{-(\rho + \lambda)^{-1}g^{(p+1)}(0)}_{:=c_{p+1}}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top) - (\rho + \lambda)^{-1}\Delta\right)^q. \tag{264}$$

Note that the triangle inequality on Eq. (45) and (46) shows

$$d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top}) = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon}),$$

where $\mathsf{off}(\cdot)$ denotes the off-diagonal component of a matrix.

The following claim now show that we can truncate higher-order terms in H^{-1} at constant order to achieve a desired error on the order of $O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$.

Claim 16. *Setting a constant* $c_H := \lceil \frac{\delta+4\epsilon}{1-\delta-2\epsilon} \rceil$, *there exists* Δ_2 *that satisfies* $\|\Delta_2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$ *such that* H^{-1} *can be written as*

$$H^{-1} = I_n + \sum_{q=1}^{c_H} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top})\right)^q + \Delta_2. \tag{265}$$

Proof. We expand $\sum_{q=1}^{\infty} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top}) + (\rho+\lambda)^{-1}\Delta\right)^q$ in Eq. (264). Note that by the triangle inequality and formula for the summation of a geometric series, the following holds:

$$\left\| \sum_{q=c_H+1}^{\infty} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top}) + (\rho+\lambda)^{-1}\Delta\right)^q \right\|_{\mathrm{op}} \leq \sum_{q=c_H+1}^{\infty} (O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon}))^q$$

$$= O_{d,\mathbb{P}}(d^{((\delta-1)/2+\epsilon)(c_H+1)})$$

$$= O_{d,\mathbb{P}}(d^{-1/2-\epsilon}).$$

Next, clearly since $c_H \geq 1$ is a constant we have

$$\sum_{q=1}^{c_H} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top}) + (\rho+\lambda)^{-1}\Delta\right)^q = \sum_{q=1}^{c_H} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top})\right)^q + \Delta_2,$$

for some Δ_2 satisfying $\|\Delta_2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$, which completes the proof. $\square$

We now plug the expression (265) for H^{-1} into G in Eq. (261). We further substitute $\frac{\Phi_{\leq p}^{\top}\Phi_{\leq p}}{n}$ for $\mathsf{off}\left(\frac{\Phi_{\leq p}^{\top}\Phi_{\leq p}}{n}\right) + I_{L(p)}$ and we obtain

$$G = (\rho+\lambda)(nD)^{-1} + I + \mathsf{off}\left(\frac{\Phi_{\leq p}^{\top}\Phi_{\leq p}}{n}\right) + \frac{\Phi_{\leq p}^{\top}}{\sqrt{n}} \sum_{q=1}^{c_H} \left(-c_{p+1}d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top})\right)^q \frac{\Phi_{\leq p}}{\sqrt{n}} + \Delta_3,$$

with $\|\Delta_3\|_{\mathrm{op}} \leq \|\Phi_{\leq p}/\sqrt{n}\|_{\mathrm{op}}^2 \|\Delta_2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$.

Computing G^{-1} using Neumann series again gives us

$$G^{-1} = I + \sum_{r=1}^{\infty} \left(-(\rho+\lambda)(nD)^{-1} - \mathsf{off}\left(\frac{\Phi_{\leq p}^{\top}\Phi_{\leq p}}{n}\right) - \frac{\Phi_{\leq p}^{\top}}{\sqrt{n}} \sum_{q=1}^{c_H} \left(-c_{p+1}\frac{\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top})}{d^{p+1}}\right)^q \frac{\Phi_{\leq p}}{\sqrt{n}} + \Delta_3\right)^r,$$

Similarly, we can truncate the infinite summation while maintaining the same error order. Note that each item in the parenthesis is of the order $o_{d,\mathbb{P}}(1)$, which can be deduced using the following estimates:

$$\|(nD)^{-1}\|_{\mathrm{op}} = O_d(d^{-\delta}), \quad \left\|\mathsf{off}\left(\frac{\Phi_{\leq p}^{\top}\Phi_{\leq p}}{n}\right)\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon}), \quad \left\|\frac{\Phi_{\leq p}^{\top}}{\sqrt{n}}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1),$$

$$\left\|d^{-p-1}\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^{\top})\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon}), \quad \|\Delta_3\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon}),$$

where particularly the second estimate can be easily deduced from Lemma A.13 since $\mathsf{diag}\left(\frac{\Phi_{\leq p}^\top \Phi_{\leq p}}{n}\right) = I_n$.

Therefore, a completely analogous argument to Claim 16 shows that we can find a constant $c_G = \max(c_H, \lceil\frac{1-\delta+4\epsilon}{\delta-2\epsilon}\rceil)$ and a matrix Δ_4 that satisfies $\|\Delta_4\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$ such that the following holds

$$G^{-1} = I + \sum_{r=1}^{c_G}\left(-(\rho+\lambda)(nD)^{-1} - \mathsf{off}\left(\frac{\Phi_{\leq p}^\top \Phi_{\leq p}}{n}\right) - \frac{\Phi_{\leq p}^\top}{\sqrt{n}}\sum_{q=1}^{c_H}\left(-c_{p+1}\frac{\mathsf{off}(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top)}{d^{p+1}}\right)^q\frac{\Phi_{\leq p}}{\sqrt{n}}\right)^r + \Delta_4. \tag{266}$$

To ease the analysis, we replace $\mathsf{off}\left(\frac{\Phi_{\leq p}^\top \Phi_{\leq p}}{n}\right)$ by $\frac{\Phi_{\leq p}^\top \Phi_{\leq p}}{n} - I_{|L(p)|}$ and replace $\frac{\mathsf{off}\left(\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top\right)}{d^{p+1}}$ by $\frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}} - h_{p+1}I_n$, where $h_{p+1} = \binom{d}{p+1}/d^{p+1} = O_d(1)$. Then Eq. (265) and (266) take the form

$$H^{-1} = I_n + \sum_{q=1}^{c_H}\left(-c_{p+1}\frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}} + c_{p+1}h_{p+1}I_n\right)^q + \Delta_2, \tag{267}$$

and

$$G^{-1} = I_n + \sum_{r=1}^{c_G}\left[-(\rho+\lambda)(nD)^{-1} + I_{|L(p)|} - \frac{\Phi_{\leq p}^\top \Phi_{\leq p}}{n}\right.$$
$$\left. + \frac{\Phi_{\leq p}^\top}{\sqrt{n}}\sum_{q=1}^{c_H}\left(-c_{p+1}\frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}} + c_{p+1}h_{p+1}I_n\right)^q\frac{\Phi_{\leq p}}{\sqrt{n}}\right]^r + \Delta_4. \tag{268}$$

Plugging the approximations of H^{-1} (Eq. (267)) and G^{-1} (Eq. (268)) into Eq. (263) and expanding the powers we see that $b_{\mathcal{E}^c}^\top\frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}}H^{-1}\frac{\Phi_{\leq p}}{\sqrt{n}}G^{-1}E_r$ can be written as a finite sum:

$$b_{\mathcal{E}^c}^\top\frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}}H^{-1}\frac{\Phi_{\leq p}}{\sqrt{n}}G^{-1}E_r = \sum_{i=1}^{c_p}b_{\mathcal{E}^c}^\top\frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}}P_i + \Delta_5, \tag{269}$$

where $c_p \in \mathbb{N}$ is a constant and each matrix $P_i \in \mathbb{R}^{L(p)\times L(p)}$ takes the form

$$P_i = C_i A_{i_1} A_{i_2}\ldots A_{i_m}\frac{\Phi_{\leq p}(nD)^{-t_i}E_r}{\sqrt{n}},$$

with each $A_{i_j} \in \left\{I_n, \frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}}\right\} \cup \left\{\frac{\Phi_{\leq p}(nD)^{-s}\Phi_{\leq p}^\top}{n}\right\}_{s=0}^{c_G}$ for $j \in [m]$. Here $C_i > 0$ is a constant and $t_i \in \mathbb{N}$. Additionally, we can deduce that Δ_5 satisfies

$$\|\Delta_5\|_2 = O_{d,\mathbb{P}}(\log^{1/2}(d))\cdot\|f_{\mathcal{E}^c}^*\|_{L_2}\cdot(\|\Delta_2\|_{\mathrm{op}} + \|\Delta_4\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(d^{-\frac{1}{2}})\cdot\|f_{\mathcal{E}^c}^*\|_{L_2},$$

where we use the estimate of $\|H^{-1}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (251)), $\|G^{-1}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (253)), $\|\Phi_{\leq p}/\sqrt{n}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Lemma A.11) and $\left\|\frac{\Phi_{\mathcal{E}^c}b_{\mathcal{E}^c}}{\sqrt{n}}\right\|_2 = O_{d,\mathbb{P}}(\log^{1/2}(d))\cdot\|f_{\mathcal{E}^c}^*\|_{L_2}$ (Proposition B.21).

The following proposition shows that each summand is small. We defer the proof to Appendix C.3.10.

Proposition C.50. *Let A_j for $j \in [m]$ be defined as one of the following matrices:*

$$I_n, \quad \frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}}, \quad \left\{\frac{\Phi_{\leq p}(nD)^{-s}\Phi_{\leq p}^\top}{n}\right\}_{s=0}^{c_G}.$$

Then for any $t \in \mathbb{N}$ the following estimate holds

$$\mathbb{E}\left[\left\|\frac{1}{n}b_{\mathcal{E}^c}^\top\Phi_{\mathcal{E}^c}^\top A_1 A_2\ldots A_m\Phi_{\leq p}(nD)^{-t}E_r\right\|_2^2\right] = O_{d,\mathbb{P}}(d^{-\delta-1})\cdot\|f_{\mathcal{E}^c}^*\|_{L_2}^2. \tag{270}$$

90

Next, using the hypercontractivity Assumption 4 we show that the estimate (270) yields an analogous bound uniformly for all $r \in [d]$. To this end, Assumption 3 imposes that f^* is a finite-degree polynomial. Consequently, we can see that $\langle b_{\mathcal{E}^c}, \phi_{\mathcal{E}^c} \rangle$ is also finite-degree. Then we can deduce that each norm in Eq. (270) is also finite-degree since each term is the product of finitely many finite-degree polynomials. Therefore, invoking Lemma D.70, for any $q_0 \geq 1$, there exists a constant $C_{q_0} > 0$ such that:

$$\mathbb{E}\left[\max_r \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_{i_1} A_{i_2} \dots A_{i_q} \Phi_{\leq p}(nD)^{-t} E_r \right\|_2^2 \right]$$

$$\leq C_{q_0} d^{1/q_0} \max_{r \in [d]} \mathbb{E}\left[\left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_{i_1} A_{i_2} \dots A_{i_q} \Phi_{\leq p}(nD)^{-t} E_r \right\|_2^2 \right]$$

$$= C_{q_0} d^{1/q_0} \cdot O_{d,\mathbb{P}}(d^{-\delta-1}) \cdot \|f^*_{\mathcal{E}^c}\|_{L_2}^2 ,$$

where the last equality follows from Eq. (270).

Using Markov's inequality, we have with probability as least $1 - 1/\log(d)$ the following holds:

$$\max_r \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_{i_1} A_{i_2} \dots A_{i_q} \Phi_{\leq p}(nD)^{-t} E_r \right\|_2^2 = O_{d,\mathbb{P}}(d^{-1-\delta+1/q_0} \log(d)) \cdot \|f^*_{\mathcal{E}^c}\|_{L_2}^2 .$$

Plugging this bound into Eq. (269) and using Cauchy-Schwartz inequality yields

$$\max_r \left\| b_{\mathcal{E}^c}^\top \frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}} H^{-1} \frac{\Phi_{\leq p}}{\sqrt{n}} G^{-1} E_r \right\|_2^2 = \left(O_{d,\mathbb{P}}(d^{-1-\delta+1/q_0} \log(d)) + O_{d,\mathbb{P}}(d^{-1}) \right) \cdot \|f^*_{\mathcal{E}^c}\|_{L_2}^2$$

$$= O_{d,\mathbb{P}}(d^{-1}) \cdot \|f^*_{\mathcal{E}^c}\|_{L_2}^2 ,$$

where we let $q_0 > 1/\delta$. Plugging this bound into Eq. (263) and applying the AM-GM inequality yields

$$\left\| \beta_{\mathcal{E}^c}^\top \Phi_{\leq p} R \right\|_2^2 \leq 2 \left\| b_{\mathcal{E}^c}^\top \frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}} H^{-1} \frac{\Phi_{\leq p}}{\sqrt{n}} G^{-1} E_r \right\|_2^2 + 2 \|\Delta_1\|_2^2 = O_{d,\mathbb{P}}(d^{-1}) \cdot \|f^*_{\mathcal{E}^c}\|_{L_2}^2 ,$$

which completes the proof.

C.3.6 Proof of Proposition C.46

Similar to Eq. (229), we compute the expectation of the norm square over ε and apply trace inequality then we obtain

$$\mathbb{E}_{\varepsilon}\left[\left\| \beta_\varepsilon^\top \Phi_{\leq p} R \right\|_2^2 \right] \leq \sigma_\varepsilon^2 \cdot \left\| K_\lambda^{-1} \Phi_{\leq p} D \right\|_{op}^2 \cdot \mathrm{Tr}\left(D^{-1} R R^\top D^{-1} \right). \tag{271}$$

For the term $\left\| K_\lambda^{-1} \Phi_{\leq p} D \right\|_{op}^2$, Eq. (249) implies

$$K_\lambda^{-1} \Phi_{\leq p} D = \frac{1}{n} H^{-1} \Phi_{\leq p} G^{-1}.$$

With $\left\| H^{-1} \right\|_{op}, \left\| G^{-1} \right\|_{op} = O_{d,\mathbb{P}}(1)$ established in Eq. (251) and (253) and $\|\Phi_{\leq p}\|_{op} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11), we conclude

$$\left\| K_\lambda^{-1} \Phi_{\leq p} D \right\|_{op}^2 = \left\| \frac{1}{n} H^{-1} \Phi_{\leq p} G \right\|_{op}^2 = O_{d,\mathbb{P}}(n^{-1}). \tag{272}$$

With Claim 15, following the arguments in the proof of Proposition C.41, we can show

$$\mathrm{Tr}\left(D^{-1} R R^\top D^{-1} \right) = O_d(\|E_r\|_F^2) = O_d(d^{p-1}). \tag{273}$$

Plugging Eq. (272) and (273) into Eq. (271) and applying Markov's inequality completes the proof.

C.3.7 Proof of Proposition C.47

The proof will mainly use the machinery developed in the proof for Proposition C.45. Applying the same decomposition of K_λ as Eq. (259), we can write

$$K_\lambda = K_p + (\rho + \lambda)H,$$

where H is defined in Eq. (260). Invoking the Sherman-Morrison-Woodbury formula for K_λ^{-1}, we obtain

$$K_\lambda^{-1} = (\rho + \lambda)^{-1}H^{-1} - \frac{1}{n}(\rho + \lambda)^{-1}H^{-1}\Phi_{\leq p}^\top G^{-1}\Phi_{\leq p}H^{-1}, \tag{274}$$

where

$$G := (\rho + \lambda)(nD)^{-1} + \frac{\Phi_{\leq p}^\top H^{-1}\Phi_{\leq p}}{n}.$$

Plugging the approximations of H^{-1} (Eq. (267)) and G^{-1} (Eq. (268)) into Eq. (274) and expanding the expression we can write it as the finite summation similar to Eq. (269):

$$\frac{1}{n}b_\mathcal{Q}^\top\Phi_\mathcal{Q}^\top K_\lambda^{-1}\Phi_{\mathcal{S}_{p+1}^1} = \sum_{i=1}^{c_p} C_i b_\mathcal{Q}^\top\frac{\Phi_\mathcal{Q}^\top}{\sqrt{n}}A_{i_1}A_{i_2}\ldots A_{i_m}\frac{\Phi_{\mathcal{S}_{p+1}^1}}{\sqrt{n}} + \Delta, \tag{275}$$

where $A_{i_j} \in \left\{I_n, \frac{\Phi_{\mathcal{S}_{p+1}}\Phi_{\mathcal{S}_{p+1}}^\top}{d^{p+1}}\right\} \cup \left\{\frac{\Phi_{\leq p}(nD)^{-s}\Phi_{\leq p}^\top}{n}\right\}_{s\geq 0}$, $C_i > 0$ is a constant, $c_p, m \in \mathbb{N}$, and Δ satisfies $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \|b_\mathcal{Q}\|_2$.

Note that the effective degree of the feature in $\Phi_{\mathcal{S}_{p+1}^1}$ is bounded by p since $\Phi_{\mathcal{S}_{p+1}^1} = \Phi_{\mathcal{S}_p^0} \cdot X_r$, following an almost identical argument with the proof of Proposition C.50, we can show each summand satisfies:

$$\mathbb{E}\left[\left\|\frac{1}{n}b_\mathcal{Q}^\top\Phi_\mathcal{Q}^\top A_{i_1}A_{i_2}\ldots A_{i_m}\Phi_{\mathcal{S}_{p+1}^1}\right\|_2^2\right] = O_{d,\mathbb{P}}(d^{-\delta}) \cdot \|b_\mathcal{Q}\|_2^2.$$

Following a similar analysis as in the proof of Proposition C.45, we use Lemma D.70 and Markov's inequality together to obtain the following bound, which holds uniformly for all $r \in [d]$:

$$\left\|\frac{1}{n}b_\mathcal{Q}^\top\Phi_\mathcal{Q}^\top A_{i_1}A_{i_2}\ldots A_{i_m}\Phi_{\mathcal{S}_{p+1}^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{-\delta+1/q_0}) \cdot \log(d) \cdot \|b_\mathcal{Q}\|_2^2,$$

where $q_0 > 0$ is an arbitrarily large constant.

Combining with Δ in Eq. (275) and applying the triangle inequality we obtain the bound

$$\left\|\frac{1}{n}b_\mathcal{Q}^\top\Phi_\mathcal{Q}^\top K_\lambda^{-1}\Phi_{\mathcal{S}_{p+1}^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{-\delta+1/q_0}) \cdot \log(d) \cdot \|b_\mathcal{Q}\|_2^2.$$

Multiplying both sides of the equation by $(n/d^{p+1})^2$ yields

$$\left\|\frac{1}{d^{p+1}}\beta_\mathcal{Q}^\top\Phi_{\mathcal{S}_{p+1}^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{\delta-2+1/q_0})\log(d) \cdot \|b_\mathcal{Q}\|_2^2 = O_{d,\mathbb{P}}(d^{-1}) \cdot \|b_\mathcal{Q}\|_2^2, \tag{276}$$

where the last equality holds by letting $q_0 > 1/\delta$.

C.3.8 Proof of Proposition C.48

Note that Eq. (248) gives the decomposition

$$K_\lambda = \Phi_{\leq p}D\Phi_{\leq p}^\top + (\rho + \lambda)H, \tag{277}$$

where $H = I_n + (\rho + \lambda)^{-1}\Delta_1$, $\rho = g(1) - g_p(1)$ and Δ_1 satisfies $\|\Delta_1\|_{\text{op}} = O_{d,\mathbb{P}}(d^{(\delta-1)/2+\epsilon})$.

Invoking Sherman-Morrison-Woodbury formula to compute the inverse of K_λ and we obtain:

$$K_\lambda^{-1} = (\rho + \lambda)^{-1}H^{-1} - (\rho + \lambda)^{-1}\frac{1}{n}H^{-1}\Phi_{\leq p}G^{-1}\Phi_{\leq p}^\top H^{-1}.$$

Invoking Eq. (251) and (253) that H^{-1} and G^{-1} are close to I, we can show

$$\left\| K_\lambda^{-1} - (\rho + \lambda)^{-1}\left(I_n - \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\right) \right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}),$$

where we use the estimate $\|\Phi_{\leq p}\|_{\text{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11).

Plugging the above bound into left side of Eq. (244) and applying triangle inequality yields

$$\left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top K_\lambda^{-1}\Phi_{\mathcal{S}_{p+1}^1} - \frac{1}{n(\rho+\lambda)}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top\left(I_n - \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\right)\Phi_{\mathcal{S}_{p+1}^1} \right\|_{\text{op}}$$
$$= O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon} + d^{(\delta-1)/2+\epsilon}) \cdot \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2. \tag{278}$$

Now we consider $\frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top\left(I_n - \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\right)\Phi_{\mathcal{S}_{p+1}^1}$. Using the triangle inequality gives

$$\left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top\left(I_n - \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\right)\Phi_{\mathcal{S}_{p+1}^1} - b_{\mathcal{S}_{p+1}^1}^\top \right\|_2$$
$$\leq \left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1} - b_{\mathcal{S}_{p+1}^1}^\top \right\| + \left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\Phi_{\mathcal{S}_{p+1}^1} \right\|_2. \tag{279}$$

Next we show that both terms on the right side of the above inequality are small.

For the first term, note that Theorem A.13 immediately gives

$$\left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1} - b_{\mathcal{S}_{p+1}^1}^\top \right\| \leq \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2 \left\| \frac{1}{n}\Phi_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1} - I_n \right\|_{\text{op}} = \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon}). \tag{280}$$

We invoke Lemma D.71 to bound the expectation of the second term, i.e.,

$$\mathbb{E}\left[\left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\Phi_{\mathcal{S}_{p+1}^1} \right\|_2^2\right].$$

Specifically, we let non-overlapping sets $\mathcal{T}_1 = \{r\}$ and $\mathcal{T}_2 = [d] \setminus \{r\}$ with $s_1 = 0$ and $s_2 = 1$. Corresponding to the partition of the coordinate set, we can decompose the index of features such that we have $S \in 2^{\mathcal{T}_1} \times 2^{\mathcal{T}_2}$ for any feature index S that appears in the product. The effective degree of the feature in $\Phi_{\mathcal{S}_{p+1}^1}$ and $\Phi_{\leq p}$ is both bounded by $s_1 \cdot 1 + s_2 \cdot p = p$. Therefore, Lemma D.71 shows the expectation is bounded by

$$\left\| b_{\mathcal{S}_{p+1}^1} \right\|_2^2 \cdot O_d(d^p/n) = \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2^2 \cdot O_d(d^{-\delta}).$$

Following a similar analysis as in the proof of Proposition C.45, we use Lemma D.70 and Markov's inequality together to obtain the following bound, which holds uniformly for all $r \in [d]$:

$$\left\| \frac{1}{n}b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \frac{\Phi_{\leq p}\Phi_{\leq p}^\top}{n}\Phi_{\mathcal{S}_{p+1}^1} \right\|_2^2 = \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2^2 \cdot O_d(d^{-\delta+1/q_0})\log(d),$$

where $q_0 > 0$ is an arbitrarily large constant.

Plugging this bound and Eq. (279) into Eq. (280) yields:

$$\left\| \frac{1}{n} b_{\mathcal{S}_{p+1}^1}^\top \Phi_{\mathcal{S}_{p+1}^1}^\top \left(I_n - \frac{\Phi_{\leq p} \Phi_{\leq p}^\top}{n} \right) \Phi_{\mathcal{S}_{p+1}^1} - b_{\mathcal{S}_{p+1}^1}^\top \right\|_2 = \left\| b_{\mathcal{S}_{p+1}^1} \right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon}),$$

where $\epsilon > 0$ is an arbitrarily small constant and we absorb factor $\log(d)$ into d^ϵ.

Finally, combining this bound with Eq. (278) completes the proof.

C.3.9 Proof of Proposition C.49

Taking the expectation of the norm square with respect to the noise ε yields

$$\mathbb{E}\left[\left\| d^{-p-1} \beta_\varepsilon^\top \Phi_{\mathcal{S}_{p+1}^1} \right\|_2^2 \right] = d^{-2p-2} \operatorname{Tr}\left(K_\lambda^{-1} \Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top K_\lambda^{-1} \right). \tag{281}$$

Applying the inequality $|\operatorname{Tr}(AB)| \leq \operatorname{Tr}(A) \|B\|_{\mathrm{op}}$ yields

$$\operatorname{Tr}\left(K_\lambda^{-1} \Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top K_\lambda^{-1} \right) = \operatorname{Tr}\left(K_\lambda^{-2} \Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top \right) \leq \left\| K_\lambda^{-2} \right\|_{\mathrm{op}} \operatorname{Tr}\left(\Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top \right). \tag{282}$$

The bound $\left\| K_\lambda^{-2} \right\|_{\mathrm{op}} = \left\| K_\lambda^{-1} \right\|_{\mathrm{op}}^2 = O_{d,\mathbb{P}}(1)$ is implied by the expression (65). Next we analyze $\operatorname{Tr}\left(\Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top \right)$.

Since $|\mathcal{S}_{p+1}^1| = O_d(d^p)$, following Theorem A.12, we have $\left\| \Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top / n - I_{|\mathcal{S}_{p+1}^1|} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\delta/2+\epsilon})$. Then, we can conclude

$$\operatorname{Tr}\left(\frac{1}{n} \Phi_{\mathcal{S}_{p+1}^1} \Phi_{\mathcal{S}_{p+1}^1}^\top \right) \leq O_d(d^p)(1 + O_{d,\mathbb{P}} \cdot (d^{-\delta/2+\epsilon})) = O_{d,\mathbb{P}}(d^p).$$

Together with Eq. (282), the right side of Eq. (281) is bounded by $O_{d,\mathbb{P}}(d^{-2+\delta})$. Applying Markov's inequality we have $\left\| d^{-p-1} \beta_\varepsilon^\top \Phi_{\mathcal{S}_{p+1}^1} \right\|_2^2 = O_{d,\mathbb{P}}(d^{-2+\delta}) \log(d)$ with probability at least $1 - 1/\log(d)$. Absorbing the logarithmic factor into ϵ completes the proof.

C.3.10 Proof of Proposition C.50

Note that $(nD)^{-t}$ is diagonal matrix that satisfies $(nD)^{-t} \lesssim I$ and E_r is a diagonal matrix. We can upper bound the norm by removing $(nD)^{-t}$:

$$\left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_1 A_2 \ldots A_m \Phi_{\leq p}(nD)^{-t} E_r \right\|_2^2$$

$$= \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_1 A_2 \ldots A_m \Phi_{\leq p} E_r (nD)^{-t} \right\|_2^2$$

$$\lesssim O_d(1) \cdot \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_1 A_2 \ldots A_m \Phi_{\mathcal{E}} \right\|_2^2$$

where we use the fact $(E_r)_{S,S} = \mathbb{1}\{S \in \mathcal{E}\}$. Clearly, the estimate holds:

$$\mathbb{E}\left[\left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_1 A_2 \ldots A_m \Phi_{\mathcal{E}} \right\|_2^2 \right] \leq \|b_{\mathcal{E}^c}\|_2^2 \cdot \left\| \mathbb{E}\left[\underbrace{\frac{1}{n^2} \Phi_{\mathcal{E}^c}^\top A_1 \ldots A_m \Phi_{\mathcal{E}} \Phi_{\mathcal{E}}^\top A_m \ldots A_1 \Phi_{\mathcal{E}^c}}_{=:K} \right] \right\|_{\mathrm{op}}. \tag{283}$$

94

We now aim to upper bound the operator norm of $\mathbb{E}[K]$. To this end, we decompose the feature matrices in K so as to ensure that distinct matrices in the product correspond to nonoverlapping index sets.

Following Eq. (69), we have the decomposition

$$\mathcal{S}_k = \mathcal{S}_k^1 \sqcup \mathcal{S}_k^0. \tag{284}$$

Clearly, the following equalities hold:

$$\mathcal{E} = \bigcup_{k=1}^{p} \mathcal{S}_k^1, \quad \mathcal{E}^c = \left(\bigcup_{k=0}^{p} \mathcal{S}_k^0 \right) \bigcup \left(\bigcup_{k=p+1}^{\ell} \mathcal{S}_k \right).$$

Define now the collection of matrices

$$\mathcal{U} := \{\Phi_{\mathcal{S}_k^1}\}_{k=1}^{p} \bigcup \{\Phi_{\mathcal{S}_k^0}\}_{k=0}^{p} \bigcup \{\Phi_{\mathcal{S}_k}\}_{k=p+1}^{\ell}.$$

Note that any two distinct matrices in $\mathcal{U}$ correspond to distinct index sets. Normalizing the matrices, we further define the collection

$$\widetilde{\mathcal{U}} := \{n^{-\frac{1}{2}}(nD)^{-s/2}\Phi_{\mathcal{S}_k^1}\}_{k\in[p],s\geq 0} \bigcup \{n^{-\frac{1}{2}}(nD)^{-s/2}\Phi_{\mathcal{S}_k^0}\}_{0\leq k\leq p,s\geq 0} \bigcup \{d^{-k/2}\Phi_{\mathcal{S}_k}\}_{p+1\leq k\leq \ell}.$$

Let us now write the two expressions:

$$\frac{\Phi_{\leq p}(nD)^{-s}\Phi_{\leq p}^{\top}}{n} = \sum_{k=0}^{p} \frac{\Phi_{\mathcal{S}_k^0}(nD)^{-s}\Phi_{\mathcal{S}_k^0}^{\top}}{n} + \sum_{k=1}^{p} \frac{\Phi_{\mathcal{S}_k^1}(nD)^{-s}\Phi_{\mathcal{S}_k^1}^{\top}}{n};$$

$$\frac{\Phi_{\mathcal{E}}\Phi_{\mathcal{E}}^{\top}}{n} = \sum_{k=1}^{p} \frac{\Phi_{\mathcal{S}_k^1}\Phi_{\mathcal{S}_k^1}^{\top}}{n}.$$

Plugging these decompositions for the corresponding A_i in the expression of K Eq. (283), we can write K as the summation of finitely many terms, where each term takes the form of some positive constant multiplied by

$$K_j := \frac{\max(n^{1/2}, d^{k_1/2}) \cdot \max(n^{1/2}, d^{k_{2m+1}/2})}{n} \cdot B_1^{\top}\left(B_2 B_2^{\top}\right) \ldots \left(B_{2m} B_{2m}^{\top}\right) B_{2m+1},$$

where $B_1 \in \mathbb{R}^{n \times k_1}, \ldots, B_{2m+1} \in \mathbb{R}^{n \times k_{2m+1}}$ are from $\widetilde{\mathcal{U}}$.

Particularity note that B_1 and B_{2m+1} corresponding to the decomposition of $\Phi_{\mathcal{E}^c}$ and thus each takes the form $d^{-k/2}\Phi_{\mathcal{S}_k}$ for some $p+1 \leq k \leq \ell$; B_{m+1} corresponds to the decomposition of $\Phi_{\mathcal{E}}$ and takes the form $n^{-\frac{1}{2}}\Phi_{\mathcal{S}_k^1}$ for some $k \in [p]$. The equation contains an additional scaling factor on the right side, which arises from the distinct scaling of B_1 and B_{2m+1} by a factor of $1/n$ in Eq. (283).

Next, we invoke Lemma D.71 to bound each $\|\mathbb{E}[K_j]\|_{\mathrm{op}}$. Specifically, we let non-overlapping sets $\mathcal{T}_1 = \{r\}$ and $\mathcal{T}_2 = [d] \setminus \{r\}$ with $s_1 = 0$ and $s_2 = 1$, since $|\mathcal{T}_1| = O_d(d^0)$ and $|\mathcal{T}_2| = O_d(d^1)$. Corresponding to this partition of the coordinate set, we can decompose the index of features such that we have $S = S_1 \sqcup S_2 \in 2^{\mathcal{T}_1} \times 2^{\mathcal{T}_2}$ for any feature index S that appears in the product. The effective degree of the feature of B_{m+1} is bounded by $s_1 \cdot 1 + s_2 \cdot (p-1) = p$ since x_r appears in all features, namely $\Phi_{\mathcal{S}_p^1} = X_r \cdot \Phi_{\mathcal{S}_{p-1}^0}$. Consequently, we yield

$$\|\mathbb{E}[K_j]\|_{\mathrm{op}} = \frac{\max(n^{1/2}, d^{k_1/2}) \cdot \max(n^{1/2}, d^{k_{2m+1}/2})}{n} \cdot \left\|\mathbb{E}\left[B_1^{\top}\left(B_2 B_2^{\top}\right) \ldots \left(B_{2m} B_{2m}^{\top}\right) B_{2m+1}\right]\right\|_{\mathrm{op}}$$

$$= \frac{\max(n^{1/2}, d^{k_1/2}) \cdot \max(n^{1/2}, d^{k_{2m+1}/2})}{n} \cdot O_d(d^{p-1}) \cdot \min(n^{-1/2}, d^{-k_1/2}) \cdot \min(n^{-1/2}, d^{-k_{2m+1}/2})$$

$$= O_d(d^{p-1}/n) = O_d(d^{-1-\delta}). \tag{285}$$

Applying triangle inequality yields $\|\mathbb{E}[K]\|_{\mathrm{op}} = O_d(d^{-1-\delta})$. Combining with Eq. (283) completes the proof.

C.4 Proof of Propositions for Theorem B.27 (Gradient error)

C.4.1 Proof of Lemma B.29

We similarly define off-diagnoal entries of $\frac{1}{d^{(s+1-\lambda)^\top m}} \Phi_{\mathcal{S}_m}(Z)(\Phi_{\mathcal{S}_m}(Z))^\top$ by

$$\Delta_{i,j} := \begin{cases} \frac{1}{d^{(s+1-\lambda)^\top m}} \left\langle \phi_{\mathcal{S}_m}(z^{(i)}), \phi_{\mathcal{S}_m}(z^{(j)}) \right\rangle, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

We omit the argument Z for simplicity. Then we split $\frac{1}{d^{(s+1-\lambda)^\top m}} \Phi_{\mathcal{S}_m}(\Phi_{\mathcal{S}_m}^\top)$ into the diagonal and off-diagonal parts. Namely, the triangle inequality yields

$$\mathbb{E}\left[\left\| \frac{1}{d^{(s+1-\lambda)^\top m}} \Phi_{\mathcal{S}_m}(\Phi_{\mathcal{S}_m})^\top - \mu_{\mathcal{S}_m} I_n \right\|_{\text{op}}\right] \tag{286}$$

$$\leq \mathbb{E}\left[\left\| \frac{1}{d^{(s+1-\lambda)^\top m}} \mathsf{Diag}(\Phi_{\mathcal{S}_m}(\Phi_{\mathcal{S}_m})^\top) - \mu_{\mathcal{S}_m} I_n \right\|_{\text{op}}\right] + \mathbb{E}\left[\|\Delta\|_{\text{op}}\right], \tag{287}$$

where

$$\mu_{\mathcal{S}_m} := \frac{1}{d^{(s+1-\lambda)^\top m}} \mathbb{E}\left[\|\phi_{\mathcal{S}_m}\|_2^2\right].$$

Now we show $\mu_{\mathcal{S}_m} = \Theta_d(1)$. Since $\mathcal{T}_1, \ldots, \mathcal{T}_k$ do not overlap, we can write the norm of $\phi_{\mathcal{S}_m}$ as the product of the norm corresponding to each $\mathcal{T}_k$:

$$\|\phi_{\mathcal{S}_m}\|_2^2 = \prod_{k=1}^N \left\|\phi_{\mathcal{T}_k^{m_k}}\right\|_2^2, \tag{288}$$

where we denote $\mathcal{T}_k^{m_k} := \{T_k \mid T_k \in 2^{\mathcal{T}_k} \text{ and } |T_k| = m_k\}$.

Recall that for any $r \in \mathcal{T}_k$, we have $w_r = \Theta_d(d^{\gamma_k}) = \Theta_d(d^{1-\lambda_k})$. Using simple combinatorial arguments we can infer

$$\left\|\phi_{\mathcal{T}_k^{m_k}}\right\|_2^2 = \sum_{\substack{1 \leq r_1 < r_2 <, \ldots, < r_{m_k} \\ r_1, \ldots, r_{m_k} \in \mathcal{T}_k}} w_{r_1} \ldots w_{r_{m_k}} = \Theta_d(d^{(1-\lambda_k)m_k}) \cdot \frac{1}{m_k!} \cdot \Theta_d(d^{s_k m_k}) = \frac{1}{m_k!} \cdot \Theta_d(d^{(s_k+1-\lambda_k)m_k}).$$

Therefore, plugging each factor into Eq. (288) yields

$$\|\phi_{\mathcal{S}_m}\|_2^2 = \frac{1}{m!} \cdot \Theta_d(d^{(s+1-\lambda)^\top m}), \tag{289}$$

which implies $\mu_{\mathcal{S}_m} = \Theta_d(d^{-(s+1-\lambda)^\top m}) \|\phi_{\mathcal{S}_m}\|_2^2 = \Theta_d(1)$.

Bounding $\mathbb{E}\left[\left\| \frac{1}{d^{(s+1-\lambda)^\top m}} \mathsf{diag}(\Phi_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}^\top) - \mu_{\mathcal{S}_m} I_n \right\|_{\text{op}}\right]$. Note that $\|\phi_{\mathcal{S}_m}\|_2^2 = \mathbb{E}\left[\|\phi_{\mathcal{S}_m}\|_2^2\right]$ holds. Therefore, we directly have $\frac{1}{d^{(s+1-\lambda)^\top m}} \mathsf{diag}(\Phi_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}^\top) = \mu_{\mathcal{S}_m} I_n$ by definition of $\mu_{\mathcal{S}_m}$ hence we can conclude a stronger result

$$\left\| \frac{1}{d^{(s+1-\lambda)^\top m}} \mathsf{Diag}(\Phi_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}^\top) - \mu_{\mathcal{S}_m} I_n \right\|_{\text{op}} = 0. \tag{290}$$

Bound $\mathbb{E}\left[\|\Delta\|_{\text{op}}\right]$. We follow the steps in the proof of Lemma A.13. Specifically, we use the same bound as Eq. (174).

$$\mathbb{E}\left[\|\Delta\|_{\text{op}}\right] \leq 4n^{1/2} \sup_{T \subseteq [n]} (\mathbb{E}_{T^c}[\Sigma(T)])^{1/2} + 4c\log^{1/2}(n) \cdot \sup_{T \subseteq [n]} (\mathbb{E}_{T^c}[\Gamma(T)])^{1/2}. \tag{291}$$

Step 1: Bound $\mathbb{E}_{T^c}[\Sigma(T)]$. Note that each entry of $\Delta_{T^c,j}\Delta_{T^c,j}^\top$ takes the form $\phi_i^\top \phi_j \phi_j^\top \phi_l/d^{2(s+1-\lambda)^\top m}$ for $i,l \in T^c, j \in T$. Since $\mathbb{E}_{z^{(j)}}[\phi_j\phi_j^\top] = I_{|\mathcal{S}_m|} \cdot \Theta_d(d^{(1-\lambda)^\top m})$, we have each entry of $\mathbb{E}_{z^{(j)}}[\Delta_{T^c,j}\Delta_{T^c,j}^\top]$ takes the form $\phi_i^\top \phi_l \cdot \Theta_d(d^{(\lambda-1-2s)^\top m})$. Therefore, we deduce

$$
\begin{aligned}
\Sigma(T) &= \left\| \mathbb{E}_{z^{(j)}}\left[\Delta_{T^c,j}\Delta_{T^c,j}^\top\right] \right\|_{\mathrm{op}} \\
&= \left\| \left(\Phi\Phi^\top\right)_{T^c,T^c} \right\|_{\mathrm{op}} \cdot \Theta_d(d^{(\lambda-1-2s)^\top m}) \\
&\leq \mathbb{E}_T\left[\left\|\Phi\Phi^\top\right\|_{\mathrm{op}}\right] \cdot \Theta_d(d^{(\lambda-1-2s)^\top m}),
\end{aligned}
$$

where in the last inequality, we use Cauchy's interlacing theorem [41]. Specifically, the theorem shows that adding a row and a column to a Hermitian matrix will only increase its operator norm.

With a similar analysis, corresponding to Eq. (175), we have

$$
\mathbb{E}_{T^c}[\Sigma(T)] \leq \Theta_d(d^{-s^\top m}) \cdot \mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] + \Theta_d(d^{(\lambda-1-2s)^\top m}) \cdot \mathbb{E}\left[\max_{j\in[n]}\left\|\phi_{\mathcal{S}_m}(z^{(j)})\right\|^2\right]. \tag{292}
$$

With Eq. (289), there exists a constant $C_\Sigma > 0$ such that the following holds

$$
\mathbb{E}_{T^c}[\Sigma(T)] \leq C_\Sigma d^{-s^\top m} \cdot \mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] + C_\Sigma d^{-s^\top m}. \tag{293}
$$

Step 2: Bound $\mathbb{E}_{T^c}[\Gamma(T)]$. With a similar analysis, corresponding to Eq. (179) and (180), we can obtain for any $q \geq 1$,

$$
\mathbb{E}_{T^c}\mathbb{E}_T\left[\max_{i\in T}\|\Delta_{i,T^c}\|^2\right] = \mathbb{E}\left[\max_{i\in T}\|\Delta_{i,T^c}\|^2\right] \leq n^{1+2/q}C_{2l,2q}^2\mathbb{E}\left[\Delta_{i,j}^{2q}\right].
$$

Since $\Delta_{i,j}$ is polynomial with degree at most $2\widetilde{p}_\gamma$, the hypercontractivity Assumption 4 implies

$$
\mathbb{E}\left[\Delta_{i,j}^{2q}\right]^{1/q} \leq C_{2\widetilde{p}_\gamma,2q}^2\mathbb{E}\left[\Delta_{i,j}^2\right]. \tag{294}
$$

Note that $|\phi_u|^2 \leq c_0 d^{(1-\lambda)^\top m}$ holds for all $u \in \mathcal{S}_m$ where $c_0 > 0$ is an absolute constant. Then we can show

$$
\mathbb{E}\left[\Delta_{i,j}^2\right] = \sum_{u\in\mathcal{S}_m}\frac{1}{d^{2(s+1-\lambda)^\top m}}|\phi_u|^4 \leq |\mathcal{S}_m| \cdot c_0^2 d^{-2s^\top m} \leq c_0^2 d^{-s^\top m}
$$

where we use the inequality $|\mathcal{S}_m| \leq d^{s^\top m}$.

Therefore, we obtain a similar bound to Eq. (181):

$$
\mathbb{E}_{T^c}[\Gamma(T)] \leq c_0^2 C_{2\widetilde{p}_\gamma,2q}^2 n^{1+2/q}d^{-s^\top m}. \tag{295}
$$

Combining Eq.(292) and Eq. (295) with Eq. (291) we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] \\
&\leq 4n^{1/2}\left(\frac{C_\Sigma}{d^{s^\top m}}\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right] + \frac{C_\Sigma}{d^{s^\top m}}\right)^{1/2} + 4c\log^{1/2}(n)\left(c_0^2 C_{2\widetilde{p}_\gamma,2q}^2 n^{1+2/q}d^{-s^\top m}\right)^{1/2} \\
&\leq 4C_\Sigma^{1/2}\left(\frac{n}{d^{s^\top m}}\right)^{1/2}\left(\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]\right)^{1/2} + 4C_\Sigma^{1/2}\frac{n^{1/2}}{d^{\frac{1}{2}s^\top m}} + 4cc_0 C_{2\widetilde{p}_\gamma,2q} \cdot \frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{\frac{1}{2}}(n)}{d^{s^\top m/2}},
\end{aligned} \tag{296}
$$

where the last inequality is obtained from the elementary estimate $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a,b \geq 0$.

Solving the quadratic inequality with respect to $\sqrt{\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]}$ we obtain:

$$\sqrt{\mathbb{E}\left[\|\Delta\|_{\mathrm{op}}\right]} \leq 2C_\Sigma^{1/2}\left(\frac{n}{d^{s^\top m}}\right)^{1/2} + 2\sqrt{\frac{n}{d^{s^\top m}} + \frac{n^{1/2}}{d^{\frac{1}{2}s^\top m}} + cc_0 C_{2\widetilde{p}_\gamma,2q} \cdot \frac{n^{\frac{1}{2}+\frac{1}{q}}\log^{\frac{1}{2}}(n)}{d^{s^\top m/2}}}$$

Squaring both sides and using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and taking into account that $n = d^{q+\delta}$, we can conclude that the right side is bounded by a constant multiple of

$$\max\left(d^{p+\delta-s^\top m}, \sqrt{\log(n)} \cdot d^{(p+\delta-s^\top m)/2+\epsilon}\right), \tag{297}$$

for any $\epsilon > 0$.

Shrinking ϵ to remove logarithmic factors and applying Markov's inequality completes the proof.

C.4.2   Proof of Theorem B.28

The proof follows the proof of Theorem A.14. Similarly, we first apply Lemma D.66 and we deduce that for any $s \in (0,1)$, for $i \neq j$, the estimate

$$\left|\frac{1}{d}\left\langle\sqrt{w}\odot x^{(i)}, \sqrt{w}\odot x^{(j)}\right\rangle - \delta_{ij}\right| \leq s, \tag{298}$$

holds with probability at least $1 - 2n^2\exp\left(-s^2 d/(2\|w\|_\infty)\right)$. When $i = j$, we have

$$\left\langle\sqrt{w}\odot x^{(i)}, \sqrt{w}\odot x^{(i)}\right\rangle = d. \tag{299}$$

We assume that the events (298) occurs for the remainder of the proof.

To simplify the notation, we denote $\bar{c} = g_{w,\widetilde{p}}(1) = \sum_{k=0}^{\widetilde{p}}\frac{g_w^{(k)}(0)}{k!}$ and we define the matrix $E_w(X) := K_w(X,X) - K_{w,\widetilde{p}}(X,X)$. Then splitting $E_M(X)$ into its diagonal and off-diagonal parts, we obtain a bound similar to Eq. (185):

$$\|E_w(X) - (g_w(1)-\bar{c})I_n\|_{\mathrm{op}} \leq \|E_w(X) - \mathsf{Diag}(E_w(X))\|_F + \max_{i=1,\ldots,n}|E_w(X)_{ii} - (g(1)-\bar{c})I_n|. \tag{300}$$

For off-diagonal entries, similar to Eq. (189), assuming $s < \min(1/2,\epsilon)$ we estimate

$$\|E_w(X) - \mathsf{Diag}(E_w(X))\|_F^2 \leq \frac{4n^2\sup_{k\geq\widetilde{p}+1}g^{(k)}(0)}{3(\widetilde{p}+1)!}s^{2\widetilde{p}+2}. \tag{301}$$

For diagonal entries, invoking Eq. (299), we immediately have

$$|E_w(X)_{ii} - (g_w(1)-\bar{c})I_n| = 0. \tag{302}$$

Then we conclude

$$\|E_w(X) - (g_w(1)-\bar{c})I_n\|_{\mathrm{op}} \leq \sqrt{\frac{4\sup_{k\geq\widetilde{p}+1}g^{(k)}(0)}{3(\widetilde{p}+1)!}}s^{\widetilde{p}+1}n. \tag{303}$$

Then for any $C > 2$, we may set $s := \sqrt{2C\log(n)\|w\|_\infty/d}$ thereby ensuring that the event (298) occurs with probability at least $1 - \frac{4}{n^{C-2}}$. Plugging these values into (303) yields the estimate

$$\|E_w(x) - (g(1)-\bar{c})I_n\|_{\mathrm{op}} \leq cd^{\frac{2p-\widetilde{p}-1}{2}+\delta}C^{\frac{\widetilde{p}+1}{2}}\|w\|_\infty^{\frac{\widetilde{p}+1}{2}}\log^{\frac{\widetilde{p}+1}{2}}(n),$$

where $c := \sqrt{\frac{4\sup_{k\geq\widetilde{p}+1}g^{(k)}(0)}{3(\widetilde{p}+1)!}}$ is a numerical constant.

To meet the assumption $s < \min(1/2,\epsilon)$, d needs to satisfy $\frac{d}{\|w\|_\infty\log(d)} > \frac{\min(1/2,\epsilon)^2 c_1}{C(p+\delta)}$. The proof is complete.

C.4.3 Proof of Lemma B.30

We will analyze the component $\Phi_{\mathcal{A}_{>p_\delta}}(Z)D_{\mathcal{A}_{>p_\delta}}\Phi_{\mathcal{A}_{>p_\delta}}(Z)^\top$ in the composition (113) therefore, in the proof, we consider $\boldsymbol{m}$ that satisfies $\|\boldsymbol{m}\|_1 \leq \widetilde{p}_{\gamma_{\max}}$.

We first present the proof for Eq. (114). We decompose the product as a summation based on $\boldsymbol{m}$:

$$\Phi_{\mathcal{A}_{>p_\delta}}(Z)D_{\mathcal{A}_{>p_\delta}}\Phi_{\mathcal{A}_{>p_\delta}}(Z)^\top = \sum_{\boldsymbol{\lambda}^\top\boldsymbol{m}>p+\delta} \Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)D_{\mathcal{S}_{\boldsymbol{m}}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top. \tag{304}$$

Since $\boldsymbol{s}^\top\boldsymbol{m} \leq \boldsymbol{\lambda}^\top\boldsymbol{m}$ and $\boldsymbol{s}^\top\boldsymbol{m}$ can be less than $p+\delta$, we split the set $\{\boldsymbol{m} \mid \boldsymbol{\lambda}^\top\boldsymbol{m} > p+\delta\}$ into two sets:

$$L_1 := \{\boldsymbol{m} \mid \boldsymbol{s}^\top\boldsymbol{m} \geq p+\delta\}, \quad L_2 := \{\boldsymbol{m} \mid \boldsymbol{s}^\top\boldsymbol{m} < p+\delta < \boldsymbol{\lambda}^\top\boldsymbol{m}\}.$$

For each $\boldsymbol{m} \in L_1$, we can show

$$\left\|d^{(\boldsymbol{\lambda}-\boldsymbol{s})^\top\boldsymbol{m}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)D_{\mathcal{S}_{\boldsymbol{m}}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top - g^{(\|\boldsymbol{m}\|_1)}(0)\mu_{\mathcal{S}_{\boldsymbol{m}}}I\right\|_{\mathrm{op}}$$

$$\overset{(a)}{\leq} \left\|g^{(\|\boldsymbol{m}\|_1)}(0)d^{(\boldsymbol{\lambda}-\boldsymbol{1}-\boldsymbol{s})^\top\boldsymbol{m}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top - g^{(\|\boldsymbol{m}\|_1)}(0)\mu_{\mathcal{S}_{\boldsymbol{m}}}I\right\|_{\mathrm{op}}$$

$$+ O_d(d^{(\boldsymbol{\lambda}-\boldsymbol{1}-\boldsymbol{s})^\top\boldsymbol{m}-1}) \cdot \left\|\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top\right\|_{\mathrm{op}}$$

$$\overset{(b)}{=} O_{d,\mathbb{P}}(d^{(p+\delta-\boldsymbol{s}^\top\boldsymbol{m})/2+\epsilon}) + O_{d,\mathbb{P}}(d^{-1}). \tag{305}$$

where (a) follows from the definition of D (Lemma A.16) and (b) follows from Lemma B.29, which implies $\left\|\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\boldsymbol{s}+\boldsymbol{1}-\boldsymbol{\lambda})^\top\boldsymbol{m}})$.

For each $\boldsymbol{m} \in L_2$, by sub-multiplicativity of the spectral norm, we have

$$\left\|\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)D_{\mathcal{S}_{\boldsymbol{m}}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top\right\|_{\mathrm{op}} \leq \|\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)\|_{\mathrm{op}}^2 \cdot \|D_{\mathcal{S}_{\boldsymbol{m}}}\|_{\mathrm{op}}$$

$$= \|\Phi_{\mathcal{S}_{\boldsymbol{m}}}(X)\|_{\mathrm{op}}^2 \cdot \|W_{\mathcal{S}_{\boldsymbol{m}}}\|_{\mathrm{op}} \cdot \|D_{\mathcal{S}_{\boldsymbol{m}}}\|_{\mathrm{op}}$$

$$\overset{(a)}{=} O_{d,\mathbb{P}}(d^{p+\delta}) \cdot d^{-\boldsymbol{\lambda}^\top\boldsymbol{m}} = O_{d,\mathbb{P}}(d^{-\zeta_2}). \tag{306}$$

where (a) follows from Lemma A.11.

Consequently, multiplying both sides of Eq. (305) by $d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top\boldsymbol{m}}$ and plugging the resulted bound and Eq. (306) back into Eq. (304) yields

$$\left\|\sum_{\boldsymbol{\lambda}^\top\boldsymbol{m}\geq p+\delta} \Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)D_{\mathcal{S}_{\boldsymbol{m}}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top - \sum_{\boldsymbol{m}\in L_1} d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top\boldsymbol{m}}g^{(\|\boldsymbol{m}\|_1)}(0)\mu_{\mathcal{S}_{\boldsymbol{m}}}I - \sum_{\boldsymbol{m}\in L_2} \Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)D_{\mathcal{S}_{\boldsymbol{m}}}\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top\right\|_{\mathrm{op}}$$

$$\leq \sum_{\boldsymbol{m}\in L_1} O_{d,\mathbb{P}}(d^{(p+\delta-\boldsymbol{s}^\top\boldsymbol{m})/2+\epsilon} \cdot d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top\boldsymbol{m}}) + \sum_{\boldsymbol{m}\in L_2} O_{d,\mathbb{P}}(d^{-\zeta_2})$$

$$\leq \sum_{\boldsymbol{m}\in L_1} O_{d,\mathbb{P}}(d^{(p+\delta-\boldsymbol{s}^\top\boldsymbol{m})/2+\epsilon} \cdot d^{(\boldsymbol{s}-\boldsymbol{\lambda})^\top\boldsymbol{m}/2}) + \sum_{\boldsymbol{m}\in L_2} O_{d,\mathbb{P}}(d^{-\zeta_2})$$

$$= O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon} + d^{-\zeta_2}).$$

Combining with Eq. (113) completes the proof for Eq. (114).

Now we present the proof for Eq. (115). Following the definition of D (Lemma A.16), we have for any

m such that $\lambda^\top m > p + \delta$, the following holds

$$\left\| \mathrm{diag}\left(\Phi_{\mathcal{S}_m}(Z) D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}(Z)^\top \right) - g^{(\|m\|_1)}(0) d^{(s-\lambda)^\top m} \mu_{\mathcal{S}_m} \cdot I_n \right\|_{\mathrm{op}}$$

$$\overset{(a)}{\leq} O_d(d^{-\|m\|_1 - 1}) \cdot \left\| \mathrm{diag}\left(\Phi_{\mathcal{S}_m}(Z) \Phi_{\mathcal{S}_m}(Z)^\top \right) \right\|_{\mathrm{op}}$$

$$+ g^{(\|m\|_1)}(0) \left\| d^{-\|m\|_1} \mathrm{diag}\left(\Phi_{\mathcal{S}_m}(Z) \Phi_{\mathcal{S}_m}(Z)^\top \right) - d^{(s-\lambda)^\top m} \mu_{\mathcal{S}_m} \cdot I_n \right\|_{\mathrm{op}}$$

$$\overset{(b)}{=} O_{d,\mathbb{P}}(d^{(s-\lambda)^\top m - 1})$$

$$\overset{(c)}{=} O_{d,\mathbb{P}}(d^{-1}), \tag{307}$$

where (a) follows from the definition of D (Lemma A.16), (b) follows from Eq. (290) and (c) follows from the inequality $s^{[i]} \leq \lambda_i$.

We split those m that satisfies $\lambda^\top m > p + \delta$ into:

$$J_1 := \{ m \mid \lambda^\top m > p + \delta \ \text{and} \ s^\top m < p + 2 \}, \tag{308}$$

$$J_2 := \{ m \mid \lambda^\top m > p + \delta \ \text{and} \ s^\top m \geq p + 2 \}. \tag{309}$$

Then we can show the following estimate holds:

$$\left\| \Phi_{\mathcal{A}_{>p_\delta}}(Z) D_{\mathcal{A}_{>p_\delta}} \Phi_{\mathcal{A}_{>p_\delta}}(Z)^\top - \sum_{m \in J_1} \mathrm{off}\left(\Phi_{\mathcal{S}_m}(Z) D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}(Z)^\top \right) - \sum_{\lambda^\top m > p + \delta} d^{(s-\lambda)^\top m} g^{(\|m\|_1)}(0) \mu_{\mathcal{S}_m} I \right\|_{\mathrm{op}}$$

$$\overset{(a)}{\leq} \left\| \sum_{m \in J_2} \Phi_{\mathcal{S}_m}(Z) D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}(Z)^\top - \sum_{m \in J_2} d^{(s-\lambda)^\top m} g^{(\|m\|_1)}(0) \mu_{\mathcal{S}_m} I \right\|_{\mathrm{op}}$$

$$+ \sum_{m \in J_1} \left\| \mathrm{diag}\left(\Phi_{\mathcal{S}_m}(Z) D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}(z)^\top \right) - g^{(\|m\|_1)}(0) d^{(s-\lambda)^\top m} \mu_{\mathcal{S}_m} \cdot I_n \right\|_{\mathrm{op}}$$

$$\overset{(b)}{\leq} \sum_{m \in J_2} \left\| \Phi_{\mathcal{S}_m}(Z) D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}(Z)^\top - d^{(s-\lambda)^\top m} g^{(\|m\|_1)}(0) \mu_{\mathcal{S}_m} I \right\|_{\mathrm{op}} + O_{d,\mathbb{P}}(d^{-1})$$

$$\overset{(c)}{\leq} \sum_{m \in J_2} O_{d,\mathbb{P}}(d^{(p+\delta - s^\top m)/2 + \epsilon}) + O_{d,\mathbb{P}}(d^{-1})$$

$$= O_{d,\mathbb{P}}(d^{(\delta-2)/2 + \epsilon}) + O_{d,\mathbb{P}}(d^{-1})$$

$$= O_{d,\mathbb{P}}(d^{(\delta-2)/2 + \epsilon}),$$

where (a) is obtained by splitting $\Phi_{\mathcal{S}_m} D_{\mathcal{S}_m} \Phi_{\mathcal{S}_m}^\top$ into diagonal and off-diagonal components for $m \in J_1$, (b) follows from Eq. (307) and (c) follows from Eq. (305).

Combining the above bound with Eq. (113) completes the proof.

C.4.4 Proof of Proposition B.31

The proof is analogous to Proposition B.25, but we will provide special treatment for the case when $\lambda^\top m = p + \delta$.

Let $\mathcal{E}(r)$ denote the union of all sets $S \subset [d]$ containing r with $S \in \mathcal{A}_{\leq p_\delta}$:

$$\mathcal{E}(r) := \{ S \mid r \in S, \ S \in \mathcal{A}_{\leq p_\delta} \}. \tag{310}$$

Then $\mathcal{E}^c(r)$ is defined as the complement of $\mathcal{E}(r)$. Particularly note that the cardinality of the set $\mathcal{E}(r)$ is bounded by

$$O_d(d^{s^\top m - s_\iota}) = O_d(d^{\lambda^\top m - \lambda_\iota}) = O_d(d^{p+\delta - \lambda_\iota}), \tag{311}$$

where we use the fact $\lambda_j \geq s_j$ for all $j \in [N]$ and $m_\iota \geq 1$.

We define a partial identity matrix E_r of size $|\mathcal{A}_{\leq p_\delta}| \times |\mathcal{A}_{\leq p_\delta}|$ as

$$(E_r)_{S,S} = \mathbb{1}\{S \in \mathcal{E}(r)\}.$$

We additionally define a diagonal weight matrix

$$V := \left((\rho + \lambda)(n\widehat{D}_{\mathcal{A}_{\leq p_\delta}})^{-1} + I\right)^{-1}, \tag{312}$$

where each entry takes the form

$$V_{S,S} = \frac{1}{(\rho + \lambda)(nD_{S,S}W_{S,S})^{-1} + 1}. \tag{313}$$

We split the set $\mathcal{A}_{\leq p_\delta}$ into $\mathcal{A}_{< p_\delta}$ and $\mathcal{A}_{= p_\delta}$:

$$\mathcal{A}_{< p_\delta} := \bigcup_{\boldsymbol{\lambda}^\top \boldsymbol{m} < p + \delta} \mathcal{S}_{\boldsymbol{m}}, \quad \mathcal{A}_{= p_\delta} := \bigcup_{\boldsymbol{\lambda}^\top \boldsymbol{m} = p + \delta} \mathcal{S}_{\boldsymbol{m}}.$$

We now write $\beta = \beta_{\mathcal{E}(r)} + \beta_{\mathcal{E}^c(r)} + \beta_\varepsilon$ where we define:

$$\beta_{\mathcal{E}(r)} := K_{w,\lambda}^{-1}(\Phi_{\mathcal{E}(r)} b_{\mathcal{E}(r)}), \quad \beta_{\mathcal{E}^c(r)} := K_{w,\lambda}^{-1}(\Phi_{\mathcal{E}^c(r)} b_{\mathcal{E}^c(r)}), \quad \beta_\varepsilon = K_{w,\lambda}^{-1}\varepsilon.$$

With the decomposition of β, we accordingly bound the norm by

$$\left\|\beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R - b_{\mathcal{A}_{\leq p_\delta}}^\top V E_r\right\|_2 \tag{314}$$

$$\leq \left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R - b_{\mathcal{A}_{\leq p_\delta}}^\top V E_r\right\|_2 + \left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2 + \left\|\beta_\varepsilon^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|. \tag{315}$$

We will show all three terms on the right hand side of the above equation are small.

The following two Propositions control the first and the second term in Eq. (314) respectively. We defer the proof to Appendix C.4.6 and C.4.8 respectively.

Proposition C.51. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$\left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R - b_{\mathcal{A}_{\leq p_\delta}}^\top V E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\|\partial_r f_{\mathcal{A}_{\leq p_\delta}}^*\right\|_{L_2}$$

Proposition C.52. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2 = O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot \left\|b_{\mathcal{E}^c(r)}\right\|_2. \tag{316}$$

Next we show that the third term in Eq. (314) is small. We defer the proof to Appendix C.4.9.

Proposition C.53. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\|\beta_\varepsilon^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2 = O_{d,\mathbb{P}}(d^{-\lambda_\iota/2}) \cdot \sigma_\varepsilon. \tag{317}$$

The remainder of proof follows the proof of Proposition B.25. Combining the results of the three propositions then the right side of Eq. (314) becomes

$$O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\|\partial_r f_{\mathcal{A}_{\leq p_\delta}}^*\right\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot (\left\|b_{\mathcal{E}^c(r)}\right\|_2 + \sigma_\varepsilon).$$

Using the elementary inequality $|\,\|a\|_2^2 - \|b\|_2^2\,| \le \|a-b\|_2^2 + 2\|b\|_2\|a-b\|_2$ for any vectors a, b yields

$$\left|\left\|\beta^\top\Phi_{\mathcal{A}_{\le p_\delta}}R\right\|_2^2 - \left\|b_{\mathcal{A}_{\le p_\delta}}^\top VE_r\right\|_2^2\right|$$

$$\le O_{d,\mathbb{P}}\left(d^{-\lambda_\iota}\right)\cdot(\|b_{\mathcal{E}^c(r)}\|_2^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon})\cdot\left\|\partial_r f_{\mathcal{A}_{\le p_\delta}}^*\right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right)\cdot(\|b_{\mathcal{E}^c(r)}\|_2 + \sigma_\varepsilon)\left\|\partial_r f_{\mathcal{A}_{\le p_\delta}}^*\right\|_{L_2}. \tag{318}$$

Note that we have the estimate

$$\left\|V_{\mathcal{A}_{<p_\delta}} - I\right\|_{\mathrm{op}} = O_d(d^{-\zeta_1}), \tag{319}$$

and for $S \in \mathcal{A}_{=p_\delta}$, since $n(D_{S,S}W_{S,S}) = \Theta_d(1) + O_d(d^{-1})$, there exists a constant $c_r > 0$ such that the following holds

$$\left|\sum_{S\in\mathcal{A}_{=p_\delta}\cap\mathcal{E}(r)} V_{S,S}^2 b_S^2 - c_r\cdot\left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2\right| = O_d(d^{-1})\cdot\left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2.$$

Therefore, we conclude

$$\left|\left\|b_{\mathcal{A}_{\le p_\delta}}^\top VE_r\right\|_2^2 - \left\|\partial_r f_{\mathcal{A}_{<p_\delta}}^*\right\|_{L_2}^2 - c_r\cdot\left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2\right| = O_{d,\mathbb{P}}(d^{-\zeta_1})\cdot\left\|b_{\mathcal{A}_{<p_\delta}}\right\|_2^2. \tag{320}$$

Plugging the bound Eq. (320) into (318) gives the bound for

$$\left|\left\|\beta_{\mathcal{E}(r)}^\top\Phi_{\mathcal{A}_{\le p_\delta}}R\right\|_2^2 - \left\|\partial_r f_{\mathcal{A}_{<p_\delta}}^*\right\|_{L_2}^2 - c_r\cdot\left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2\right|,$$

which is

$$O_{d,\mathbb{P}}\left(d^{-\lambda_\iota}\right)\cdot(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon})\cdot\left\|\partial_r f_{\mathcal{A}_{\le p_\delta}}^*\right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right)\cdot(\|f^*\|_{L_2} + \sigma_\varepsilon)\left\|\partial_r f_{\mathcal{A}_{\le p_\delta}}^*\right\|_{L_2}.$$

where we use the inequality $\|b_{\mathcal{E}^c(r)}\|_2^2 \le \|f^*\|_{L_2}^2$. Then we conclude the proof.

C.4.5 Proof of Proposition B.32

Invoking Eq. (131), we can deduce

$$\left\|\beta^\top\Phi_{\mathcal{S}_m^1}R_m^1 - g^{(\|m\|_1)}(0)\cdot\beta^\top\Phi_{\mathcal{S}_m^1}W_{\mathcal{S}_m^1}d^{-\|m\|_1}\right\|_2$$

$$\le \left\|\beta^\top\Phi_{\mathcal{S}_m^1}\right\|\cdot O_d(d^{-\boldsymbol{\lambda}^\top m-1})$$

$$\le \|y\|_2\left\|K_{w,\lambda}^{-1}\right\|\left\|\Phi_{\mathcal{S}_m^1}\right\|_{\mathrm{op}}\cdot O_d(d^{-\boldsymbol{\lambda}^\top m-1})$$

$$\le O_{d,\mathbb{P}}(\sqrt{n}\log(d))\cdot(\|f^*\|_{L_2} + \sigma_\varepsilon)\cdot O_{d,\mathbb{P}}(1)\cdot O_{d,\mathbb{P}}(\sqrt{n})\cdot O_d(d^{-\boldsymbol{\lambda}^\top m-1})$$

$$= O_{d,\mathbb{P}}(d^{-1-\zeta_1}\log(d))\cdot(\|f^*\|_{L_2} + \sigma_\varepsilon),$$

where we used the fact that $\left\|\Phi_{\mathcal{S}_m^1}\right\|_{\mathrm{op}} = \|X_r\|_{\mathrm{op}}\cdot\left\|\Phi_{\mathcal{S}_{m-e_\iota}^0}\right\|_{\mathrm{op}} \le \left\|\Phi_{\mathcal{A}_{\le p_\delta}}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ from Lemma A.11 and the basic estimate $\|y\|_2 = O_{d,\mathbb{P}}(\sqrt{n}\log^{1/2}(d))(\|f^*\|_{L_2} + \sigma_\varepsilon)$. The bound $\left\|K_{w,\lambda}^{-1}\right\|_{\mathrm{op}} \le 1/(\rho + \lambda - \|\Delta_1\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(1)$ is implied by the expression (121).

Next, we analyze $d^{-\|\boldsymbol{m}\|_1}\beta^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1}$. We split the whole index set of features $\cup_{j=0}^\ell \mathcal{S}_j$ into $\mathcal{Q} \sqcup \mathcal{S}_m^1$. Correspondingly, we split β into

$$\beta_{\mathcal{Q}} := K_{w,\lambda}^{-1} \Phi_{\mathcal{Q}} b_{\mathcal{Q}}, \quad \beta_{\mathcal{S}_m^1} := K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1} b_{\mathcal{S}_m^1}, \quad \text{and} \quad \beta_\varepsilon := K_{w,\lambda}^{-1} \boldsymbol{\varepsilon},$$

then we obtain the bound

$$\left\| d^{-\|\boldsymbol{m}\|_1}\beta^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} - (\rho+\lambda)^{-1} d^{p+\delta-\|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2$$
$$\leq \left\| d^{-\|\boldsymbol{m}\|_1}\beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} \right\|_2 + \left\| d^{-\|\boldsymbol{m}\|_1}\beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} - (\rho+\lambda)^{-1} d^{p+\delta-\|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2$$
$$+ \left\| d^{-\|\boldsymbol{m}\|_1}\beta_\varepsilon^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} \right\|_2. \tag{321}$$

For the first term, we apply the following proposition. The proof is deferred to Appendix C.4.10.

Proposition C.54. *For any $\epsilon > 0$, the following estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\| d^{-\|\boldsymbol{m}\|_1}\beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} \right\|_2^2 = O_{d,\mathbb{P}}(d^{-\zeta_2-\lambda_\iota+\epsilon}) \cdot \|b_{\mathcal{Q}}\|_2^2. \tag{322}$$

Now we proceed to analyze the second term. The following proposition gives an estimate for the second term. We defer the proof to Appendix C.4.11.

Proposition C.55. *For any $\epsilon > 0$, the following estimate holds uniformly over all coordinates $r \in [d]$*

$$\left\| \frac{1}{n}\beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} - (\rho+\lambda)^{-1} b_{\mathcal{S}_m^1}^\top \right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot \|b_{\mathcal{S}_m^1}\|_2. \tag{323}$$

Similarly, after rescaling and multiplying both terms on the left side of the above equation by $W_{\mathcal{S}_m^1}$, we yield

$$\left\| d^{-\|\boldsymbol{m}\|_1}\beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} - (\rho+\lambda)^{-1} d^{p+\delta-\|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2$$
$$= d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}} \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot \|b_{\mathcal{S}_m^1}\|_2. \tag{324}$$

Next, we show that the third term is small. We defer the proof of the following proposition to Appendix C.4.12.

Proposition C.56. *For any $\epsilon > 0$, the following estimate holds uniformly for all $r \in \mathcal{T}_\iota$:*

$$\left\| d^{-\|\boldsymbol{m}\|_1}\beta_\varepsilon^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} \right\|_2^2 = O_{d,\mathbb{P}}(d^{-\zeta_2-\lambda_\iota+\epsilon}) \cdot \sigma_\varepsilon^2 \tag{325}$$

Lastly, plugging Eq. (322), (324) and (325) into (321) yields:

$$\left\| d^{-\|\boldsymbol{m}\|_1}\beta^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} - (\rho+\lambda)^{-1} d^{p+\delta-\|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2$$
$$\leq d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}} \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot \|b_{\mathcal{S}_m^1}\|_2 + O_{d,\mathbb{P}}(d^{-\zeta_2/2-\lambda_\iota/2+\epsilon/2}) \cdot (\|b_{\mathcal{Q}}\|_2 + \sigma_\varepsilon).$$

Following a similar argument with the proof of Proposition B.26, we can derive a bound for

$$\left| \left\| d^{-\|\boldsymbol{m}\|_1}\beta^\top \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1} \right\|_2^2 - \left\| (\rho+\lambda)^{-1} d^{p+\delta-\|\boldsymbol{m}\|_1} b_{\mathcal{S}_m^1}^\top W_{\mathcal{S}_m^1} \right\|_2^2 \right|,$$

which takes the form

$$O_{d,\mathbb{P}}(d^{-\zeta_2-\lambda_\iota+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + d^{2(p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m})} \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot \|b_{\mathcal{S}_m^1}\|_2^2$$
$$+ O_{d,\mathbb{P}}(d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}-\zeta_2/2-\lambda_\iota/2+\epsilon/2}) \cdot \|b_{\mathcal{S}_m^1}\|_2 (\|f^*\|_{L_2} + \sigma_\varepsilon),$$

where we use the fact $\|b_{\mathcal{Q}}\|_2 \leq \|f^*\|_{L_2}$.

Combining with the fact $\|b_{\mathcal{S}_m^1}\|_2 = \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}$ and using the condition $\zeta_2 > \epsilon$ completes the proof.

103

C.4.6 Proof of Proposition C.51

In the proof, we will frequently use the following estimate for any Δ with $\|\Delta\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$:

$$\left\|(I + \Delta)^{-1} - I\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}}), \tag{326}$$

which can be deduced using Neumann series.

Notice that Eq. (121) shows

$$K_w = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + \rho I_n + \Delta_1,$$

where $\rho \geq 0$ is a constant and Δ_1 satisfies $\|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$.

We write

$$K_{w,\lambda} = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + (\rho + \lambda)H, \tag{327}$$

where

$$H := I_n + (\rho + \lambda)^{-1}\Delta_1. \tag{328}$$

Invoking Sherman-Morrison-Woodbury formula to compute $K_{w,\lambda}^{-1}$ in Eq. (327), we can obtain

$$\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R$$

$$= \frac{1}{n} b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \underbrace{\left((\rho + \lambda)(n\widehat{D}_{\mathcal{A}_{\leq p_\delta}})^{-1} + \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}}{n} \right)^{-1}}_{G} \left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R \right). \tag{329}$$

Now we analyze G^{-1}. Notice that $\left\| \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}}{n} - I_n \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon})$ (Theorem A.12) and

$$\left\|H^{-1} - I_n\right\|_{\mathrm{op}} = O_d(\|\Delta_1\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon}), \tag{330}$$

which follows from Eq. (247).

Then similar to Eq. (252), we can show

$$\left\| \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}}{n} - I \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}), \tag{331}$$

which immediately implies

$$\left\| G - \left(\underbrace{(\rho + \lambda)(n\widehat{D}_{\mathcal{A}_{\leq p_\delta}})^{-1} + I}_{=V^{-1}} \right) \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}). \tag{332}$$

Next we bound $\left\|G^{-1}\right\|_{\mathrm{op}}$. Since $G = V^{-1} + G - V^{-1} = V^{-1}(I + V(G - V^{-1}))$, we have

$$G^{-1} = (I + V(G - V^{-1}))^{-1}V,$$

where $\left\|V(G - V^{-1})\right\|_{\mathrm{op}} \leq \left\|G - V^{-1}\right\|_{\mathrm{op}} \|V\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon})$.

Therefore, by inducing the Neumann series, we have

$$G^{-1} = \sum_{k=0}^{\infty} (-V(G - V^{-1}))^k V,$$

which implies the bound

$$\left\|G^{-1}\right\|_{\mathrm{op}} \leq \frac{\|V\|_{\mathrm{op}}}{1 - \left\|V(G - V^{-1})\right\|_{\mathrm{op}} \|V\|_{\mathrm{op}}} = O_d(1). \tag{333}$$

From this, we can derive

$$\begin{aligned}
\left\|G^{-1} - V\right\|_{\mathrm{op}} &= \left\|G^{-1}(V^{-1} - G)V\right\|_{\mathrm{op}} \\
&\leq \left\|G^{-1}\right\|_{\mathrm{op}} \left\|G - V^{-1}\right\|_{\mathrm{op}} \|V\|_{\mathrm{op}} \\
&= O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}).
\end{aligned} \tag{334}$$

The following claim estimates the term in the second parentheses of Eq. (329). The proof is deferred to Appendix C.4.7.

Claim 17. *The following estimate holds*

$$\left\|\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R - E_r\right\|_{\mathrm{op}} = O_d(d^{-1} + w_r^2 d^{-2}). \tag{335}$$

Eq. (335) and (334) together give

$$\begin{aligned}
\left\|G^{-1}\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R - V E_r\right\|_{\mathrm{op}} &\leq \left\|G^{-1} - V\right\|_{\mathrm{op}} \|E_r\|_{\mathrm{op}} + \left\|G^{-1}\right\|_{\mathrm{op}} \left\|\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R - E_r\right\|_{\mathrm{op}} \\
&= O_{d,\mathbb{P}}(w_r^2 d^{-2} + d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}).
\end{aligned}$$

Applying this estimate to Eq. (329) and obtain

$$\left\|\beta_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}} R - \frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + w_r^2 d^{-2}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2, \tag{336}$$

where we use the fact that $\left\|H^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (330)), $\left\|G^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (333)), and $\left\|\Phi_{\mathcal{E}(r)}\right\|_{\mathrm{op}} \leq \left\|\Phi_{\mathcal{A}_{\leq p_\delta}}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11).

In the following, we show

$$\left\|\frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r - b_{\mathcal{A}_{\leq p_\delta}} V E_r\right\|_2 = o_{d,\mathbb{P}}(1), \tag{337}$$

then plugging it back to Eq. (336) completes the proof.

First, applying the estimate Eq. (330) yields

$$\left\|\frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r - \frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2. \tag{338}$$

Then invoking the bound $\left\|\Phi_{\mathcal{A}_{\leq p_\delta}}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}}/n - I\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon})$ (Theorem A.12), we have

$$\begin{aligned}
\left\|\frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r - b_{\mathcal{A}_{\leq p_\delta}}^{\top} V E_r\right\|_2 &= \left\|\frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^{\top} E_r \Phi_{\mathcal{A}_{\leq p_\delta}}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r - b_{\mathcal{A}_{\leq p_\delta}}^{\top} V E_r\right\|_2 \\
&\leq \left\|b_{\mathcal{A}_{\leq p_\delta}}^{\top} E_r \left(\Phi_{\mathcal{A}_{\leq p_\delta}}^{\top} \Phi_{\mathcal{A}_{\leq p_\delta}}/n - I\right) V E_r\right\|_2 \\
&\leq \left\|b_{\mathcal{A}_{\leq p_\delta}}^{\top} E_r V E_r\right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}) \\
&= \left\|b_{\mathcal{E}(r)}\right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}).
\end{aligned}$$

Plugging this estimate into Eq. (338) yields

$$\left\|\frac{1}{n} b_{\mathcal{E}(r)}^{\top} \Phi_{\mathcal{E}(r)}^{\top} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} V E_r - b_{\mathcal{A}_{\leq p_\delta}}^{\top} V E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + w_r^2 d^{-2}) \cdot \left\|b_{\mathcal{E}(r)}\right\|_2.$$

Combining with Eq. (336) completes the proof.

C.4.7 Proof of Claim 17

Note that R contains blocks $R_{\mathcal{S}_m^1}$ and $R_{\mathcal{S}_m^0}$. We discuss each kind of block separately. Also recall that D is a diagonal matrix with entry $D_{\mathcal{S}_m} = g^{(\|m\|_1)}(0)d^{-\|m\|_1} + O_d(d^{-\|m\|_1-1})$ and each entry of $W_{\mathcal{S}_m}$ is of the order $d^{\|m\|_1-\lambda^\top m}$ based on Eq. (118). We will use these two properties in the analysis.

For $S \in \mathcal{S}_m^1$ or $\mathcal{S}_m^0$, based on the definition of D Eq. (108) and W Eq. (118) we have

$$(\widehat{D})_{S,S}^{-1} = (DW)_{S,S}^{-1} = \left(g^{(\|m\|_1)}(0)\right)^{-1} W_{S,S}^{-1} d^{\|m\|_1} + O_d(d^{-\lambda^\top m-1}).$$

Since Eq. (131) gives

$$R_{S,S} = g^{(\|m\|_1)}(0)W_{S,S}d^{-\|m\|_1} + O_d(d^{-\lambda^\top m-1}),$$

by a straightforward calculation, we obtain

$$(\widehat{D}^{-1}R)_{S,S} = 1 + O_d(d^{-1}).$$

Similarly, for $S \in \mathcal{S}_m^0$, Eq. (130) gives

$$R_{S,S} = g^{(\|m\|_1+2)}(0)W_{S,S}w_r^2 d^{-\|m\|_1-2} + w_r^2 \cdot O_d(d^{-\lambda^\top m-3}).$$

From this, we derive

$$(\widehat{D}^{-1}R)_{S,S} = w_r^2 \cdot O_d(d^{-2}).$$

Combining these two cases yields

$$(\widehat{D}^{-1}R)_{S,S} = \begin{cases} 1 + O_d(d^{-1}) & \text{if } S \in \cup_m \mathcal{S}_m^1; \\ w_r^2 \cdot O_d(d^{-2}), & \text{if } S \in \cup_m \mathcal{S}_m^0; \\ 0 & \text{otherwise.} \end{cases} \tag{339}$$

Combining this with the expression E_r:

$$(E_r)_{S,S} = \mathbb{1}\left\{S \in \cup_m \mathcal{S}_m^1\right\}, \tag{340}$$

we complete the proof.

C.4.8 Proof of Proposition C.52

The proof is analogous to that of Proposition C.45. Similarly in the proof, we will frequently use the following estimate for any Δ with $\|\Delta\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, which can be deduced from the Neumann series:

$$\left\|(I+\Delta)^{-1} - I\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}}), \tag{341}$$

therefore we have $(I+\Delta)^{-1} = I + \Delta'$ with some Δ' satisfying $\|\Delta'\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\|\Delta\|_{\mathrm{op}})$.

Now we are ready to present the proof. For simplicity of notation, we omit (r) in $\mathcal{E}(r)$ and $\mathcal{E}^c(r)$. We first expand $K_{w,\lambda}^{-1}$ using Eq. (115) in Lemma B.30. There exists a Δ that satisfies $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{(\delta-2)/2+\epsilon}) = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$ when ϵ is sufficiently small such that the following holds

$$K_{w,\lambda} = K_{w,p} + \underbrace{(\rho+\lambda)I_n + \sum_{m \in J_1} \mathrm{off}\left(\Phi_{\mathcal{S}_m}(Z)D_{\mathcal{S}_m}\Phi_{\mathcal{S}_m}(Z)^\top\right) + \Delta}_{:=(\rho+\lambda)H}, \tag{342}$$

where J_1 is defined in Eq. (308).

We additionally denote

$$H := I_n + (\rho + \lambda)^{-1} \sum_{\boldsymbol{m} \in J_1} \mathsf{off}\left(\Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z) D_{\mathcal{S}_{\boldsymbol{m}}} \Phi_{\mathcal{S}_{\boldsymbol{m}}}(Z)^\top\right) + (\rho + \lambda)^{-1}\Delta. \tag{343}$$

Then we can write $K_{w,\lambda} = K_{w,p} + (\rho + \lambda)H$ and use Sherman-Morrison-Woodbury formula to compute $K_{w,\lambda}^{-1}$:

$$\beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R$$

$$= \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \underbrace{\left((\rho + \lambda)(n\widehat{D}_{\mathcal{A}_{\leq p_\delta}})^{-1} + \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}}{n}\right)^{-1}}_{G} \left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R\right). \tag{344}$$

We first simplify Eq. (344). Following Claim 17 which shows $\left\|\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R - E_r\right\|_{\mathrm{op}} = O_d(d^{-1} + w_r^2 d^{-2})$ we can obtain

$$\left\|\beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R - \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} E_r\right\|_2 = O_{d,\mathbb{P}}(d^{-1} + w_r^2 d^{-2}) \cdot \log(d) \|b_{\mathcal{E}^c}\|_2, \tag{345}$$

where we use the fact that $\left\|H^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (330)), $\left\|G^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (333)), and $\|\Phi_{\mathcal{E}} b_{\mathcal{E}^c}\|_2 = O_{d,\mathbb{P}}(\sqrt{n}\log(d)) \|b_{\mathcal{E}^c}\|_2$ (Proposition B.21).

Consequently, there exists Δ_1 that satisfies $\|\Delta_1\|_2 = O_{d,\mathbb{P}}((d^{-1} + w_r^2 d^{-2})\log(d)) \cdot \|f_{\mathcal{D}^c}^*\|_{L_2}$ such that the following holds:

$$\beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R = \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} E_r + \Delta_1. \tag{346}$$

Corresponding to Claim 16, we can similarly estimate H^{-1} using Neumann series, that is, there exists a constant $c_H > 0$ such that the following holds:

$$H^{-1} = I_n + \sum_{q=1}^{c_H} \left(\sum_{\boldsymbol{m} \in J_1} -c_{\boldsymbol{m}} B_{\boldsymbol{m}} + c_{\boldsymbol{m}} h_{\boldsymbol{m}} I_n\right)^q + \Delta_2, \tag{347}$$

where $B_{\boldsymbol{m}} := \Phi_{\mathcal{S}_{\boldsymbol{m}}} \widehat{D}_{\mathcal{S}_{\boldsymbol{m}}} \Phi_{\mathcal{S}_{\boldsymbol{m}}}^\top$ and $c_{\boldsymbol{m}}, h_{\boldsymbol{m}}$ are constants, and $\|\Delta_2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$ holds.

We plug the above expression of H^{-1} into G in Eq. (344). We further substitute $\frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}}{n}$ for $\mathsf{off}\left(\frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}}{n}\right) + I_{|\mathcal{A}_{\leq p_\delta}|}$ and we obtain

$$G = V^{-1} + \mathsf{off}\left(\frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}}{n}\right) + \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{\sqrt{n}} \sum_{q=1}^{c_H} \left(\sum_{\boldsymbol{m} \in J_1} -c_{\boldsymbol{m}} B_{\boldsymbol{m}} + c_{\boldsymbol{m}} h_{\boldsymbol{m}} I_n\right)^q \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}}{\sqrt{n}} + \Delta_3,$$

where V is defined in Eq. (312) and Δ_3 satisfies $\|\Delta_3\|_{\mathrm{op}} \leq \left\|\Phi_{\mathcal{A}_{\leq p_\delta}}/\sqrt{n}\right\|_{\mathrm{op}}^2 \|\Delta_2\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$.

Note that V^{-1} is a diagonal matrix with each entry greater than 1. For a small perturbation Δ with $\|\Delta\|_{\mathrm{op}} = o_d(1)$, a modified Neumann series takes the form:

$$(V^{-1} + \Delta)^{-1} = (I + V\Delta)^{-1} V = V + \sum_{r=1}^{\infty} (-1)^r (V\Delta)^r V.$$

Then similar to Eq. (268), we express G^{-1} using modified Neumann series with higher-order terms truncated. Consequently, we can find a consntant $c_G > 0$ such that the following holds:

$$G^{-1} = V + \sum_{r=1}^{c_G} V^r \left(I - \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right.$$
$$\left. + \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{\sqrt{n}} \sum_{q=1}^{c_H} \left(\sum_{\boldsymbol{m} \in J_1} -c_{\boldsymbol{m}} B_{\boldsymbol{m}} + c_{\boldsymbol{m}} h_{\boldsymbol{m}} I_n \right)^q \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}}{\sqrt{n}} \right)^r V + \Delta_4, \qquad (348)$$

Δ_4 satisfies $\|\Delta_4\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}-\epsilon})$.

Plugging the approximations of H^{-1} (Eq. (347)) and G^{-1} (Eq. (348)) into Eq. (346) and expanding the expression we can write it as a finite sum:

$$b_{\mathcal{E}^c}^\top \frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}} H^{-1} \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}}{\sqrt{n}} G^{-1} E_r = \sum_{i=1}^{c_p} b_{\mathcal{E}^c}^\top \frac{\Phi_{\mathcal{E}^c}^\top}{\sqrt{n}} P_i + \Delta_5,$$

where each P_i takes the form

$$P_i = C_i A_{i_1} A_{i_2} \dots A_{i_m} \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V^{t_i} E_r}{\sqrt{n}}$$

with each $A_{i_j} \in \{I_n\} \cup \left\{ \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V^s \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right\}_{s=0}^{c_G+1} \cup \{B_{\boldsymbol{m}}\}_{\boldsymbol{m} \in J_1}$ for $j \in [m]$.

Here $t_i \in \mathbb{N}$, c_p, C_i are constants. Additionally, Δ_5 satisfies

$$\|\Delta_5\|_2 = O_{d,\mathbb{P}}(\log(d)) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2} \cdot (\|\Delta_2\|_{\mathrm{op}} + \|\Delta_4\|_{\mathrm{op}}) = O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2},$$

where we use the estimate of H^{-1} (Eq. (330)), G^{-1} (Eq. (333)), $\left\| \Phi_{\mathcal{A}_{\leq p_\delta}} / \sqrt{n} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Lemma A.11) and $\left\| \frac{\Phi_{\mathcal{E}^c} b_{\mathcal{E}^c}}{\sqrt{n}} \right\|_2 = O_{d,\mathbb{P}}(\log^{1/2}(d)) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}$ (Proposition B.21).

We can prove the following result similar to Proposition C.50. The proof appears in Appendix C.4.13.

Proposition C.57. *Let A_j be defined as one of the following matrices:*

$$\{I_n\}, \quad \left\{ \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V^s \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right\}_{s=0}^{c_G+1}, \quad \left\{ \Phi_{\mathcal{S}_{\boldsymbol{m}}} \widehat{D}_{\mathcal{S}_{\boldsymbol{m}}} \Phi_{\mathcal{S}_{\boldsymbol{m}}}^\top \right\}_{\boldsymbol{m} \in J_1}.$$

Then for any $t \in \mathbb{N}$ the following estimate holds

$$\mathbb{E}\left[\left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top A_1 A_2 \dots A_m \Phi_{\mathcal{A}_{\leq p_\delta}} V^t E_r \right\|_2^2 \right] = O_d\left(d^{-\lambda_\iota - \zeta_1} \right) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}^2 \qquad (349)$$

The remaining steps of the proof proceed following Proposition C.45. We can demonstrate that this estimate holds uniformly across all $r \in [d]$ with an extra multiplicative error factor, then we can conclude the bound:

$$\max_r \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} E_r \right\|_2^2 = \left(O_{d,\mathbb{P}}(d^{-\lambda_\iota - \zeta_1 + 1/q_0} \log(d)) + O_{d,\mathbb{P}}(d^{-1}) \right) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}^2,$$

where $q_0 > 0$ is an arbitrarily large constant. For complete details of this argument, we direct readers to the referenced proposition.

Plugging this bound into Eq. (346) and applying the AM-GM inequality yields

$$\left\|\beta_{\mathcal{E}^c}^\top \Phi_{\leq p} R\right\|_2^2 \leq 2 \left\|\frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} E_r\right\|_2^2 + 2 \left\|\Delta_1\right\|_2^2$$
$$= \left(O_{d,\mathbb{P}}(d^{-\lambda_\iota - \zeta_1 + \epsilon}) + O_{d,\mathbb{P}}(d^{-1}) + O_{d,\mathbb{P}}(w_r^4 d^{-4+\epsilon})\right) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}^2,$$

where we absorb d^{1/q_0} and $\log(d)$ into d^ϵ. With the fact $\lambda_\iota \leq 1$ and $\zeta_1 > \epsilon$, the bound is reduced to $O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \cdot \|f_{\mathcal{E}^c}^*\|_{L_2}^2$ thus we conclude the proof.

C.4.9 Proof of Proposition C.53

The proof follows closely the proof of Proposition C.46 hence we omit intermediate derivations. We refer the readers to Appendix C.3.6 for the complete derivation.

We can similarly establish

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\|\beta_\varepsilon^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2^2\right] \leq \sigma_\varepsilon^2 \cdot \left\|K_\lambda^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}}\right\|_{\mathrm{op}}^2 \cdot \mathrm{Tr}\left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R R^\top \widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1}\right), \tag{350}$$

and successively can show

$$\left\|K_\lambda^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}}\right\|_{\mathrm{op}}^2 = \left\|\frac{1}{n} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G\right\|_{\mathrm{op}}^2 = O_{d,\mathbb{P}}(n^{-1}), \tag{351}$$

implied by Eq. (329), and

$$\mathrm{Tr}\left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R R^\top \widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1}\right) = O_d(\|E_r\|_F^2) = O_d(d^{p+\delta-\lambda_\iota-\zeta_1}), \tag{352}$$

which follows Eq. (311).

Plugging Eq. (351) and (352) into Eq. (350) and applying Markov's inequality completes the proof.

C.4.10 Proof of Proposition C.54

The proof will mainly use the machinery developed in the proof for Proposition C.52.

Applying the same decomposition of K_λ as Eq. (342), we can write

$$K_{w,\lambda} = K_{w,p} + (\rho + \lambda) H,$$

where H is defined in Eq. (343).

Invoking Sherman-Morrison-Woodbury formula for $K_{w,\lambda}^{-1}$, we can obtain

$$K_{w,\lambda}^{-1} = (\rho + \lambda)^{-1} H^{-1} - \frac{1}{n}(\rho + \lambda)^{-1} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top G^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} H^{-1}, \tag{353}$$

where

$$G := (\rho + \lambda)(n\widehat{D}_{\mathcal{A}_{\leq p_\delta}})^{-1} + \frac{\Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}}{n}.$$

Plugging the approximations of H^{-1} (Eq. (347)) and G^{-1} (Eq. (348)) into Eq. (353) and expanding the expression we can write it as the finite summation:

$$\frac{1}{n} b_{\mathcal{Q}}^\top \Phi_{\mathcal{Q}}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1} = \sum_{i=1}^{c_p} C_i b_{\mathcal{Q}}^\top \frac{\Phi_{\mathcal{Q}}^\top}{\sqrt{n}} A_{i_1} A_{i_2} \dots A_{i_m} \frac{\Phi_{\mathcal{S}_m^1}}{\sqrt{n}} + \Delta, \tag{354}$$

where $A_{i_j} \in \{I_n\} \cup \left\{\frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V^s \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n}\right\}_{s=0}^{c_G+1} \cup \left\{\Phi_{\mathcal{S}_{m'}} \widehat{D}_{\mathcal{S}_{m'}} \Phi_{\mathcal{S}_{m'}}^\top\right\}_{\boldsymbol{m'} \in J_1}$ for $j \in [m]$. Here $C_i > 0$ is a constant, $c_p, m \in \mathbb{N}$, and Δ satisfies $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\frac{1}{2}}) \cdot \|b_{\mathcal{Q}}\|_2$.

Note that the adjusted effective degree of the feature in $\Phi_{\mathcal{S}_m^1}$ is bounded by $\boldsymbol{s}^\top \boldsymbol{m} - s_\iota$, following an almost identical argument with the proof of Proposition C.57 by using Lemma D.71, we can show each summand satisfies:

$$\mathbb{E}\left[\left\|\frac{1}{n} b_{\mathcal{Q}}^\top \Phi_{\mathcal{Q}}^\top A_{i_1} A_{i_2} \dots A_{i_m} \Phi_{\mathcal{S}_m^1}\right\|_2^2\right] = O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - p - \delta}) \cdot \|b_{\mathcal{Q}}\|_2^2.$$

Using a similar analysis as in the proof of Proposition C.52, we can show that the estimates hold uniformly for all $r \in T_\iota$. Specifically, invoking Lemma D.70 and Markov's inequality together yield the bound

$$\left\|\frac{1}{n} b_{\mathcal{Q}}^\top \Phi_{\mathcal{Q}}^\top A_{i_1} A_{i_2} \dots A_{i_q} \Phi_{\mathcal{S}_m^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - p - \delta + 1/q_0}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2,$$

where $q_0 > 0$ is an arbitrarily large constant.

Combining with Δ in Eq. (354) and applying the triangle inequality and applying the triangle inequality we obtain the bound

$$\left\|\frac{1}{n} b_{\mathcal{Q}}^\top \Phi_{\mathcal{Q}}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - p - \delta + 1/q_0}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2. \tag{355}$$

Multiplying both sides of the equation by $(n/d^{\|\boldsymbol{m}\|_1})^2$ and taking $W_{\mathcal{S}_m^1}$ into account, which satisfies $\left\|W_{\mathcal{S}_m^1}\right\|_{\mathrm{op}} = O_d(d^{\|\boldsymbol{m}\|_1 - \boldsymbol{\lambda}^\top \boldsymbol{m}})$ yields:

$$\left\|d^{-\|\boldsymbol{m}\|_1} b_{\mathcal{Q}}^\top \Phi_{\mathcal{Q}}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1} W_{\mathcal{S}_m^1}\right\|_2^2 = O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - p - \delta + 1/q_0 + 2(p+\delta) - 2\|\boldsymbol{m}\|_1 + 2\|\boldsymbol{m}\|_1 - 2\boldsymbol{\lambda}^\top \boldsymbol{m}}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2,$$
$$= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - 2\boldsymbol{\lambda}^\top \boldsymbol{m} - s_\iota + p + \delta + 1/q_0}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2$$
$$= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - \boldsymbol{\lambda}^\top \boldsymbol{m} - s_\iota - \zeta_2 + 1/q_0}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2$$
$$= O_{d,\mathbb{P}}(d^{-\zeta_2 - \lambda_\iota + 1/q_0}) \cdot \log(d) \cdot \|b_{\mathcal{Q}}\|_2^2,$$

where in the last equality we use the fact that $\boldsymbol{s}^\top \boldsymbol{m} - \boldsymbol{\lambda}^\top \boldsymbol{m} \leq s_\iota - \lambda_\iota \leq 0$.

Absorbing the factor $d^{1/q_0} \log(d)$ into d^ϵ where $\epsilon > 0$ is an arbitrarily small constant yields the desired bound.

C.4.11 Proof of Proposition C.55

Note that Eq. (327) gives the decomposition

$$K_{w,\lambda} = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + (\rho + \lambda) H, \tag{356}$$

where $H = I_n + (\rho + \lambda)^{-1} \Delta_1$, ρ is defined in Eq. (116) and Δ_1 satisfies $\|\Delta_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$.

Invoking Sherman-Morrison-Woodbury formula to compute the inverse of $K_{w,\lambda}$ and we obtain:

$$K_{w,\lambda}^{-1} = (\rho + \lambda)^{-1} H^{-1} - (\rho + \lambda)^{-1} \frac{1}{n} H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1},$$

where G takes the same form as the one in Eq. (329).

Invoking Eq. (330) and (334) that H^{-1} is close to I_n and G^{-1} is close to V, we can show

$$\left\|K_{w,\lambda}^{-1} - (\rho + \lambda)^{-1}\left(I_n - \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n}\right)\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}),$$

where we use the estimate $\left\|\Phi_{\mathcal{A}_{\leq p_\delta}}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11).

Plugging the above bound into the left side of Eq. (323) and applying the triangle inequality yields

$$\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1} - \frac{1}{n(\rho+\lambda)} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \left(I_n - \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right) \Phi_{\mathcal{S}_m^1} \right\|_{\mathrm{op}}$$
$$= O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2, \tag{357}$$

where we use the estimate $\left\| \Phi_{\mathcal{S}_m^1} \right\|_{\mathrm{op}} = \| X_r \|_{\mathrm{op}} \cdot \left\| \Phi_{\mathcal{S}_{m-e_\iota}^0} \right\|_{\mathrm{op}} \leq \left\| \Phi_{\mathcal{A}_{\leq p_\delta}} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ from Lemma A.11.

Next we analyze $\frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \left(I_n - \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right) \Phi_{\mathcal{S}_m^1}$. Using the triangle inequality gives

$$\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \left(I_n - \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \right) \Phi_{\mathcal{S}_m^1} - b_{\mathcal{S}_m^1}^\top \right\|_2$$
$$\leq \left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} - b_{\mathcal{S}_m^1}^\top \right\| + \left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \Phi_{\mathcal{S}_m^1} \right\|_2. \tag{358}$$

Next we show that both terms on the right side of the above inequality are small.

For the first term, we use Theorem A.13 which gives

$$\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} - b_{\mathcal{S}_m^1}^\top \right\| \leq \left\| b_{\mathcal{S}_m^1} \right\|_2 \left\| \frac{1}{n} \Phi_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} - I_n \right\|_{\mathrm{op}}$$
$$= \left\| b_{\mathcal{S}_m^1} \right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}). \tag{359}$$

For the second term, invoking Lemma D.71, since the adjusted effective degree of feature $\phi_{\mathcal{A}_{\leq p_\delta}}$ is bounded by $s^\top m'$ where m' satisfies $\lambda^\top m' \leq p + \delta$, we can deduce

$$\mathbb{E}\left[\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \Phi_{\mathcal{S}_m^1} \right\|_2^2 \right] = \left\| b_{\mathcal{S}_m^1} \right\|_2^2 \cdot \max_{\lambda^\top m' \leq p+\delta} O_d(d^{s^\top m'-p-\delta})$$
$$= \left\| b_{\mathcal{S}_m^1} \right\|_2^2 \cdot O_d(d^{-\zeta_1}),$$

where the last equality follows from the definition of ζ_1, i.e., Eq. (104).

Using a similar analysis as in the proof of Proposition C.52, we can show that the estimates hold uniformly for all $r \in [d]$. Specifically, invoking Lemma D.70 and Markov's inequality together yield the bound

$$\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \frac{\Phi_{\mathcal{A}_{\leq p_\delta}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} \Phi_{\mathcal{S}_m^1} \right\|_2^2 = \left\| b_{\mathcal{S}_m^1} \right\|_2^2 \cdot O_d(d^{-\zeta_1+1/q_0}) \cdot \log(d),$$

where $q_0 > 0$ is an arbitrarily large constant.

Plugging this bound and Eq. (359) into Eq. (358) yields:

$$\left\| \frac{1}{n} b_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1}^\top \left(I_n - \frac{\Phi_{\leq p} V \Phi_{\leq p}^\top}{n} \right) \Phi_{\mathcal{S}_m^1} - b_{\mathcal{S}_m^1}^\top \right\|_2 = \left\| b_{\mathcal{S}_m^1} \right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}).$$

Combining this bound with Eq. (357) completes the proof.

C.4.12 Proof of Proposition C.56

The proof follows closely the proof of Proposition C.49 hence we omit intermediate derivations. We refer the readers to Appendix C.3.9 for the complete derivation.

In parallel to Eq. (282), we can obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\|d^{-\|\boldsymbol{m}\|_1}\beta_{\varepsilon}^{\top}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}W_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_2^2\right] \le \left\|W_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_{\mathrm{op}}^2 \cdot d^{-2\|\boldsymbol{m}\|_1}\cdot\mathbb{E}\left[\left\|\beta_{\varepsilon}^{\top}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_2^2\right]$$

$$\le O_d(d^{-2\boldsymbol{\lambda}^{\top}\boldsymbol{m}})\cdot\left\|K_{w,\lambda}^{-2}\right\|_{\mathrm{op}}\mathrm{Tr}\left(\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^{\top}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\right). \tag{360}$$

The bound

$$\left\|K_{w,\lambda}^{-2}\right\|_{\mathrm{op}} = \left\|K_{w,\lambda}^{-1}\right\|_{\mathrm{op}}^2 = O_{d,\mathbb{P}}(1) \tag{361}$$

is implied by the expression (121). Next we analyze $\mathrm{Tr}\left(\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^{\top}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\right)$. We first control the size $|\mathcal{S}_{\boldsymbol{m}}^1| = O_d(d^{\boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota})$. The following claim implies $|\mathcal{S}_{\boldsymbol{m}}^1|\lesssim n$ holds:

Claim 18. *The constant*

$$\zeta_3 := \min_{\boldsymbol{m}:\boldsymbol{\lambda}^{\top}\boldsymbol{m}\in(p+\delta,p+\delta+\lambda_\iota],\ m_\iota\ge 1}\left(p+\delta-\boldsymbol{s}^{\top}\boldsymbol{m}+s_\iota\right)$$

is bounded from 0.

Proof. Since $\lambda_j \ge s_j$ for all $j\in[N]$, we have $\zeta_3 \ge p+\delta-(\boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota) \ge p+\delta-(\boldsymbol{\lambda}^{\top}\boldsymbol{m}-\lambda_\iota) \ge 0$. The equality holds when (1) $\boldsymbol{\lambda}^{\top}\boldsymbol{m}-\lambda_\iota = p+\delta$ and (2) $\boldsymbol{\lambda}^{\top}\boldsymbol{m}-\lambda_\iota = \boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota$ hold. For condition (2) to hold, it requires either $m_\iota = 1$ and $\lambda_j = s_j = 1$ for $j\ne\iota$; or $m_\iota \ge 2$ and $\lambda_j = s_j = 1$ for all $j\in[N]$. Both cases contradict condition (1), hence the inequality stands strictly, which completes the proof. $\square$

Consequently, following Theorem A.12, we have $\left\|\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^{\top}/n - I_{|\mathcal{S}_{p+1}^1|}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_3/2+\epsilon})$. Then, we can conclude

$$\mathrm{Tr}\left(\frac{1}{n}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^{\top}\right) \le O_d(d^{\boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota})(1+O_{d,\mathbb{P}}(d^{-\zeta_3/2+\epsilon})) = O_d(d^{\boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota}).$$

Together with Eq. (361), the right side of Eq. (360) is bounded by

$$O_{d,\mathbb{P}}(d^{\boldsymbol{s}^{\top}\boldsymbol{m}-s_\iota-2\boldsymbol{\lambda}^{\top}\boldsymbol{m}+p+\delta}) = O_{d,\mathbb{P}}(d^{-\lambda_\iota-\zeta_2}),$$

where we use the fact $\boldsymbol{s}^{\top}\boldsymbol{m}-\boldsymbol{\lambda}^{\top}\boldsymbol{m}\le s_\iota-\lambda_\iota$.

Applying Markov's inequality we obtain the bound $O_{d,\mathbb{P}}(d^{-\lambda_\iota-\zeta_2})\log(d)$ with probability at least $1-1/\log(d)$. Absorbing the logarithmic factor into ϵ completes the proof.

C.4.13 Proof of Proposition C.57

Since V is a diagonal matrix with $V\lesssim I$, then we can upper bound the norm by removing V^t in the product:

$$\left\|\frac{1}{n}b_{\mathcal{E}^c}^{\top}\Phi_{\mathcal{E}^c}^{\top}A_1A_2\ldots A_m\Phi_{\mathcal{A}_{\le p_\delta}}V^t E_r\right\|_2^2 \lesssim O_d(1)\cdot\left\|\frac{1}{n}b_{\mathcal{E}^c}^{\top}\Phi_{\mathcal{E}^c}^{\top}A_1A_2\ldots A_m\Phi_{\mathcal{E}}\right\|_2^2,$$

where we use the fact $(E_r)_{S,S} = \mathbb{1}\{S\in\mathcal{E}\}$.

Similar to Eq. (283), the following estimate holds:

$$\mathbb{E}\left[\left\|\frac{1}{n}b_{\mathcal{E}^c}^{\top}\Phi_{\mathcal{E}^c}^{\top}A_1A_2\ldots A_m\Phi_{\mathcal{E}}\right\|_2^2\right] \le \|b_{\mathcal{E}^c}\|_2^2\cdot\left\|\mathbb{E}\left[\underbrace{\frac{1}{n^2}\Phi_{\mathcal{E}^c}^{\top}A_1\ldots A_m\Phi_{\mathcal{E}}\Phi_{\mathcal{E}}^{\top}A_m\ldots A_1\Phi_{\mathcal{E}^c}}_{=:K}\right]\right\|_{\mathrm{op}}. \tag{362}$$

We now aim to upper bound the operator norm of $\mathbb{E}[K]$. To this end, we decompose the feature matrices in K so as to ensure that distinct matrices in the product correspond to nonoverlapping index sets.

Following Eq. (126) we have the decomposition

$$\mathcal{A}_{\leq p_\delta} = \sqcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \mathcal{S}_{\boldsymbol{m}}, \quad \mathcal{S}_{\boldsymbol{m}} = \mathcal{S}_{\boldsymbol{m}}^1 \sqcup \mathcal{S}_{\boldsymbol{m}}^0, \tag{363}$$

Clearly, the following equalities hold:

$$\mathcal{E} = \bigcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \mathcal{S}_{\boldsymbol{m}}^1, \quad \mathcal{E}^c = \left\{ \bigcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \mathcal{S}_{\boldsymbol{m}}^0 \right\} \bigcup \left\{ \bigcup_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta, \|\boldsymbol{m}\|_1 \leq \ell} \mathcal{S}_{\boldsymbol{m}} \right\}$$

Define now the collection of matrices

$$\mathcal{U} := \{\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \bigcup \{\Phi_{\mathcal{S}_{\boldsymbol{m}}^0}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \bigcup \{\Phi_{\mathcal{S}_{\boldsymbol{m}}}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta}.$$

Note that any two distinct matrices in $\mathcal{U}$ correspond to distinct index sets. Normalizing the matrices, we further define the collection

$$\widetilde{\mathcal{U}} := \{n^{-\frac{1}{2}} \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \bigcup \{n^{-\frac{1}{2}} \Phi_{\mathcal{S}_{\boldsymbol{m},j}^0}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \bigcup \{(\widehat{D}_{\mathcal{S}_{\boldsymbol{m}}})^{-\frac{1}{2}} \Phi_{\mathcal{S}_{\boldsymbol{m}}}\}_{\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta},$$

where $\left\|\widehat{D}_{\mathcal{S}_{\boldsymbol{m}}}\right\|_{\mathrm{op}} = \Theta_d(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m}})$.

Let us now write the two expressions:

$$\frac{\Phi_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top}{n} = \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \left(\frac{\Phi_{\mathcal{S}_{\boldsymbol{m}}^0} \Phi_{\mathcal{S}_{\boldsymbol{m}}^0}^\top}{n} + \frac{\Phi_{\mathcal{S}_{\boldsymbol{m}}^1} \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^\top}{n} \right);$$

$$\frac{\Phi_{\mathcal{E}} \Phi_{\mathcal{E}}^\top}{n} = \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \frac{\Phi_{\mathcal{S}_{\boldsymbol{m}}^1} \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^\top}{n}.$$

Plugging these decompositions for the corresponding A_i in the expression of K Eq. (362), we can write K as the summation of finitely many terms, where each term takes the form of some positive constant multiplied by

$$K_j := \frac{\max(n^{1/2}, d^{k_1/2}) \cdot \max(n^{1/2}, d^{k_{2m+1}/2})}{n} \cdot B_1^\top \left(B_2 B_2^\top \right) \ldots \left(B_{2m} B_{2m}^\top \right) B_{2m+1},$$

where $B_1 \in \mathbb{R}^{n \times k_1}, \ldots, B_{2m+1} \in \mathbb{R}^{n \times k_{2m+1}}$ are from $\widetilde{\mathcal{U}}$.

Particularly note that B_1 and B_{2m+1} corresponding to the decomposition of $\Phi_{\mathcal{E}^c}$ and thus each takes the form $(\widehat{D}_{\mathcal{S}_{\boldsymbol{m}}})^{-\frac{1}{2}} \Phi_{\mathcal{S}_{\boldsymbol{m}}}$ for some $\boldsymbol{m}$ that satisfies $\boldsymbol{\lambda}^\top \boldsymbol{m} > p + \delta$; B_{m+1} corresponds to the decomposition of $\Phi_{\mathcal{E}}$ and takes the form $n^{-\frac{1}{2}} \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}$ for some $\boldsymbol{m}$ that satisfies $\boldsymbol{\lambda}^\top \boldsymbol{m} > p + \delta$. The equation contains an additional scaling factor on the right side, which arises from the distinct scaling of B_1 and B_{2m+1} by a factor of $1/n$ in Eq. (362).

Next, we invoke Lemma D.71 to bound each $\|\mathbb{E}[K_j]\|_{\mathrm{op}}$. According to the definition of feature $\phi_{\mathcal{S}_{\boldsymbol{m}}}$ Eq. (101), we naturally have non-overlapping sets $\mathcal{T}_1, \ldots \mathcal{T}_N$ with $|\mathcal{T}_i| = O_d(d^{s_i})$. We further split the block $\mathcal{T}_\iota$ where $r \in \mathcal{T}_\iota$ into $\mathcal{T}_\iota^1 := \{r\}$ and $\mathcal{T}_\iota^0 := \mathcal{T}_\iota \setminus \{r\}$. Then we have in total $N + 1$ many blocks of $[d]$ and any feature index S that appears in the product can be accordingly decomposed. The adjusted effective degree of the feature of B_{m+1} is thus bounded by

$$\sum_{i \in [N], i \neq \iota} s_i m_i + (m_\iota - 1) s_\iota^0 + 1 \cdot s_\iota^1,$$

where $|\mathcal{T}_\iota^1| = \Theta_d(d^{s_\iota^1}) = \Theta_d(d^0)$ and $|\mathcal{T}_\iota^0| = \Theta_d(d^{s_\iota^0}) = \Theta_d(d^{s_\iota})$. Consequently, similar to Eq. (285) we can deduce

$$\|\mathbb{E}[K_j]\|_{\mathrm{op}} = O_d(d^{\sum_{i \in [N], i \neq \iota} s_i m_i + (m_\iota - 1) s_\iota - p - \delta}).$$

113

Since $s_i \le \lambda_i$ for all $i \in [N]$ and $m_\iota > 0$, the exponent can be further bounded by

$$\sum_{i=1}^{N} \lambda_i m_i - \lambda_\iota - p - \delta \le -\lambda_\iota - \zeta_1.$$

Consequently, we obtain the bound for each $\|\mathbb{E}[K_j]\|_{\mathrm{op}}$. Applying triangle inequality yields $\|\mathbb{E}[K]\|_{\mathrm{op}} = O_d(d^{-\lambda_\iota - \zeta_1})$. Combining with Eq. (362) completes the proof.

C.5 Proof of Propositions for Theorem B.27 (DN estimator error)

C.5.1 Proof of Proposition B.33

The proof is analogous to that of Proposition B.31, and throughout we follow the conventions and notation established there.

We decompose the diagonal matrix $(\widehat{D}_{\mathcal{A}_{\le p_\delta}}^{-1} R)$ into two diagonal matrices $\widehat{E}_r$ and $\widehat{E}_r^\circ$ of size $|\mathcal{A}_{\le p_\delta}| \times |\mathcal{A}_{\le p_\delta}|$ as

$$(\widehat{E}_r)_{S,S} = (\widehat{D}_{\mathcal{A}_{\le p_\delta}}^{-1} R)_{S,S} \cdot \mathbb{1}\{S \in \mathcal{E}(r)\}$$
$$(\widehat{E}_r^\circ)_{S,S} = (\widehat{D}_{\mathcal{A}_{\le p_\delta}}^{-1} R)_{S,S} \cdot \mathbb{1}\{S \in \mathcal{E}^c(r)\},$$

where sets $\mathcal{E}(r)$ and $\mathcal{E}^c(r)$ are defined in Eq. (310).

The following claim estimates the term in $\widehat{E}_r$ and $\widehat{E}_r^\circ$. The proof is deferred to Appendix C.5.4.

Claim 19. *The following estimate holds for any $S \in \mathcal{A}_{\le p_\delta}$:*

$$(\widehat{E}_r)_{S,S} = \left(g^{(\|\boldsymbol{m}\|_1)}(0)\right)^{-\frac{1}{2}} W_{S,S}^{-\frac{1}{2}} d^{(\|\boldsymbol{m}\|_1 - p - \delta)/2} \cdot (1 + O_d(d^{-\frac{1}{2}})) \cdot \mathbb{1}\{S \in \mathcal{S}_{\boldsymbol{m}}^1\}, \tag{364}$$

$$(\widehat{E}_r^\circ)_{S,S} = w_r \cdot O_d(d^{(\boldsymbol{\lambda}^\top \boldsymbol{m} - p - \delta - 2)/2}) \cdot \mathbb{1}\{S \in \mathcal{S}_{\boldsymbol{m}}^0\}. \tag{365}$$

We similarly write $\beta = \beta_{\mathcal{E}(r)} + \beta_{\mathcal{E}^c(r)} + \beta_\varepsilon$ where we define:

$$\beta_{\mathcal{E}(r)} := K_{w,\lambda}^{-1}(\Phi_{\mathcal{E}(r)} b_{\mathcal{E}(r)}), \quad \beta_{\mathcal{E}^c(r)} := K_{w,\lambda}^{-1}(\Phi_{\mathcal{E}^c(r)} b_{\mathcal{E}^c(r)}), \quad \beta_\varepsilon = K_{w,\lambda}^{-1}\varepsilon.$$

With the decomposition of β, we accordingly bound the norm by

$$\left\|\beta^\top \Phi_{\mathcal{A}_{\le p_\delta}} R - b_{\mathcal{A}_{\le p_\delta}}^\top V \widehat{E}_r\right\|_2$$
$$\le \left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\le p_\delta}} R - b_{\mathcal{A}_{\le p_\delta}}^\top V \widehat{E}_r\right\|_2 + \left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\mathcal{A}_{\le p_\delta}} R\right\|_2 + \left\|\beta_\varepsilon^\top \Phi_{\mathcal{A}_{\le p_\delta}} R\right\|, \tag{366}$$

where the matrix V is defined in Eq. (312).

We will show all three terms on the right hand side of the above equation are small. Before doing so, we introduce some notation to simplify the proof. We define the matrix $R_{\boldsymbol{m}} \in \mathbb{R}^{|\mathcal{A}_{\le p_\delta}| \times |\mathcal{S}_{\boldsymbol{m}}|}$, which is the submatrix of R consisting of the columns indexed by $\mathcal{S}_{\boldsymbol{m}}$. Similarly we let $(\widehat{E}_r)_{\mathcal{S}_{\boldsymbol{m}}}$ denote the matrix formed by the columns of $\widehat{E}_r$ with indices in $\mathcal{S}_{\boldsymbol{m}}$. Therefore, we can decompose the first term squared into

$$\left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\le p_\delta}} R - b_{\mathcal{A}_{\le p_\delta}}^\top V \widehat{E}_r\right\|_2^2 = \sum_{\boldsymbol{m}: \boldsymbol{\lambda}^\top \boldsymbol{m} \le p + \delta} \left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\le p_\delta}} R_{\boldsymbol{m}} - b_{\mathcal{A}_{\le p_\delta}}^\top V (\widehat{E}_r)_{\mathcal{S}_{\boldsymbol{m}}}\right\|_2^2.$$

The following two propositions provide estimates for the first and the second term in Eq. (366) respectively. We defer the proof to Appendix C.5.3 and C.5.5 respectively.

Proposition C.58. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates r:*

$$\left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\le p_\delta}} R_{\boldsymbol{m}} - b_{\mathcal{A}_{\le p_\delta}}^\top V (\widehat{E}_r)_{\mathcal{S}_{\boldsymbol{m}}}\right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon + (\boldsymbol{\lambda}^\top \boldsymbol{m} - p - \delta)/2}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2}$$

Proposition C.59. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\|\beta_{\mathcal{E}^c(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2 = O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot \left\|b_{\mathcal{E}^c(r)}\right\|_2. \tag{367}$$

Next we show that the third term in Eq. (366) is small. We defer the proof to Appendix C.5.6.

Proposition C.60. *For any $\epsilon > 0$, the estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\|\beta_\varepsilon^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2 = O_{d,\mathbb{P}}(d^{-\lambda_\iota/2}) \cdot \sigma_\varepsilon. \tag{368}$$

The remainder of proof follows the proof of Proposition B.25. Combining the results of the three propositions then the right side of Eq. (366) becomes

$$\sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta-2)/2}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2} + O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot (\left\|b_{\mathcal{E}^c(r)}\right\|_2 + \sigma_\varepsilon).$$

Using the elementary inequality $| \|a\|_2^2 - \|b\|_2^2 | \leq \|a-b\|_2^2 + 2\|b\|_2 \|a-b\|_2$ for any vectors a, b yields

$$\left| \left\|\beta^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2^2 - \left\|b_{\mathcal{A}_{\leq p_\delta}}^\top V\widehat{E}_r\right\|_2^2 \right|$$

$$\leq O_{d,\mathbb{P}}\left(d^{-\lambda_\iota}\right) \cdot (\left\|b_{\mathcal{E}^c(r)}\right\|_2^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta)}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot (\left\|b_{\mathcal{E}^c(r)}\right\|_2 + \sigma_\varepsilon) \left\|\partial_r f_{\mathcal{A}_{\leq p_\delta}}^*\right\|_{L_2}. \tag{369}$$

Note that we have the estimate

$$\left\|V_{\mathcal{A}_{<p_\delta}} - I\right\|_{\text{op}} = O_d(d^{-\zeta_1}),$$

and for $S \in \mathcal{A}_{=p_\delta}$, since $n \cdot (D_{S,S} W_{S,S}) = \Theta_d(1) + O_d(d^{-1})$, there exists a constant $c_r > 0$ such that the following holds

$$\left| \sum_{S \in \mathcal{A}_{=p_\delta} \cap \mathcal{E}(r)} V_{S,S}^2 b_S^2 - c_r \cdot \left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2 \right| = O_d(d^{-1}) \cdot \left\|\partial_r f_{\mathcal{A}_{=p_\delta}}^*\right\|_{L_2}^2.$$

With the scaling of $\widehat{E}_r$ (364), we can deduce

$$\left\|b_{\mathcal{S}_{\boldsymbol{m}}^1}^\top (\widehat{E}_r)_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_2^2 = (g^{(\|\boldsymbol{m}\|_1)}(0))^{-1} \left\|b_{\mathcal{S}_{\boldsymbol{m}}^1}^\top W_{\mathcal{S}_{\boldsymbol{m}}^1}^{-\frac{1}{2}} d^{(\|\boldsymbol{m}\|_1-p-\delta)/2}(1+O_d(d^{-\frac{1}{2}}))\right\|_2^2 = \Theta_d(d^{\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta}) \left\|b_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_2^2.$$

Therefore, using triangle inequality we conclude

$$\left\|b_{\mathcal{A}_{\leq p_\delta}}^\top V\widehat{E}_r\right\|_2^2 = \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \Theta_d(d^{\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2}^2. \tag{370}$$

Plugging Eq. (370) into (369) gives the bound for

$$\left| \left\|\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R\right\|_2^2 - \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \Theta_d(d^{\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2}^2 \right|,$$

which is

$$O_{d,\mathbb{P}}\left(d^{-\lambda_\iota}\right) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta)}) \cdot \left\|\partial_r f_{\mathcal{S}_{\boldsymbol{m}}^1}^*\right\|_{L_2}^2$$

$$+ O_{d,\mathbb{P}}\left(d^{-\lambda_\iota/2}\right) \cdot (\|f^*\|_{L_2} + \sigma_\varepsilon) \left\|\partial_r f_{\mathcal{A}_{\leq p_\delta}}^*\right\|_{L_2},$$

where we use the inequality $\left\|b_{\mathcal{E}^c(r)}\right\|_2^2 \leq \|f^*\|_{L_2}^2$.

115

C.5.2 Proof of Proposition B.34

Invoking Eq. (144), we can deduce

$$\left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 - \beta^\top \Phi_{\mathcal{S}_m^1} \left(n^{-1} g^{(\|\boldsymbol{m}\|_1)}(0) W_{\mathcal{S}_m^1} d^{-\|\boldsymbol{m}\|_1} \right)^{\frac{1}{2}} \right\|_2$$
$$\leq \left\| \beta^\top \Phi_{\mathcal{S}_m^1} \left(n^{-1} g^{(\|\boldsymbol{m}\|_1)}(0) W_{\mathcal{S}_m^1} d^{-\|\boldsymbol{m}\|_1} \right)^{\frac{1}{2}} \right\|_2 \cdot O_d(d^{-1/2}). \tag{371}$$

For the subsequent analysis, we temporarily omit the multiplicative constant $\sqrt{g^{(\|\boldsymbol{m}\|_1)}(0)}$ and focus on the term $n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}}$. We split the whole index set of features $\cup_{j=0}^\ell \mathcal{S}_j$ into $\mathcal{Q} \sqcup \mathcal{S}_m^1$. Correspondingly, we split β into

$$\beta_{\mathcal{Q}} := K_{w,\lambda}^{-1} \Phi_{\mathcal{Q}} b_{\mathcal{Q}}, \quad \beta_{\mathcal{S}_m^1} := K_{w,\lambda}^{-1} \Phi_{\mathcal{S}_m^1} b_{\mathcal{S}_m^1}, \quad \text{and} \quad \beta_\varepsilon := K_{w,\lambda}^{-1} \varepsilon,$$

then we obtain the bound

$$\left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} - (\rho + \lambda)^{-1} d^{(p+\delta-\|\boldsymbol{m}\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2$$
$$\leq \left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2$$
$$+ \left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} - (\rho + \lambda)^{-1} d^{(p+\delta-\|\boldsymbol{m}\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2$$
$$+ \left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2^2. \tag{372}$$

For the first term, we apply the following proposition. The proof is deferred to Appendix C.5.7.

Proposition C.61. *For any $\epsilon > 0$, the following estimate holds uniformly over all coordinates $r \in \mathcal{T}_\iota$:*

$$\left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2^2 = O_{d,\mathbb{P}}(d^{-\lambda_\iota + \epsilon}) \cdot \|b_{\mathcal{Q}}\|_2^2. \tag{373}$$

Now we proceed to analyze the second term. After rescaling and multiplying both terms on the left side of the equation (323) by $(W_{\mathcal{S}_m^1})^{1/2}$, we yield

$$\left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{S}_m^1}^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} - (\rho + \lambda)^{-1} d^{(p+\delta-\|\boldsymbol{m}\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2$$
$$= d^{(p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m})/2} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2. \tag{374}$$

Next, we show that the third term is small. We defer the proof of the following proposition to Appendix C.5.8.

Proposition C.62. *For any $\epsilon > 0$, the following estimate holds uniformly for all $r \in \mathcal{T}_\iota$:*

$$\left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_\varepsilon^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2^2 = O_{d,\mathbb{P}}(d^{-\lambda_\iota + \epsilon}) \cdot \sigma_\varepsilon^2 \tag{375}$$

Lastly, plugging Eq. (373), (374) and (375) into (372) yields:

$$\left\| n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} - (\rho + \lambda)^{-1} d^{(p+\delta-\|\boldsymbol{m}\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2$$
$$\leq d^{(p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m})/2} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2 + \epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota/2 + \epsilon/2}) \cdot \left(\|b_{\mathcal{Q}}\|_2 + \sigma_\varepsilon \right).$$

116

Then combining the above bound with Eq. (371) gives

$$\left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 - (\rho + \lambda)^{-1} \sqrt{g^{(\|m\|_1)}(0)} \cdot d^{(p+\delta-\|m\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top (W_{\mathcal{S}_m^1})^{\frac{1}{2}} \right\|_2$$

$$\overset{(a)}{\leq} \left\| \beta^\top \Phi_{\mathcal{S}_m^1} R_m^1 - \beta^\top \Phi_{\mathcal{S}_m^1} \left(n^{-1} g^{(\|m\|_1)}(0) W_{\mathcal{S}_m^1} d^{-\|m\|_1} \right)^{\frac{1}{2}} \right\|_2$$

$$+ \sqrt{g^{(\|m\|_1)}(0)} \cdot \left\| n^{-\frac{1}{2}} d^{-\frac{\|m\|_1}{2}} \beta^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} - (\rho + \lambda)^{-1} d^{(p+\delta-\|m\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top (W_{\mathcal{S}_m^1})^{\frac{1}{2}} \right\|_2$$

$$\overset{(b)}{\leq} O_d(d^{-\frac{1}{2}}) \cdot \left\| (\rho + \lambda)^{-1} \sqrt{g^{(\|m\|_1)}(0)} \cdot d^{(p+\delta-\|m\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top (W_{\mathcal{S}_m^1})^{\frac{1}{2}} \right\|_2 + \left(1 + O_d(d^{-\frac{1}{2}}) \right)$$

$$\cdot \left(d^{(p+\delta-\lambda^\top m)/2} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota/2+\epsilon/2}) \cdot (\|b_{\mathcal{Q}}\|_2 + \sigma_\varepsilon) \right)$$

$$= O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2 + O_{d,\mathbb{P}}(d^{-\lambda_\iota/2+\epsilon/2}) \cdot (\|b_{\mathcal{Q}}\|_2 + \sigma_\varepsilon),$$

where both (a) and (b) follow from the triangle inequality.

Following a similar argument with the proof of Proposition B.32, we can derive a bound for

$$\left| \left\| n^{-\frac{1}{2}} d^{-\frac{\|m\|_1}{2}} \beta^\top \Phi_{\mathcal{S}_m^1} \left(W_{\mathcal{S}_m^1} \right)^{\frac{1}{2}} \right\|_2^2 - \left\| (\rho + \lambda)^{-1} d^{(p+\delta-\|m\|_1)/2} \cdot b_{\mathcal{S}_m^1}^\top (W_{\mathcal{S}_m^1})^{\frac{1}{2}} \right\|_2^2 \right|, \tag{376}$$

which takes the form

$$O_{d,\mathbb{P}}(d^{-\lambda_\iota+\epsilon}) \cdot (\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2) + d^{p+\delta-\lambda^\top m} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2^2$$

$$+ O_{d,\mathbb{P}}(d^{(p+\delta-\lambda^\top m)/2 - \lambda_\iota/2 + \epsilon/2}) \cdot \left\| b_{\mathcal{S}_m^1} \right\|_2 (\|f^*\|_{L_2} + \sigma_\varepsilon),$$

where we use the fact $\|b_{\mathcal{Q}}\|_2 \leq \|f^*\|_{L_2}$.

Combining with the fact $\left\| b_{\mathcal{S}_m^1} \right\|_2 = \left\| \partial_r f_{\mathcal{S}_m^1}^* \right\|_{L_2}$ completes the proof.

C.5.3 Proof of Proposition C.58

The proof is analogous to the proof of Proposition C.51. First, following Eq. (327) we write

$$K_{w,\lambda} = \Phi_{\mathcal{A}_{\leq p_\delta}} \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top + (\rho + \lambda) H, \tag{377}$$

where $H := I_n + (\rho + \lambda)^{-1} \Delta_1$.

Then following Eq. (329), we have

$$\beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R_m = \frac{1}{n} b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R_m \right). \tag{378}$$

With the decomposition $\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R_m = (\widehat{E}_r)_{\mathcal{S}_m} + (\widehat{E}_r^\circ)_{\mathcal{S}_m}$, Eq. (364) and (334) together give

$$\left\| G^{-1} \widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R_m - V(\widehat{E}_r)_{\mathcal{S}_m} \right\|_{\text{op}} \leq \|G^{-1}\|_{\text{op}} \left\| (\widehat{E}_r^\circ)_{\mathcal{S}_m} \right\|_{\text{op}} + \|G^{-1} - V\|_{\text{op}} \left\| (\widehat{E}_r)_{\mathcal{S}_m} \right\|_{\text{op}}$$

$$= O_{d,\mathbb{P}}(w_r \cdot d^{(\lambda^\top m - p - \delta - 2)/2} + d^{-\zeta/2+\epsilon+(\lambda^\top m - p - \delta)/2})$$

$$= O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\lambda^\top m - p - \delta)/2}),$$

where we use the fact $\zeta \leq \lambda_\iota$.

Corresponding to Eq. (336), we obtain

$$\left\| \beta_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R_m - \frac{1}{n} b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} V(\widehat{E}_r)_{\mathcal{S}_m} \right\|_2 = O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon+(\lambda^\top m - p - \delta)/2}) \cdot \left\| b_{\mathcal{E}(r)} \right\|_2.$$

Note that Eq. (338) correspondingly becomes

$$\left\|\frac{1}{n}b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top H^{-1}\Phi_{\mathcal{A}_{\leq p_\delta}}V(\widehat{E}_r)_{\mathcal{S}_m} - \frac{1}{n}b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2$$

$$= O_{d,\mathbb{P}}\left(\left\|b_{\mathcal{E}(r)}^\top V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2\right) = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta-2)/2})\cdot\left\|b_{\mathcal{E}(r)}\right\|_2.$$

Then invoking the bound $\left\|\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}/n - I\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon})$ (Theorem A.12), we have

$$\left\|\frac{1}{n}b_{\mathcal{E}(r)}^\top \Phi_{\mathcal{E}(r)}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}V(\widehat{E}_r)_{\mathcal{S}_m} - b_{\mathcal{A}_{\leq p_\delta}}^\top V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2 = \left\|\frac{1}{n}b_{\mathcal{A}_{\leq p_\delta}}^\top E_r \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}V(\widehat{E}_r)_{\mathcal{S}_m} - b_{\mathcal{A}_{\leq p_\delta}}^\top V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2$$

$$\leq \left\|b_{\mathcal{A}_{\leq p_\delta}}^\top E_r \left(\Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}}/n - I\right)V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2$$

$$\leq \left\|b_{\mathcal{A}_{\leq p_\delta}}^\top E_r V(\widehat{E}_r)_{\mathcal{S}_m}\right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon})$$

$$= \left\|b_{\mathcal{E}(r)}\right\|_2 \cdot O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon+(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta)/2}).$$

Then the remainder of the analysis is identical to the proof of Proposition C.51.

C.5.4 Proof of Claim 19

Note that R contains blocks $R_{\mathcal{S}_m^1}$ and $R_{\mathcal{S}_m^0}$. We discuss each kind of block separately.

Also recall that D is a diagonal matrix with entry $D_{\mathcal{S}_m} = g^{(\|\boldsymbol{m}\|_1)}(0)d^{-\|\boldsymbol{m}\|_1} + O_d(d^{-\|\boldsymbol{m}\|_1-1})$ and each entry of $W_{\mathcal{S}_m}$ is of the order $d^{\|\boldsymbol{m}\|_1-\boldsymbol{\lambda}^\top \boldsymbol{m}}$ based on Eq. (118). We will use these two properties in the analysis.

For $S \in \mathcal{S}_m^1$ or $\mathcal{S}_m^0$, based on the definition of D Eq. (108) and W Eq. (118) we have

$$(\widehat{D})_{S,S}^{-1} = (DW)_{S,S}^{-1} = \left(g^{(\|\boldsymbol{m}\|_1)}(0)\right)^{-1}W_{S,S}^{-1}d^{\|\boldsymbol{m}\|_1} + O_d(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m}-1}).$$

Since Eq. (131) gives

$$R_{S,S} = \sqrt{g^{(\|\boldsymbol{m}\|_1)}(0)}\cdot W_{S,S}^{\frac{1}{2}}d^{-(\|\boldsymbol{m}\|_1+p+\delta)/2} + O_d(d^{-(\boldsymbol{\lambda}^\top \boldsymbol{m}+p+\delta+1)/2}),$$

by a straightforward calculation, we obtain

$$(\widehat{D}^{-1}R)_{S,S} = \left(g^{(\|\boldsymbol{m}\|_1)}(0)\right)^{-\frac{1}{2}}W_{S,S}^{-\frac{1}{2}}d^{(\|\boldsymbol{m}\|_1-p-\delta)/2}\cdot(1+O_d(d^{-\frac{1}{2}})).$$

Similarly, for $S \in \mathcal{S}_m^0$, Eq. (130) gives

$$R_{S,S} = \sqrt{g^{(\|\boldsymbol{m}\|_1+2)}(0)}\cdot w_r \cdot W_{S,S}^{\frac{1}{2}}d^{-(\|\boldsymbol{m}\|_1+p+\delta+2)/2} + w_r \cdot O_d(d^{-(\boldsymbol{\lambda}^\top \boldsymbol{m}+p+\delta+3)/2}).$$

From this, we derive

$$(\widehat{D}^{-1}R)_{S,S} = w_r \cdot O_d(d^{(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta-2)/2}).$$

Combining these two cases yields

$$(\widehat{D}^{-1}R)_{S,S} = \begin{cases} \left(g^{(\|\boldsymbol{m}\|_1)}(0)\right)^{-\frac{1}{2}}W_{S,S}^{-\frac{1}{2}}d^{(\|\boldsymbol{m}\|_1-p-\delta)/2}\cdot(1+O_d(d^{-\frac{1}{2}})) & \text{if } S \in \cup_{\boldsymbol{m}}\mathcal{S}_m^1; \\ w_r \cdot O_d(d^{(\boldsymbol{\lambda}^\top \boldsymbol{m}-p-\delta-2)/2}), & \text{if } S \in \cup_{\boldsymbol{m}}\mathcal{S}_m^0; \\ 0 & \text{otherwise.} \end{cases} \tag{379}$$

Combining this with the expression E_r:

$$(E_r)_{S,S} = \mathbb{1}\left\{S \in \cup_{\boldsymbol{m}}\mathcal{S}_m^1\right\}, \tag{380}$$

we complete the proof.

118

C.5.5 Proof of Proposition C.59

The proof is analogous to that of Proposition C.52. First, following Eq. (344) we have

$$\beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R = \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R \right). \tag{381}$$

Following Claim 19 which shows $\left\| \widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R - \widehat{E}_r \right\|_{\mathrm{op}} = \left\| \widehat{E}_r^\circ \right\|_{\mathrm{op}} = O_d(w_r \cdot d^{-1})$ we can obtain

$$\left\| \beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R - \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \widehat{E}_r \right\|_2 = O_{d,\mathbb{P}}(w_r \cdot d^{-1}) \cdot \log(d) \left\| b_{\mathcal{E}^c} \right\|_2, \tag{382}$$

where we use the fact that $\left\| H^{-1} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (330)), $\left\| G^{-1} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (Eq. (333)), and $\left\| \Phi_{\mathcal{E}} b_{\mathcal{E}^c} \right\|_2 = O_{d,\mathbb{P}}(\sqrt{n} \log(d)) \left\| b_{\mathcal{E}^c} \right\|_2$ (Proposition B.21).

Consequently, there exists Δ_1 that satisfies $\left\| \Delta_1 \right\|_2 = O_{d,\mathbb{P}}(w_r \cdot d^{-1} \log(d)) \cdot \left\| f_{\mathcal{E}^c}^* \right\|_{L_2}$ such that the following holds:

$$\beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R = \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \widehat{E}_r + \Delta_1. \tag{383}$$

Comparing $\widehat{E}_r$ with E_r appearing in Eq. (346), they have the same support. By Claim 19, we have each non-zero entry of $\widehat{E}_r$ is of the order $O_d(1)$. Therefore, following the identical analysis in the proof of Proposition C.52, we can yield the bound

$$\begin{aligned} \left\| \beta_{\mathcal{E}^c}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 &\leq 2 \left\| \frac{1}{n} b_{\mathcal{E}^c}^\top \Phi_{\mathcal{E}^c}^\top H^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} G^{-1} \widehat{E}_r \right\|_2^2 + 2 \left\| \Delta_1 \right\|_2^2 \\ &= \left(O_{d,\mathbb{P}}(d^{-\lambda_\iota - \zeta_1 + \epsilon}) + O_{d,\mathbb{P}}(d^{-1}) + O_{d,\mathbb{P}}(w_r^2 d^{-2+\epsilon}) \right) \cdot \left\| f_{\mathcal{E}^c}^* \right\|_{L_2}^2 \\ &= O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \cdot \left\| f_{\mathcal{E}^c}^* \right\|_{L_2}^2 \end{aligned}$$

Thus we conclude the proof.

C.5.6 Proof of Proposition C.60

The proof is almost identical to that of Proposition C.53. The only difference appears in the bound Eq. (352) which becomes

$$\mathrm{Tr} \left(\widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} R R^\top \widehat{D}_{\mathcal{A}_{\leq p_\delta}}^{-1} \right) = O_d \left(\left\| \widehat{E}_r \right\|_F^2 + \left\| \widehat{E}_r^\circ \right\|_F^2 \right) = O_d(d^{p+\delta-\lambda_\iota-\zeta_1}) + O_d(d^{p+\delta-1}). \tag{384}$$

Then, applying a similar analysis to the proof of Proposition C.53, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \beta_\varepsilon^\top \Phi_{\mathcal{A}_{\leq p_\delta}} R \right\|_2^2 \right] = O_d(d^{-\lambda_\iota - \zeta_1} + d^{-1}).$$

Since we have $\zeta_1 > 0$, applying Markov's inequality completes the proof.

C.5.7 Proof of Proposition C.61

The proof is analogous to that of Proposition C.54, and throughout we follow the conventions and notation established there.

Multiplying both sides of the equation (355) by $\frac{n}{d^{\|\boldsymbol{m}\|_1}}$ and taking $W_{\mathcal{S}_{\boldsymbol{m}}^1}$ into account, which satisfies $\left\| W_{\mathcal{S}_{\boldsymbol{m}}^1} \right\|_{\mathrm{op}} = O_d(d^{\|\boldsymbol{m}\|_1 - \boldsymbol{\lambda}^\top \boldsymbol{m}})$ yields:

$$\begin{aligned} \left\| \beta_{\mathcal{Q}}^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}} \left(n^{-1} W_{\mathcal{S}_{\boldsymbol{m}}^1} d^{-\|\boldsymbol{m}\|_1} \right)^{\frac{1}{2}} \right\|_2^2 &= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - p - \delta + 1/q_0 + (p+\delta) - \|\boldsymbol{m}\|_1 + \|\boldsymbol{m}\|_1 - \boldsymbol{\lambda}^\top \boldsymbol{m}}) \cdot \log(d) \cdot \left\| b_{\mathcal{Q}} \right\|_2^2, \\ &= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - \boldsymbol{\lambda}^\top \boldsymbol{m} - s_\iota + 1/q_0}) \cdot \log(d) \cdot \left\| b_{\mathcal{Q}} \right\|_2^2 \\ &= O_{d,\mathbb{P}}(d^{-\lambda_\iota + 1/q_0}) \cdot \log(d) \cdot \left\| b_{\mathcal{Q}} \right\|_2^2, \end{aligned}$$

where in the last equality we replace s_ι by λ_ι since $m_\iota \geq 1$ and we use the fact $\boldsymbol{s}^\top \boldsymbol{m} - \boldsymbol{\lambda}^\top \boldsymbol{m} \leq s_\iota - \lambda_\iota \leq 0$.

Absorbing the factor $d^{1/q_0} \log(d)$ into d^ϵ where $\epsilon > 0$ is an arbitrarily small constant yields the desired bound.

C.5.8 Proof of Proposition C.62

The proof is almost identical to that of Proposition C.56.

In parallel to Eq. (360), we can obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\|n^{-\frac{1}{2}} d^{-\frac{\|\boldsymbol{m}\|_1}{2}} \beta_{\mathcal{S}_{\boldsymbol{m}}^1}^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^1} \left(W_{\mathcal{S}_{\boldsymbol{m}}^1}\right)^{\frac{1}{2}}\right\|_2^2\right] \leq \left\|W_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_{\mathrm{op}} \cdot d^{-\|\boldsymbol{m}\|_1 - p - \delta} \cdot \mathbb{E}\left[\left\|\beta_\varepsilon^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\right\|_2^2\right]$$

$$\leq O_d(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m} - p - \delta}) \cdot \left\|K_{w,\lambda}^{-2}\right\|_{\mathrm{op}} \mathrm{Tr}\left(\Phi_{\mathcal{S}_{\boldsymbol{m}}^1}^\top \Phi_{\mathcal{S}_{\boldsymbol{m}}^1}\right). \tag{385}$$

Then the remainder of the proof follows exactly with the proof of Proposition C.56 and we can get the bound

$$O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m} - s_\iota - \boldsymbol{\lambda}^\top \boldsymbol{m} - p - \delta + p + \delta}) = O_{d,\mathbb{P}}(d^{-\lambda_\iota}).$$

Applying Markov's inequality we obtain the bound $O_{d,\mathbb{P}}(d^{-\lambda_\iota}) \log(d)$ with probability at least $1 - 1/\log(d)$. Absorbing the logarithmic factor into ϵ completes the proof.

C.6 Proof of Propositions for Theorem B.27 (Generalization bound)

C.6.1 Proof of Lemma B.35

For kernel $K_w(x, y) = g(\frac{x^\top \mathsf{Diag}(w) y}{d})$, we can associate a self adjoint operator $\mathcal{H}_d : L_2(\mathbb{H}^d, \tau_d) \to L_2(\mathbb{H}^d, \tau_d)$:

$$(\mathcal{H}_d f)(x) = \mathbb{E}_y \left[K_w(x, y) f(y)\right]. \tag{386}$$

We denote the Fourier basis by $\phi_{\mathcal{I}}(x) := x^{\mathcal{I}}$ for any multi-index $\mathcal{I} \in [d]$ and denote $z_i = x_i y_i$ for $i = 1, \ldots, d$. Then we have

$$(\mathcal{H}_d \phi_{\mathcal{I}})(x) = \mathbb{E}_y \left[K(x, y) \phi_{\mathcal{I}}(y)\right]$$

$$\overset{(a)}{=} \frac{1}{2^d} \sum_{z \in \mathbb{H}^d} \left(g\left(\frac{w^\top z}{d}\right) \phi_{\mathcal{I}}(x) \phi_{\mathcal{I}}(z)\right)$$

$$= \underbrace{\frac{1}{2^d} \sum_{z \in \mathbb{H}^d} \left(g\left(\frac{w^\top z}{d}\right) \phi_{\mathcal{I}}(z)\right)}_{\lambda_{\mathcal{I}}} \phi_{\mathcal{I}}(x),$$

where (a) follows from the fact $y^{\mathcal{I}} = x^{\mathcal{I}} z^{\mathcal{I}}$.

Now we analyze the eigenvalue $\lambda_{\mathcal{I}}$. We take the Taylor expansion of $g\left(\frac{w^\top z}{d}\right)$ around 0 and we obtain

$$\lambda_{\mathcal{I}} = \sum_{m=0}^\infty \frac{g^{(m)}(0)}{m!} \underbrace{\frac{1}{2^d} \sum_{z \in \mathbb{H}^d} \left(\frac{w^\top z}{d}\right)^m \phi_{\mathcal{I}}(z)}_{E_{m,\mathcal{I}}}. \tag{387}$$

We now consider

$$E_{m,\mathcal{I}} = \mathbb{E}_z \left[\left(\frac{w^\top z}{d}\right)^m \phi_{\mathcal{I}}(z)\right] = \frac{1}{d^m} \mathbb{E}_z \left[\left(\sum_{j_1, \ldots, j_m \in [d]^m} (w_{j_1} z_{j_1}) \ldots (w_{j_m} z_{j_m})\right) \phi_{\mathcal{I}}(z)\right].$$

Denoting $r = |\mathcal{I}|$, with simple calculation, we observe that $E_{m,\mathcal{I}} = 0$ unless $m - r \geq 0$ and $(m - r)$ is even. Since $\|w\|_\infty = O_d(d^\gamma)$ with $\gamma < 1$, we have

$$|E_{m,\mathcal{I}}| = O_d(d^{-m} \|w\|_\infty^{m-r} |w^\mathcal{I}|) = o_d(d^{-m+(m-r)\gamma})|w^\mathcal{I}|.$$

Therefore, the series is dominated by $m = r$ and we can conclude

$$E_{m,\mathcal{I}} = \frac{m!}{d^m} w^\mathcal{I} + O_d(d^{-m-2}) \cdot w^\mathcal{I} \cdot d^{2\gamma}, \tag{388}$$

where the main term comes from matching r indices of $\mathcal{I}$ and the error is driven by the first non-vanishing $m = r + 2$ contribution.

Plugging (388) back into Eq. (387) yields

$$\lambda_\mathcal{I} = g^{(|\mathcal{I}|)}(0)d^{-|\mathcal{I}|}w^\mathcal{I} + O_d(d^{-|\mathcal{I}|-2+2\gamma}) \cdot w^\mathcal{I}.$$

Then writing $K_w(x,y)$ as the summation of $\lambda_\mathcal{I}\phi_\mathcal{I}(x)\phi_\mathcal{I}(y)$ over all $\mathcal{I}$ completes the proof.

C.6.2 Proof of Proposition B.36

Invoking Lemma B.35 we can derive the decomposition

$$\mathbb{E}_x[K(Z, \sqrt{w} \odot x)K(\sqrt{w} \odot x, Z)] = \sum_{k=0}^\infty \sum_{\boldsymbol{m}: \|\boldsymbol{m}\|_1 = k} \Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top, \tag{389}$$

where $\Lambda_{\mathcal{S}_m}$ is a diagonal matrix with entries $\Lambda_{S,S} = g^{-\|\boldsymbol{m}\|_1}(0)d^{-\|\boldsymbol{m}\|_1}w^S + O_d(d^{-\|\boldsymbol{m}\|_1-2+2\gamma}) \cdot w^S$ for $S \in \mathcal{S}_{\boldsymbol{m}}$.

We split Eq. (389) into two parts based on $\boldsymbol{m}$: $\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p + \delta$ and $\boldsymbol{\lambda}^\top \boldsymbol{m} > p + \delta$:

$$\sum_{k=0}^\infty \sum_{\boldsymbol{m}: |\boldsymbol{m}| = k} \Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top$$
$$= \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \leq p+\delta} \Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top + \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} > p+\delta} \Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top. \tag{390}$$

Next, we show that the second term is small. Similar to Theorem B.28, we split the set:

$$\{\boldsymbol{m} \mid \boldsymbol{\lambda}^\top \boldsymbol{m} > p + \delta\} = \{\boldsymbol{m} \mid \boldsymbol{\lambda}^\top \boldsymbol{m} \geq 2p + 3\delta + 1\} \bigsqcup \{\boldsymbol{m} \mid p + \delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p + 3\delta + 1\}$$

and we apply different techniques for each set.

First, note the Frobeniusm norm of $\Phi_{\mathcal{S}_m}(X)$ can be bound as

$$\|\Phi_{\mathcal{S}_m}(X)\|_F^2 = n \cdot d^{\boldsymbol{s}^\top \boldsymbol{m}}$$

Then for $\boldsymbol{m} \in \{\boldsymbol{m} \mid \boldsymbol{\lambda}^\top \boldsymbol{m} \geq 2p + 3\delta + 1\}$ we have the bound

$$\begin{aligned}
\left\|\Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top\right\|_{\mathrm{op}} &\leq \|\Phi_{\mathcal{S}_m}(X)\|_F^2 \cdot \|\Lambda_{\mathcal{S}_m}\|_{\mathrm{op}}^2 \\
&= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m}} \cdot d^{p+\delta})O_d(d^{-2\|\boldsymbol{m}\|_1}) \cdot O_d(d^{2(\mathbf{1}-\boldsymbol{\lambda})^\top \boldsymbol{m}}) \\
&= O_{d,\mathbb{P}}(d^{p+\delta+(\boldsymbol{s}-2\boldsymbol{\lambda})^\top \boldsymbol{m}}) = O_{d,\mathbb{P}}(d^{p+\delta-\boldsymbol{\lambda}^\top \boldsymbol{m}}),
\end{aligned} \tag{391}$$

where the last inequality follows from the fact that $s_k \leq \lambda_k$ for all $k \in [N]$.

We proceed to use triangle inequality to bound the summation

$$\left\| \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \geq 2p+3\delta+1} \Phi_{\mathcal{S}_m}(Z)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(Z)^\top \right\|_{\mathrm{op}} \leq \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \geq 2p+3\delta+1} \left\| \Phi_{\mathcal{S}_m}(Z)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(Z)^\top \right\|_{\mathrm{op}}$$

$$\leq \sum_{k=2p}^{\infty} \sum_{\boldsymbol{m}:\boldsymbol{\lambda}^\top \boldsymbol{m} \in [k+3\delta+1,k+3\delta+2)} \left\| \Phi_{\mathcal{S}_m}(Z)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(Z)^\top \right\|_{\mathrm{op}}$$

$$\overset{(a)}{\leq} \sum_{k=2p}^{\infty} O_d((k+5)^N) O_{d,\mathbb{P}}(d^{p-k-2\delta-1})$$

$$= \sum_{k=0}^{\infty} O_{d,\mathbb{P}}(d^{-p-2\delta-1-k+N\log_d(2p+k+5)}), \tag{392}$$

where (a) follows from the fact that all $\boldsymbol{m} \in \mathbb{R}^N$ that satisfies $\boldsymbol{\lambda}^\top \boldsymbol{m} \in [k+3\delta+1, k+3\delta+2)$ is contained in the set $\{\boldsymbol{m} : \|\boldsymbol{m}\|_1 \leq (k+5)/\|\boldsymbol{\lambda}\|_\infty\}$, whose cardinality is bounded by $O_d((k+5)^N)$.

Now we consider the sequence

$$\{d^{-p-2\delta-1-k+N\log_d(2p+k+5)}\}_{k=0}^{\infty}.$$

Observe that the ratio r of consecutive terms satisfies

$$r \leq d^{-1+N\log_d((2p+6)/(2p+5))}.$$

Therefore, the summation of the sequence is bounded by the summation a geometric series with first term $d^{-p-2\delta-1+N\log_d(2p+5)}$ and a common ratio $d^{-1+N\log_d((2p+6)/(2p+5))}$.

Using the formula to compute the summation of the geometric series we can show

$$\sum_{k=0}^{\infty} O_{d,\mathbb{P}}(d^{-p-2\delta-1-k+N\log_d(2p+k+5)}) = O_{d,\mathbb{P}}(d^{-p-2\delta-1+N\log_d(2p+5)}).$$

Since $\delta \gtrsim N\log_d(2p+5)$, plugging the bound back into Eq. (392) yields

$$\left\| \sum_{\boldsymbol{\lambda}^\top \boldsymbol{m} \geq 2p+3\delta+1/2} \Phi_{\mathcal{S}_m}(Z)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(Z)^\top \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-p-\delta-1}). \tag{393}$$

Next we consider the set $\{\boldsymbol{m} \mid p+\delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p+3\delta+1\}$. Lemma A.11 shows the following hold for any $\boldsymbol{m} \in \{\boldsymbol{m} \mid p+\delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p+3\delta+1\}$

$$\|\Phi_{\mathcal{S}_m}(X)\|_{\mathrm{op}}^2 = O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m}} + d^{p+\delta}). \tag{394}$$

Then for each $\boldsymbol{m}$ we have the bound

$$\left\| \Phi_{\mathcal{S}_m}(X)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(X)^\top \right\|_{\mathrm{op}} \leq \|\Phi_{\mathcal{S}_m}(X)\|_{\mathrm{op}}^2 \cdot \|\Lambda_{\mathcal{S}_m}\|_{\mathrm{op}}^2$$

$$= O_{d,\mathbb{P}}(d^{\boldsymbol{s}^\top \boldsymbol{m}} + d^{p+\delta}) O_d(d^{-2\|\boldsymbol{m}\|_1}) \cdot O_d(d^{2(\boldsymbol{1}-\boldsymbol{\lambda})^\top \boldsymbol{m}})$$

$$= O_{d,\mathbb{P}}(d^{(\boldsymbol{s}-2\boldsymbol{\lambda})^\top \boldsymbol{m}} + d^{p+\delta-2\boldsymbol{\lambda}^\top \boldsymbol{m}}) = O_{d,\mathbb{P}}(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m}}). \tag{395}$$

Invoking the triangle inequality yields:

$$\left\| \sum_{p+\delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p+3\delta+1} \Phi_{\mathcal{S}_m}(Z)\Lambda_{\mathcal{S}_m}^2 \Phi_{\mathcal{S}_m}(Z)^\top \right\|_{\mathrm{op}} \leq \sum_{p+\delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p+3\delta+1} O_{d,\mathbb{P}}(d^{-\boldsymbol{\lambda}^\top \boldsymbol{m}}) \tag{396}$$

$$= O_{d,\mathbb{P}}(d^{-p-\delta-\zeta_2}), \tag{397}$$

122

where the second equality follows from the fact that the cardinality of the set $\{\boldsymbol{m} \mid p + \delta < \boldsymbol{\lambda}^\top \boldsymbol{m} < 2p + 3\delta + 1\}$ is finite.

Combining Eq. (393) and (396) with (390) yields:

$$\mathbb{E}_x[K(Z, \sqrt{w} \odot x)K(\sqrt{w} \odot x, Z)] = \Phi_{\mathcal{A}_{\leq p_\delta}}(X)\Lambda_{\mathcal{A}_{\leq p_\delta}}^2 \Phi_{\mathcal{A}_{\leq p_\delta}}(X)^\top + \Delta,$$

with $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-p-\delta-\zeta_2})$, which completes the proof.

$\square$

C.6.3   Proof of Proposition B.37

We define $\beta := K_{w,\lambda}^{-1} y$ and decompose β into

$$\beta = \underbrace{K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} b_{\mathcal{A}_{\leq p_\delta}}}_{:=\beta_{\mathcal{A}_{\leq p_\delta}}} + \underbrace{K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{> p_\delta}} b_{\mathcal{A}_{> p_\delta}}}_{:=\beta_{\mathcal{A}_{> p_\delta}}} + \underbrace{K_{w,\lambda}^{-1} \varepsilon}_{:=\beta_\varepsilon} . \tag{398}$$

and we have the decomposition of the absolute difference

$$\left| T_{11} - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right|$$
$$\leq \left| \beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}} \beta_{\mathcal{A}_{\leq p_\delta}} - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right| + \left| \beta_{\mathcal{A}_{> p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}} \beta_{\mathcal{A}_{> p_\delta}} \right| + \left| \beta_\varepsilon^\top M_{\mathcal{A}_{\leq p_\delta}} \beta_\varepsilon \right|$$
$$+ 2\left| \beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}} \beta_{\mathcal{A}_{> p_\delta}} \right| + 2\left| (\beta_{\mathcal{A}_{\leq p_\delta}} + \beta_{\mathcal{A}_{> p_\delta}})^\top M_{\mathcal{A}_{\leq p_\delta}} \beta_\varepsilon \right| . \tag{399}$$

For the bound above, we will show that the first three terms are small hence the last two terms are also small followed by Cauchy-Schwartz inequality.

We first establish the following key estimates. The proof appears in Appendix C.6.5.

Proposition C.63. *For any $\epsilon > 0$, the following estimates hold:*

$$\left\| nK_{w,\lambda}^{-1} M_{\mathcal{A}_{\leq p_\delta}} K_{w,\lambda}^{-1} - \frac{1}{n} \Phi_{\mathcal{A}_{\leq p_\delta}} V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + d^{-2+2\gamma_{\max}}) \tag{400}$$

$$\left\| \Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} M_{\mathcal{A}_{\leq p_\delta}} K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} - V^2 \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + d^{-2+2\gamma_{\max}}), \tag{401}$$

where $\gamma_{\max} := \max_{k \in [N]} \gamma_k$.

Following Eq. (106), we denote $\zeta := \min(\zeta_1, \zeta_2)$ to simplify the notation. Then we can immediately obtain that the first term in Eq. (399) is bounded by

$$\left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \cdot \left\| V^2 - \Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} M_{\mathcal{A}_{\leq p_\delta}} K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \right\|_{\mathrm{op}}$$
$$= O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}}) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \tag{402}$$

For the second term in Eq. (399), invoking Proposition C.63 we can obtain the estimate

$$\left| y_{\mathcal{A}_{> p_\delta}}^\top K_{w,\lambda}^{-1} M_{\mathcal{A}_{\leq p_\delta}} K_{w,\lambda}^{-1} y_{\mathcal{A}_{> p_\delta}} - \left\| \frac{1}{n} \Phi_{\mathcal{A}_{\leq p_\delta}} y_{\mathcal{A}_{> p_\delta}} \right\|_2^2 \right|$$
$$\leq O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}}) \cdot \frac{1}{n} \left\| y_{\mathcal{A}_{> p_\delta}} \right\|_2^2$$
$$= O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}}) \cdot \log(d) \cdot \left\| f_{\mathcal{A}_{> p_\delta}}^* \right\|_{L_2}^2 , \tag{403}$$

where the last equality follows from Proposition B.21.

123

Similar to the analysis to Eq. (224), we can see that

$$
\mathbb{E}\left[\left\|\frac{1}{n}\Phi_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{>p_\delta}}b_{\mathcal{A}_{>p_\delta}}\right\|_2^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\langle\Phi_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}^\top, \Phi_{\mathcal{A}_{>p_\delta}}b_{\mathcal{A}_{>p_\delta}}b_{\mathcal{A}_{>p_\delta}}^\top\Phi_{\mathcal{A}_{>p_\delta}}^\top\right\rangle\right]
$$

$$
= \frac{|\mathcal{A}_{\leq p_\delta}|}{n}\left\|b_{\mathcal{A}_{>p_\delta}}\right\|_2^2 = O_d(d^{-\zeta_1})\cdot\left\|b_{\mathcal{A}_{>p_\delta}}\right\|_2^2.
$$

Applying Markov's inequality we obtain

$$
\left\|\frac{1}{n}\Phi_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{>p_\delta}}b_{\mathcal{A}_{>p_\delta}}\right\|_2^2 = O_{d,\mathbb{P}}(d^{-\zeta_1}\log(d))\cdot\left\|b_{\mathcal{A}_{>p_\delta}}\right\|_2^2.
$$

Plugging it back into Eq. (403) yields

$$
\left|y_{\mathcal{A}_{>p_\delta}}^\top K_{w,\lambda}^{-1}M_{\mathcal{A}_{\leq p_\delta}}K_{w,\lambda}^{-1}y_{\mathcal{A}_{\leq p_\delta}}\right| = O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})\cdot\log(d)\cdot\left\|f_{\mathcal{A}_{>p_\delta}}^*\right\|_{L_2}^2. \tag{404}
$$

For the third term in Eq. (399), taking the expectation over $\boldsymbol{\varepsilon}$ yields

$$
\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\beta_\varepsilon^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right] = \sigma_\varepsilon^2\cdot\mathrm{Tr}\left(K_{w,\lambda}^{-1}M_{\mathcal{A}_{\leq p_\delta}}K_{w,\lambda}^{-1}\right)
$$

$$
\overset{(a)}{=} \frac{\sigma_\varepsilon^2}{n}\left(\mathrm{Tr}\left(\frac{1}{n}\Phi_{\mathcal{A}_{\leq p_\delta}}V^2\Phi_{\mathcal{A}_{\leq p_\delta}}^\top\right) + n\cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})\right)
$$

$$
\overset{(b)}{\leq} \frac{\sigma_\varepsilon^2}{n}\mathrm{Tr}\left(\frac{1}{n}\Phi_{\mathcal{A}_{\leq p_\delta}}^\top\Phi_{\mathcal{A}_{\leq p_\delta}}\right)\cdot\|V\|_{\mathrm{op}}^2 + \sigma_\varepsilon^2\cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})
$$

$$
\overset{(c)}{\leq} O_d(d^{-\zeta_1})\cdot\sigma_\varepsilon^2 + \sigma_\varepsilon^2\cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})
$$

$$
= \sigma_\varepsilon^2\cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}}),
$$

where (a) follows from Proposition C.63, (b) follow from the inequality $|\mathrm{Tr}(AB)| \leq \mathrm{Tr}(A)\|B\|_{\mathrm{op}}$ and (c) follows from the bound $\|V\|_{\mathrm{op}} = O_d(1)$ (Eq. (312)) and the equality $\mathrm{Tr}\left(\frac{1}{n}\Phi_{\mathcal{A}_{\leq p_\delta}}^\top\Phi_{\mathcal{A}_{\leq p_\delta}}\right) = \frac{1}{n}\left\|\Phi_{\mathcal{A}_{\leq p_\delta}}\right\|_F^2 = |\mathcal{A}_{\leq p_\delta}|$.

Using Markov's inequality, we conclude the bound

$$
\left|\beta_\varepsilon^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right| = \sigma_\varepsilon^2\cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}}). \tag{405}
$$

With Eq. (402) and (404), we apply Cauchy-Schwarz inequality for the fourth term in Eq. (399)

$$
\left|\beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{>p_\delta}}\right| \leq \sqrt{\left|\beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{\leq p_\delta}}\right|}\sqrt{\left|\beta_{\mathcal{A}_{>p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{>p_\delta}}\right|}
$$

$$
\leq 2\left(\left|\beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{\leq p_\delta}}\right| + \left|\beta_{\mathcal{A}_{>p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{>p_\delta}}\right|\right)
$$

$$
\leq O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})\cdot\log(d)\cdot\|f^*\|_{L_2}^2, \tag{406}
$$

where the second inequality follows from the AM-GM inequality.

Similarly, we bound the last term in Eq. (399) using Cauchy-Schwarz and AM-GM inequality:

$$
\left|(\beta_{\mathcal{A}_{\leq p_\delta}} + \beta_{\mathcal{A}_{>p_\delta}})^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right|
$$

$$
\leq \left|\beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right| + \left|\beta_{\mathcal{A}_{>p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right|
$$

$$
\leq 2\left(\left|\beta_{\mathcal{A}_{\leq p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{\leq p_\delta}}\right| + \left|\beta_{\mathcal{A}_{>p_\delta}}^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_{\mathcal{A}_{>p_\delta}}\right|\right) + 4\left|\beta_\varepsilon^\top M_{\mathcal{A}_{\leq p_\delta}}\beta_\varepsilon\right|
$$

$$
= O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon} + d^{-2+2\gamma_{\max}})\cdot\log(d)\cdot(\|f^*\|_{L_2}^2 + \sigma_\varepsilon^2). \tag{407}
$$

Following a similar analysis with the proof of Proposition B.31, in parallel to Eq. (320), we can show that with constant $c > 0$ (152) the following holds:

$$\left| \left\| b_{\mathcal{A}_{\leq p_\delta}} V \right\|_2^2 - \left\| f^*_{\mathcal{A}_{< p_\delta}} \right\|_{L_2}^2 - c \cdot \left\| f^*_{\mathcal{A}_{= p_\delta}} \right\|_{L_2}^2 \right| = O_{d,\mathbb{P}}(d^{-\zeta_1}) \cdot \left\| b_{\mathcal{A}_{< p_\delta}} \right\|_2^2. \tag{408}$$

Plugging the above bound together with with Eq. (402), (404), (405), (406) and (407) into Eq. (399) and absorbing the logarithmic factor into ϵ completes the proof.

C.6.4 Proof of Proposition B.38

Note that with Lemma B.35 we can deduce

$$\mathbb{E}_x[(f^*_{\mathcal{A}_{< p_\delta}}(x) + \sqrt{c} \cdot f^*_{\mathcal{A}_{= p_\delta}}(x)) K(\sqrt{w} \odot x, Z)] = b_{\mathcal{A}_{\leq p_\delta}}^\top \widetilde{V} \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}(X)^\top, \tag{409}$$

where $\widetilde{V}_{\mathcal{A}_{< p_\delta}} = I$ and $\widetilde{V}_{\mathcal{A}_{= p_\delta}} = V_{\mathcal{A}_{= p_\delta}}$. With the decomposition

$$\beta = \underbrace{K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} b_{\mathcal{A}_{\leq p_\delta}}}_{:=\beta_{\mathcal{A}_{\leq p_\delta}}} + \underbrace{K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{> p_\delta}} b_{\mathcal{A}_{> p_\delta}}}_{:=\beta_{\mathcal{A}_{> p_\delta}}} + \underbrace{K_{w,\lambda}^{-1} \varepsilon}_{:=\beta_\varepsilon}, \tag{410}$$

we decompose the absolute difference:

$$\left| T_2 - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right|$$
$$\leq \left| b_{\mathcal{A}_{\leq p_\delta}}^\top (V^2 - \widetilde{V}^2) b_{\mathcal{A}_{\leq p_\delta}} \right| + \left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_{\mathcal{A}_{\leq p_\delta}} - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right|$$
$$+ \left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_{\mathcal{A}_{> p_\delta}} \right| + \left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_\varepsilon \right|, \tag{411}$$

where the diagonal matrix V is defined in Eq. (151).

It is straightforward to bound the first term using Eq. (319):

$$\left| b_{\mathcal{A}_{\leq p_\delta}}^\top (V^2 - \widetilde{V}^2) b_{\mathcal{A}_{\leq p_\delta}} \right| \leq \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \cdot \left\| V^2 - \widetilde{V}^2 \right\|_{\mathrm{op}} = O_d(d^{-\zeta_1}) \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2.$$

In the following, we will show that the remaining three terms on the right side of the above inequality are small. We first establish the following key estimate. We defer the proof to Appendix C.6.6.

Proposition C.64. *For any $\epsilon > 0$, the following estimate holds:*

$$\left\| \sqrt{n} \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} - \frac{1}{\sqrt{n}} V \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}). \tag{412}$$

Following Eq. (106), we denote $\zeta := \min(\zeta_1, \zeta_2)$ to simplify the notation. For the second term in Eq. (411), we can obtain

$$\left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} b_{\mathcal{A}_{\leq p_\delta}} - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right|$$
$$\overset{(a)}{=} \left| \frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} b_{\mathcal{A}_{\leq p_\delta}} - b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 b_{\mathcal{A}_{\leq p_\delta}} \right| + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \frac{1}{\sqrt{n}} \Phi_{\mathcal{A}_{\leq p_\delta}} \right\|_{\mathrm{op}} \cdot \left\| V b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2$$
$$\overset{(b)}{=} \left| b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \left(\frac{1}{n} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{\leq p_\delta}} - I \right) b_{\mathcal{A}_{\leq p_\delta}} \right| + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| V b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2$$
$$\overset{(c)}{=} O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon}) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \|V\|_{\mathrm{op}}^2 + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \|V\|_{\mathrm{op}}^2$$
$$= O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2, \tag{413}$$

125

where (a) follows from Proposition C.64, (b) follows from Lemma A.11 and (c) follows from Theorem A.12.

For the third term of (411), we similar invoke Proposition C.64 and we can obtain

$$\left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right|$$

$$= \left| \frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right| + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| \frac{1}{\sqrt{n}} \Phi_{\mathcal{A}_{\leq p_\delta}} \right\|_{\mathrm{op}} \left\| \frac{1}{\sqrt{n}} \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right\|_2 \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2 \|V\|_{\mathrm{op}}$$

$$= \left| \frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right| + O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \log^{1/2}(d) \left\| b_{\mathcal{A}_{>p_\delta}} \right\|_2 \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2, \tag{414}$$

where the last equality follows from Lemma A.11 and Proposition B.21.

Since $V_{j,j} \leq 1$ for all $j \in \mathcal{A}_{\leq p_\delta}$, a simple calculation shows

$$\mathbb{E}\left[\left| \frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right|^2 \right]$$

$$= \frac{1}{n^2} \mathbb{E}\left[\left(\sum_{i\in[n]} \sum_{j\in\mathcal{A}_{\leq p_\delta}} \sum_{k\in\mathcal{A}_{>p_\delta}} b_j V_{j,j}^2 \phi_j(x^{(i)}) \phi_k(x^{(i)}) b_k \right)^2 \right]$$

$$= \frac{1}{n^2} \mathbb{E}\left[\sum_{i\in[n]} \sum_{j\in\mathcal{A}_{\leq p_\delta}} \sum_{k\in\mathcal{A}_{>p_\delta}} b_j^2 V_{j,j}^4 b_k^2 \phi_j(x^{(i)})^2 \phi_k(x^{(i)})^2 \right]$$

$$\leq \frac{1}{n^2} \mathbb{E}\left[\sum_{i\in[n]} \sum_{j\in\mathcal{A}_{\leq p_\delta}} \sum_{k\in\mathcal{A}_{>p_\delta}} b_j^2 b_k^2 \phi_j(x^{(i)})^2 \phi_k(x^{(i)})^2 \right]$$

$$= \frac{\left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \left\| b_{\mathcal{A}_{>p_\delta}} \right\|_2^2}{n}.$$

Using Markov's inequality, we obtain

$$\left| \frac{1}{n} b_{\mathcal{A}_{\leq p_\delta}}^\top V^2 \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \Phi_{\mathcal{A}_{>p_\delta}} b_{\mathcal{A}_{>p_\delta}} \right| = O_{d,\mathbb{P}}(n^{-\frac{1}{2}} \log^{\frac{1}{2}}(d)) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2 \left\| b_{\mathcal{A}_{>p_\delta}} \right\|_2.$$

Combining this bound with Eq. (414) completes the bound for the second term in Eq. (411):

$$\left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_{\mathcal{A}_{>p_\delta}} \right| = O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \log^{1/2}(d) \cdot \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2 \left\| b_{\mathcal{A}_{>p_\delta}} \right\|_2$$

$$\leq O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \log^{1/2}(d) \cdot \|f^*\|_{L_2}^2. \tag{415}$$

For the last term in Eq. (411), taking the expectation of the square over ε yields

$$\mathbb{E}_{\varepsilon}\left[\left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_\varepsilon \right|^2 \right] = \sigma_\varepsilon^2 \cdot \left\| K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \Lambda_{\mathcal{A}_{\leq p_\delta}} V b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2. \tag{416}$$

Invoking Proposition C.64 gives

$$\left\| K_{w,\lambda}^{-1} \Phi_{\mathcal{A}_{\leq p_\delta}} \Lambda_{\mathcal{A}_{\leq p_\delta}} V b_{\mathcal{A}_{\leq p_\delta}} \right\|_2 = \frac{1}{n} \left\| \Phi_{\mathcal{A}_{\leq p_\delta}} V^2 b_{\mathcal{A}_{\leq p_\delta}} \right\|_2 + \frac{1}{\sqrt{n}} \cdot O_{d,\mathbb{P}}(d^{-\zeta/2+\epsilon}) \cdot \left\| V b_{\mathcal{A}_{\leq p_\delta}} \right\|_2$$

$$= O_{d,\mathbb{P}}(1) \cdot \frac{1}{\sqrt{n}} \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2,$$

where the second equality follows the bound $\left\| \Phi_{\mathcal{A}_{\leq p_\delta}} \right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\sqrt{n})$ (Lemma A.11) and $\|V\|_{\mathrm{op}} = O_d(1)$ (Eq. (312)).

Plugging the above equation into Eq. (416) and invoking Markov's inequality yields

$$\left| b_{\mathcal{A}_{\leq p_\delta}}^\top V \Lambda_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}^\top \beta_\varepsilon \right| = \frac{1}{\sqrt{n}} \cdot O_{d,\mathbb{P}}(1) \cdot \left(\sigma_\varepsilon^2 + \left\| b_{\mathcal{A}_{\leq p_\delta}} \right\|_2^2 \right). \tag{417}$$

Lastly, combining Eq. (408) with the bound for Eq. (411) using Eq. (413), (415) and (417), and absorbing the logarithmic factors into ϵ completes the proof.

C.6.5 Proof of Proposition C.63

By Lemma B.30, we have the estimate for K_w:

$$K_w = \Phi_{\mathcal{A}_{\leq p_\delta}}(Z) D_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}(Z) + \rho(I + \Delta),$$

with $\|\Delta\|_{\mathrm{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$.

Using the representation $\Phi_{\mathcal{A}_{\leq p_\delta}}(Z) = \Phi_{\mathcal{A}_{\leq p_\delta}}(X) W_{\mathcal{A}_{\leq p_\delta}}^{1/2}$ and adding the ridge parameter, we can write

$$K_{w,\lambda} = \Phi_{\mathcal{A}_{\leq p_\delta}}(X) \widehat{D}_{\mathcal{A}_{\leq p_\delta}} \Phi_{\mathcal{A}_{\leq p_\delta}}(X) + (\rho + \lambda)H,$$

where we denote $\widehat{D}_{\mathcal{A}_{\leq p_\delta}} := D_{\mathcal{A}_{\leq p_\delta}} W_{\mathcal{A}_{\leq p_\delta}}$ and $H := I + (\rho + \lambda)^{-1}\Delta$

For simplicity of notation, we omit the subscript and the argument X when it is clear from the context. Invoking Sherman–Morrison–Woodbury formula, we can obtain

$$K^{-1} = (\rho + \lambda)^{-1} H^{-1} - (\rho + \lambda)^{-1} H^{-1} \Phi((\rho + \lambda)\widehat{D}^{-1} + \Phi^\top H^{-1}\Phi)^{-1}\Phi^\top H^{-1}$$

Next we show $K^{-1}\Phi\Lambda$ and $\Lambda\Phi^\top K^{-1}$ can be further simplified. With the formula $I - (A + B)^{-1}B = (A + B)^{-1}A$ and $I - B(A + B)^{-1} = A(A + B)^{-1}$ we can get the following respectively:

$$K^{-1}\Phi\Lambda = H^{-1}\Phi((\rho + \lambda)\widehat{D}^{-1} + \Phi^\top H^{-1}\Phi)^{-1}(\widehat{D}^{-1}\Lambda), \tag{418}$$

$$\Lambda\Phi^\top K^{-1} = (\widehat{D}^{-1}\Lambda)((\rho + \lambda)\widehat{D}^{-1} + \Phi^\top H^{-1}\Phi)^{-1}\Phi^\top H^{-1}. \tag{419}$$

Combining these two equations, we obtain the following equalities corresponding to terms in Eq. (400) and (401):

$$nK^{-1}\Phi\Lambda^2\Phi^\top K^{-1} = \frac{1}{n}H^{-1}\Phi G^{-1}\left(\widehat{D}^{-1}\Lambda\right)^2 G^{-1}\Phi^\top H^{-1}, \tag{420}$$

$$\Phi^\top K^{-1}\Phi\Lambda^2\Phi^\top K^{-1}\Phi = \left(\frac{1}{n}\Phi^\top H^{-1}\Phi\right) G^{-1}\left(\widehat{D}^{-1}\Lambda\right)^2 G^{-1}\left(\frac{1}{n}\Phi^\top H^{-1}\Phi\right), \tag{421}$$

where $G := \left((\rho + \lambda)(n\widehat{D})^{-1} + \Phi^\top H^{-1}\Phi/n\right)$.

Note that any entry of $\widehat{D}^{-1}\Lambda$ takes the form

$$\left(\left(g^{|S|}(0)d^{-|S|} + O_d(d^{-|S|-1})\right) \cdot w^S\right)^{-1} \left(g^{(|S|)}(0)d^{-|S|}w^S + O_d(d^{-|S|-2+2\gamma_{\max}}w^S)\right)$$

$$= \left(d^{-|S|}w^S\right)^{-1}\left(g^{|S|}(0) + O_d(d^{-1})\right)^{-1} \cdot \left(g^{(|S|)}(0)d^{-|S|}w^S + O_d(d^{-|S|-2+2\gamma_{\max}}w^S)\right)$$

$$= \left(g^{|S|}(0) + O_d(d^{-1})\right)^{-1}\left(g^{|S|}(0) + O_d(d^{-2+2\gamma_{\max}})\right)$$

$$= 1 + O_d(d^{-2+2\gamma_{\max}}).$$

Then we immediately get

$$\left\|\widehat{D}^{-1}\Lambda - I\right\|_{\mathrm{op}} = O_d(d^{-2+2\gamma_{\max}}). \tag{422}$$

Together with Eq. (330) and (334) and i.e., $\left\|H^{-1} - I\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$ and $\left\|G^{-1} - V\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon})$, using triangle inequalities yields

$$\left\|\frac{1}{n}H^{-1}\Phi G^{-1}\left(\widehat{D}^{-1}\Lambda\right)^2 G^{-1}\Phi^\top H^{-1} - \frac{1}{n}\Phi V^2 \Phi^\top\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + d^{-2+2\gamma_{\max}}),$$

where we use the fact that $\|\Phi/\sqrt{n}\|_{\text{op}} = O_{d,\mathbb{P}}(1)$ holds (Lemma A.11).

Similarly, together with Eq. (331) and i.e., $\left\|\frac{1}{n}\Phi^\top H^{-1}\Phi - I\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2-\zeta_2/2+\epsilon})$ and (334) , using triangle inequalities yields

$$\left\|\frac{1}{n^2}\Phi^\top H^{-1}\Phi G^{-1}\left(\widehat{D}^{-1}\Lambda\right)^2 G^{-1}\Phi^\top H^{-1}\Phi - V^2\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon} + d^{-2+2\gamma_{\max}}),$$

which completes the proof.

C.6.6   Proof of Proposition C.64

Similar to the proof of Proposition C.63, we use the estimate for $K_{w,\lambda}$ and use Sherman-Morrison-Woodbury formula to compute the inverse. We can obtain the following equality similar to Eq. (419):

$$\widehat{D}_{\mathcal{A}_{\leq p_\delta}}\Phi_{\mathcal{A}_{\leq p_\delta}}^\top K_{w,\lambda}^{-1} = ((\rho + \lambda)\widehat{D}^{-1} + \Phi^\top H^{-1}\Phi)^{-1}\Phi^\top H^{-1} \tag{423}$$

where $H := I + (\rho + \lambda)^{-1}\Delta$ and Δ satisfies $\|\Delta\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$.

We similarly denote $G := \left((\rho + \lambda)(n\widehat{D})^{-1} + \Phi^\top H^{-1}\Phi/n\right)$. Eq. (330) and (334) give $\left\|H^{-1} - I\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_2/2+\epsilon})$ and $\left\|G^{-1} - V\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon})$. Therefore, using the triangle inequality we have

$$\left\|G^{-1}\Phi_{\mathcal{A}_{\leq p_\delta}}^\top H^{-1}/\sqrt{n} - V\Phi_{\mathcal{A}_{\leq p_\delta}}^\top/\sqrt{n}\right\|_{\text{op}} = O_{d,\mathbb{P}}(d^{-\zeta_1/2+\epsilon} + d^{-\zeta_2/2+\epsilon}).$$

Combining this bound with Eq. (423) completes the proof.

D   Technical lemmas

Lemma D.65. *Consider independent, mean zero, isotropic random vectors $x^{(1)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$ and suppose that the coordinates of each vector $x^{(i)}$ are independent and σ-subGaussian. Fix a constant $s \in (0, 1)$. Then with probability at least $1 - 2n^2 \exp(-cs^2 d)$ the estimate*

$$\left|\frac{1}{d}\langle x^{(i)}, x^{(j)}\rangle - \delta_{ij}\right| \leq s\sigma^2,$$

holds simultaneously for all $i, j = 1, \ldots n$.

Proof. For any fixed i and j and s, Bernstein's inequality [69, Corollary 2.8.3] gives

$$\mathbb{P}(|\tfrac{1}{d}\langle x^{(i)}, x^{(j)}\rangle - \delta_{ij}| \geq s\sigma^2) \leq 2\exp(-cs^2 d).$$

Taking a union bound over $i, j = 1, \ldots, n$ completes the proof.   □

Lemma D.66. *Consider random vectors $x^{(1)}, \ldots, x^{(n)}$ i.i.d. drawn from the uniform distribution over $\mathbb{H}_d = \{-1, +1\}^d$. Fix a constant $s \in (0, 1)$. Given a vector $w \in \mathbb{R}^d$ with $\|w\|_1 = d$ and $w_r \geq 0$ for $r = 1, 2, \ldots, d$, then with probability at least $1 - 2n^2 \exp(-s^2 d/(2\|w\|_\infty))$ the estimate*

$$\left|\frac{1}{d}\left\langle x^{(i)}, w \odot x^{(j)}\right\rangle - \delta_{ij}\right| \leq s, \tag{424}$$

holds simultaneously for all $i, j = 1, 2, \ldots, n$.

Proof. For any fixed i, j, if $i = j$, then we have

$$\frac{1}{d} \left\langle x^{(i)}, w \odot x^{(i)} \right\rangle = \|w\|_1 / d = 1.$$

For $i \neq j$ and any fixed s, Hoeffding's inequality [69, Theorem 2.2.6] gives

$$\mathbb{P}\left(\left| \frac{1}{d} \left\langle x^{(i)}, w \odot x^{(j)} \right\rangle \right| \geq s \right) \leq 2 \exp\left(-\frac{s^2 d}{2 \|w\|_\infty} \right),$$

where we use the fact that each coordinate of x satisfies $w_k x_k^{(i)} x_k^{(j)} \in [-w_k, w_k]$ hence $\sum_{k=1}^d w_k^2 \leq \|w\|_1 \|w\|_\infty = d \|w\|_\infty$.

Taking a union bound over $i, j = 1, \ldots, n$ completes the proof. $\qquad \square$

Lemma D.67. *Consider independent, mean zero, isotropic random vector $x \in \mathbb{R}^d$ and suppose that the coordinates of x are independent and σ-sub-Gaussian. Fix a constant $s \in (0, \sigma^{-2})$. Then with probability at least $1 - 2 \exp\left(-cds^2 \right)$ the following estimate holds:*

$$\left| \frac{1}{d^q} \|x\|_2^{2q} - 1 \right| \leq q 2^{2q} s \sigma^2 \qquad \forall q > 0.$$

Proof. Concentration of the norm [69, Theorem 3.1.1] yields for any $s > 0$ the estimate

$$\mathbb{P}(|\frac{1}{\sqrt{d}} \|x\|_2 - 1| \geq s\sigma^2) \leq 2 \exp\left(-cds^2 \right).$$

Next, we will use the elementary observation: for any $z > 0$, $t \in (0, 1)$, and $q \geq 1$ the implication holds:

$$|z^q - 1| \geq q \cdot 2^{q-1} t \qquad \Longrightarrow \qquad |z - 1| \geq t.$$

Therefore, setting $z = \frac{1}{\sqrt{d}} \|x\|_2$ and $t = s\sigma^2$ we deduce

$$\mathbb{P}(|z^q - 1| \geq q \cdot 2^{q-1} t) \leq \mathbb{P}(|z - 1| \geq t) \leq 2 \exp\left(-cds^2 \right).$$

Replacing q by $2q$ completes the proof. $\qquad \square$

Lemma D.68 (Matrix decoupling (Lemma 4 from [53]). *Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric random matrix. For $T_1, T_2 \subseteq [n]$, we denote $A_{T_1, T_2} = (A_{i,j})_{i \in T_1, j \in T_2}$. Then we have*

$$\mathbb{E}\left[\|A - \mathsf{Diag}(A)\|_{\mathrm{op}} \right] \leq 4 \max_{T \subseteq [n]} \mathbb{E}\left[\|A_{T, T^c}\|_{\mathrm{op}} \right]. \tag{425}$$

Lemma D.69 (Theorem 5.48 from [70]). *Let $A \in \mathbb{R}^{N \times n}$ with $A^\top = (\mathbf{a}_1, \cdots, \mathbf{a}_N)$ where $\mathbf{a}_i$ are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\Sigma = \mathbb{E}\left[\mathbf{a} \mathbf{a}_i^\top \right]$. Let $m := \mathbb{E}\left[\max_{i \in [N]} \|\mathbf{a}_i\|^2 \right]$. Then there exists an absolute constant $C > 0$ such that the following holds*

$$\mathbb{E}\left[\|A\|_{\mathrm{op}}^2 \right]^{1/2} \leq \|\Sigma\|_{\mathrm{op}}^{1/2} n^{1/2} + C m^{1/2} \log^{1/2}(\min(N, n)). \tag{426}$$

Lemma D.70 (Uniform bound using hypercontractivity). *Let $\phi_1, \ldots, \phi_n$ be finite-degree polynomials with maximum degree l that satisfy the hypercontractivity Assumption 4 with respect to measure μ_d. Then the following holds for any $q \geq 1$:*

$$\mathbb{E}\left[\max_{i \in [n]} |\phi_i|^2 \right] \leq n^{1/q} C_{l,2q}^{1/2} \max_{i \in [n]} \mathbb{E}\left[|\phi_i|^2 \right],$$

where $C_{l,2q} > 0$ is some constant.

Proof. Using Lyapunov's inequality for any $q \geq 1$, we deduce

$$\mathbb{E}\left[\max_{i \in [n]} |\phi_i|^2\right] \leq \mathbb{E}\left[\max_{i \in [n]} |\phi_i|^{2q}\right]^{1/q}$$

$$\overset{(a)}{\leq} \mathbb{E}\left[\sum_{i \in [n]} |\phi_i|^{2q}\right]^{1/q}$$

$$\leq n^{1/q} \max_{i \in [n]} \mathbb{E}\left[|\phi_i|^{2q}\right]^{1/q}, \tag{427}$$

where (a) follows from bounding the maximum by a sum. The hypercontractivity assumption implies there exists a constant $C_{l,2q} > 0$ such that the following holds

$$\mathbb{E}\left[|\phi_i|^{2q}\right]^{1/(2q)} \leq C_{l,2q} \mathbb{E}\left[|\phi_i|^2\right]^{1/2}, \tag{428}$$

where l is the maximum degree of $\phi_1, \ldots, \phi_n$. Plugging Eq. (428) into Eq. (427) yields

$$\mathbb{E}\left[\max_{i \in [n]} |\phi_i|^2\right] \leq n^{1/q} C_{l,2q}^{1/2} \max_{i \in [n]} \mathbb{E}\left[|\phi_i|^2\right],$$

which completes the proof. $\qquad\square$

Lemma D.71 (Expectation of Fourier feature matrices product (Theorem 2 from [75]))**.** *Suppose there exist non-overlapping sets $\mathcal{T}_1, \ldots, \mathcal{T}_\ell \subset [d]$ that satisfy $|\mathcal{T}_k| = \Theta(d^{s_k})$ for all $k \in [\ell]$ where $s_i \in [0,1]$. Fix the set families $\mathcal{S}_1^{[k]}, \ldots \mathcal{S}_{m+1}^{[k]} \in 2^{\mathcal{T}_k}$ for $k \in [\ell]$. Consider collections of sets $\mathcal{S}_1, \ldots, \mathcal{S}_{m+1} \in 2^{[d]}$ and weights $w^{(i)} \in \mathbb{R}_+^{\mathcal{S}_i}$, where $S_i = \sqcup_{k=1}^\ell S_i^{[k]}$ for all $S_i \in \mathcal{S}_i$. Define the matrix product*

$$M = A_1^\top \left(A_2 A_2^\top\right) \cdots \left(A_m A_m^\top\right) A_{m+1},$$

where $A_i = X_{\mathcal{S}_i} \mathsf{Diag}(w^{(i)})$ are the scaled Fourier-Walsh matrices. Define the adjusted effective degree for A_i:

$$p_i^{\mathrm{aeff}} := \sum_{k=1}^\ell s_k \cdot p_i^{[k]}.$$

Assume that the following regularity conditions hold for all indices $i, j \in [m+1]$:

1. **(degree bound)** *The inequalities $|S_i^{[k]}| \leq p_i^{[k]}$ hold for all $S_i^{[k]} \in \mathcal{S}_i^{[k]}$,*

2. **(trivial intersection)** *Whenever $\mathcal{S}_i$ intersects $\mathcal{S}_j$, the equality $\mathcal{S}_i = \mathcal{S}_j$ holds,*

3. **(small weights)** *The weights satisfy $w^{(i)} = O_d(n^{-1/2} \wedge d^{-p_i^{\mathrm{aeff}}/2})$ for all $i \in [m+1]$.*

Then the estimate

$$\|\mathbb{E}M\|_{\mathrm{op}} = O_d\left(d^{p_j^{\mathrm{aeff}}}\right) \|w^{(1)}\|_\infty \|w^{(m+1)}\|_\infty, \tag{429}$$

holds for any index $j \in [m+1]$ such that $\mathcal{S}_j$ is distinct from $\mathcal{S}_1$ and $\mathcal{S}_{m+1}$.