Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation

Lee H. Dicker Dean P. Foster Daniel Hsu

Department of Statistics and Biostatistics
Rutgers University
Piscataway, NJ 08854
e-mail: ldicker@stat.rutgers.edu

Department of Statistics
Wharton School, University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: dean@foster.net

Department of Computer Science
Columbia University
New York, NY 10027, USA
e-mail: djhsu@cs.columbia.edu

Abstract: Regularization is an essential element of virtually all kernel methods for nonparametric regression problems. A critical factor in the effectiveness of a given kernel method is the type of regularization that is employed. This article compares and contrasts members from a general class of regularization techniques, which notably includes ridge regression and principal component regression. We first derive risk bounds for these techniques that match the minimax rates in several settings, using recent large deviations machinery and a natural bias-variance decomposition. We then show that certain regularization techniques are more adaptable than others to favorable regularity properties that the true regression function may possess. This, in particular, demonstrates a striking difference between kernel ridge regression and kernel principal component regression.

Keywords and phrases: Learning theory, Principal component regression, Reproducing kernel Hilbert space, Ridge regression.

1. Introduction

Suppose that the observed data consists of $\mathbf{z}_i = (y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ and $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$. Suppose further that $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \rho$ are iid from some probability distribution ρ on $\mathcal{Y} \times \mathcal{X}$. Let $\rho(\cdot|\mathbf{x})$ denote the conditional distribution of y_i given $\mathbf{x}_i = \mathbf{x} \in \mathcal{X}$ and let $\rho_{\mathcal{X}}$

denote the marginal distribution of $\mathbf{x}_i$. Our goal is to use the available data to estimate the regression function of y on $\mathbf{x}$,

$$f^\dagger(\mathbf{x}) = \int_{\mathcal{Y}} y \; d\rho(y|\mathbf{x}),$$

which minimizes the mean-squared prediction error

$$\int_{\mathcal{Y}\times\mathcal{X}} \{y - f(\mathbf{x})\}^2 \; d\rho(y, \mathbf{x})$$

over $\rho_{\mathcal{X}}$-measurable functions $f : \mathcal{X} \to \mathbb{R}$. More specifically, for an estimator $\hat{f}$ define the risk

$$\mathcal{R}_\rho(\hat{f}) = \mathbb{E}\left[\int \{f^\dagger(\mathbf{x}) - \hat{f}(\mathbf{x})\}^2 \; d\rho_{\mathcal{X}}(\mathbf{x})\right] = \mathbb{E}\left(||f^\dagger - \hat{f}||^2_{\rho_{\mathcal{X}}}\right), \tag{1}$$

where the expectation is computed over $\mathbf{z}_1, \ldots, \mathbf{z}_n$ and $||\cdot||_{\rho_{\mathcal{X}}}$ denotes the norm on $L^2(\rho_{\mathcal{X}})$; we seek estimators $\hat{f}$ which minimize $\mathcal{R}_\rho(\hat{f})$.

This is a version of the random design nonparametric regression problem. There is a vast literature on nonparametric regression, along with a huge variety of corresponding methods (e.g., Gyorfi et al., 2002; Wasserman, 2006). In this paper, we focus on regularization and kernel methods for estimating $f^\dagger$ (here, we mean "kernel" as in *reproducing kernel Hilbert space*, rather than *kernel-smoothing*, which is another popular approach to nonparametric regression). Most of our results apply to general regularization operators. However, our motivating examples are two well-known regularization techniques: Kernel ridge regression (which we refer to as "KRR"; KRR is also known as Tikhonov regularization) and kernel principal component regression ("KPCR"; also known as spectral cut-off regularization).

Our main contribution is two-fold. First, we derive new risk bounds for a general class of kernel-regularization methods, which includes both KRR and KPCR. These bounds imply that the corresponding regularization methods achieve the minimax rate for estimating $f^\dagger$ with respect to a variety of kernels and settings; this is illustrated by example in Corollaries 1–4 of Section 6.1. In each example, the minimax rate is determined by the eigenvalues of the kernel operator, computed with respect to $\rho_{\mathcal{X}}$. Second, we show precisely how some types of regularization operators are able to adapt to the regularity of $f^\dagger$, while others are known to have more limitations in this regard and suffer from the *saturation* effect (Bauer et al., 2007; Mathé, 2005; Neubauer, 1997). More specifically, we show that if $f^\dagger$ has an especially simplified expansion in the eigenbasis of the kernel operator, then some regularization estimators are able to leverage this to obtain even faster (minimax) convergence rates. As a consequence of our adaptation results, we conclude that KPCR is fully minimax adaptive in a range of settings, while KRR is known to saturate. This illustrates a striking advantage that KPCR may have over KRR in these settings.

Related Work

Kernel ridge regression and other regularization methods have been widely studied. Indeed, it is well-known that KRR is minimax in many of the settings we consider in this paper, such as those described in Corollaries 1–4 below (Caponnetto and De Vito, 2007; Zhang, 2005; Zhang et al., 2013). However, our risk bounds and corresponding minimax results apply to more general regularization operators (not just KRR/Tikhonov regularization), which appears to be new. Existing risk bounds for other regularization operators tend to be looser (Bauer et al., 2007) (see the comments after Theorem 1 below) or require auxiliary information (e.g., the bounds in (Caponnetto and Yao, 2010) apply to semi-supervised settings where an additional pool of unlabeled data is available).

Others have also observed that KPCR may have advantages over KRR, as discussed above. Indeed, others have even observed that Tikhonov regularization (KRR) saturates, while spectral cut-off regularization (KPCR) does not (Bauer et al., 2007; Lo Gerfo et al., 2008; Mathé, 2005). Our results (Propositions 4–5) extend beyond this observation to precisely quantify the advantages of unsaturated regularization operators in terms of adaptation and minimaxity. In other related work, Dhillon et al. (2013) have illustrated the potential advantages of KPCR over KRR in finite-dimensional problems with linear kernels; though their work is not framed in terms of saturation and general regularization operators, it essentially relies on similar concepts.

The main engine behind the technical results in this paper is a collection of large-deviation results for Hilbert-Schmidt operators. The required machinery is developed in Appendix A (especially Lemmas 3–5 and Corollary 5); these results build on straightforward extensions of results from (Tropp, 2015). Additionally, our Proposition 5, on finite-rank adaptation, relies on well-known eigenvalue perturbation results that have been adapted to handle Hilbert-Schmidt operators (e.g. the Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970)).

2. Statistical Setting and Assumptions

Our basic assumption on the distribution of $\mathbf{z} = (y, \mathbf{x}) \sim \rho$ is that the residual variance is bounded; more specifically, we assume that there exists a constant $\sigma^2 > 0$ such that

$$\int_{\mathcal{Y}} \{y - f^\dagger(\mathbf{x})\}^2 \, d\rho(y|\mathbf{x}) \leq \sigma^2 \tag{2}$$

for almost all $\mathbf{x} \in \mathcal{X}$. Zhang et al. (2013) also assume (2); this assumption is slightly weaker than the analogous assumption in (Bauer et al., 2007) (equation (1) in their paper). Note that (2) holds if y is bounded almost surely.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive-definite kernel function. We assume that K is bounded, i.e., that there exists $\kappa^2 > 0$ such that

$$\sup_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x}, \mathbf{x}) \leq \kappa^2.$$

Additionally, we assume that there is a countable basis of eigenfunctions $\{\psi_j\}_{j=1}^{\infty} \subseteq L^2(\rho_{\mathcal{X}})$ and a sequence of corresponding eigenvalues $t_1^2 \geq t_2^2 \geq \cdots \geq 0$ such that

$$K(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{j=1}^{\infty} t_j^2 \psi_j(\mathbf{x}) \psi_j(\tilde{\mathbf{x}}), \quad \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X} \tag{3}$$

and the convergence is absolute. Mercer's theorem and various generalizations give conditions under which representations like (3) are known to hold (Carmeli et al., 2006); one of the simplest examples is when $\mathcal{X}$ is a compact Hausdorff space, $\rho_{\mathcal{X}}$ is a probability measure on the Borel sets of $\mathcal{X}$, and K is continuous. Observe that

$$\sum_{j=1}^{\infty} t_j^2 = \sum_{j=1}^{\infty} t_j^2 \int_{\mathcal{X}} \psi_j(\mathbf{x})^2 \, d\rho_{\mathcal{X}}(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}) \, d\rho_{\mathcal{X}}(\mathbf{x}) \leq \kappa^2;$$

in particular, $\{t_j^2\} \in \ell^1(\mathbb{N})$.

Let $\mathcal{H} \subseteq L^2(\rho_{\mathcal{X}})$ be the reproducing kernel Hilbert space (RKHS) corresponding to K (Aronszajn, 1950) and let

$$\phi_j = t_j \psi_j, \quad j = 1, 2, \ldots \tag{4}$$

It follows from basic facts about RKHSs that $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$ (if $t_J^2 > t_{J+1}^2 = 0$, then $\{\phi_j\}_{j=1}^{J}$ is an orthonormal basis for $\mathcal{H}$). Our main assumption on the relationship between y, $\mathbf{x}$, and the kernel K is that

$$f^{\dagger} \in \mathcal{H}. \tag{5}$$

This is a regularity or smoothness assumption on $f^{\dagger}$. We represent $f^{\dagger}$ in the bases $\{\psi_j\}_{j=1}^{\infty}$ and $\{\phi_j\}_{j=1}^{\infty}$ as

$$f^{\dagger} = \sum_{j=1}^{\infty} \gamma_j \psi_j = \sum_{j=1}^{\infty} \beta_j \phi_j, \tag{6}$$

where $\beta_j = \gamma_j / t_j$. Then the assumption (5) is equivalent to

$$\sum_{j=1}^{\infty} \beta_j^2 = \sum_{j=1}^{\infty} \frac{\gamma_j^2}{t_j^2} < \infty.$$

Many of the results in this paper can be modified, so that they apply to settings where $f^\dagger \notin \mathcal{H}$, by replacing $f^\dagger$ with an appropriate projection of $f^\dagger$ onto $\mathcal{H}$ and including an approximation error term in the corresponding bounds. This approach leads to the study of *oracle inequalities* (Hsu et al., 2014; Koltchinskii, 2006; Steinwart et al., 2009; Zhang, 2005; Zhang et al., 2013), which we do not pursue in detail here. However, investigating oracle inequalities for general regularization operators may be of interest for future research, as most existing work focuses on ridge regularization.

3. Regularization

As discussed in Section 1, our goal is find estimators $\hat{f}$ that minimize the risk (1). In this paper, we focus on regularization-based estimators for $f^\dagger$. In order to precisely describe these estimators, we require some additional notation for various operators that will be of interest, and some basic definitions from regularization theory.

3.1. Finite-Rank Operators of Interest

For $\mathbf{x} \in \mathcal{X}$, define $K_\mathbf{x} \in \mathcal{H}$ by $K_\mathbf{x}(\tilde{\mathbf{x}}) = K(\mathbf{x}, \tilde{\mathbf{x}})$, $\tilde{\mathbf{x}} \in \mathcal{X}$, and let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Additionally, define the finite-rank linear operators $S_X : \mathcal{H} \to \mathbb{R}^n$ and $T_X : \mathcal{H} \to \mathcal{H}$ (both depending on X) by

$$S_X \phi = (\langle \phi, K_{\mathbf{x}_1} \rangle_\mathcal{H}, \ldots, \langle \phi, K_{\mathbf{x}_n} \rangle_\mathcal{H})^\top = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))^\top,$$
$$T_X \phi = \frac{1}{n} \sum_{i=1}^n \langle \phi, K_{\mathbf{x}_i} \rangle_\mathcal{H} K_{\mathbf{x}_i} = \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{x}_i) K_{\mathbf{x}_i},$$

where $\langle \cdot, \cdot \rangle_\mathcal{H}$ is the inner-product on $\mathcal{H}$ and $\phi \in \mathcal{H}$. Let $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ denote the normalized inner-product on $\mathbb{R}^n$, defined by $\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle_{\mathbb{R}^n} = n^{-1} \mathbf{v}^\top \tilde{\mathbf{v}}$ for $\mathbf{v} = (v_1, \ldots, v_n)^\top$, $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_n)^\top \in \mathbb{R}^n$. Then the adjoint of S_X with respect to $\langle \cdot, \cdot \rangle_\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$, $S_X^* : \mathbb{R}^n \to \mathcal{H}$, is given by

$$S_X^* \mathbf{v} = \frac{1}{n} \sum_{i=1}^n v_i K_{\mathbf{x}_i}.$$

Additionally, $T_X = S_X^* S_X$. Finally, observe that $S_X S_X^* : \mathbb{R}^n \to \mathbb{R}^n$ is given by the $n \times n$ matrix $S_X S_X^* = n^{-1} \mathbf{K}$, where $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$; $\mathbf{K}$ is the *kernel matrix*, which is ubiquitous in kernel methods and enables finite computation.

3.2. Basic Definitions

A family of functions $g_\lambda : [0, \infty) \to [0, \infty)$ indexed by $\lambda > 0$ is called a *regularization family* if it satisfies the following three conditions:

(R1) $\sup_{0<t\leq\kappa^2} |tg_\lambda(t)| < 1$.
(R2) $\sup_{0<t\leq\kappa^2} |1 - tg_\lambda(t)| \leq 1$.
(R3) $\sup_{0<t\leq\kappa^2} |g_\lambda(t)| < \lambda^{-1}$.

(This definition follows (Bauer et al., 2007; Engl et al., 1996), but is slightly more restrictive.) The main idea behind a regularization family is that it "looks" similar to $g(t) = 1/t$, but is better-behaved near $t = 0$, i.e., it is bounded by λ^{-1}. An important quantity that is related to the adaptability and saturation point of a regularization family (mentioned in Section 1), is the *qualification* of the regularization. The qualification of the regularization family $\{g_\lambda\}_{\lambda>0}$ is defined to be the maximal $\xi \geq 0$ such that

$$\sup_{0<t\leq\kappa^2} |1 - tg_\lambda(t)|t^\xi \leq \lambda^\xi. \tag{7}$$

The two primary examples of regularization considered in this paper are ridge (Tikhonov) regularization, where

$$g_\lambda(t) = r_\lambda(t) = \frac{1}{t + \lambda}$$

and principal component (spectral cut-off) regularization, where

$$g_\lambda(t) = s_\lambda(t) = \frac{1}{t} I\{t \geq \lambda\}. \tag{8}$$

Observe that ridge regularization has qualification 1 and principal component regularization has qualification ∞. Another example of a regularization family is the Landweber iteration, which can be viewed as a special case of gradient descent (see, e.g., Bauer et al., 2007; Lo Gerfo et al., 2008; Rosasco et al., 2005). Throughout the paper, all regularization families are assumed to have qualification at least 1.

3.3. Estimators

Given a regularization family $\{g_\lambda\}_{\lambda>0}$, we define the g_λ-regularized estimators for $f^\dagger$,

$$\hat{f}_\lambda = g_\lambda(T_X)S_X^*\mathbf{y}. \tag{9}$$

Here, g_λ acts on the spectrum (eigenvalues) of the finite-rank operator T_X. The dependence of $\hat{f}_\lambda$ on the regularization family is implicit; our results hold for any regularization family

except where explicitly stated (in particular, Section 6.2). The estimators $\hat{f}_\lambda$ are the main focus of this paper.

To provide some intuition behind $\hat{f}_\lambda$, define the positive self-adjoint operator $T : \mathcal{H} \to \mathcal{H}$ by

$$T\phi = \int_{\mathcal{X}} \langle \phi, K_{\mathbf{x}} \rangle_{\mathcal{H}} K_{\mathbf{x}} \, d\rho_{\mathcal{X}}(\mathbf{x}) = \sum_{j=1}^{\infty} t_j^2 \langle \phi, \phi_j \rangle_{\mathcal{H}} \phi_j, \quad \phi \in \mathcal{H}.$$

Observe that T is a "population" version of the operator T_X. Unlike T_X, T often has infinite rank; however, we still might expect that

$$T \approx T_X \tag{10}$$

for large n (where the approximation holds in some suitable sense).

We also have a large-n approximation for $S_X^* \mathbf{y}$. For $\phi \in \mathcal{H}$,

$$\begin{aligned} \langle \phi, S_X^* \mathbf{y} \rangle_{\mathcal{H}} &= \frac{1}{n} \sum_{i=1}^{n} y_i \langle \phi, K_{\mathbf{x}_i} \rangle_{\mathcal{H}} \approx \int_{\mathcal{Y} \times \mathcal{X}} y \phi(\mathbf{x}) \, d\rho(y, \mathbf{x}) \\ &= \int_{\mathcal{X}} f^\dagger(\mathbf{x}) \phi(\mathbf{x}) \, d\rho_{\mathcal{X}}(\mathbf{x}) = \langle \phi, f^\dagger \rangle_{L^2(\rho_{\mathcal{X}})} = \langle \phi, T f^\dagger \rangle_{\mathcal{H}}, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{L^2(\rho_{\mathcal{X}})}$ denotes the inner-product on $L^2(\rho_{\mathcal{X}})$ and we have used (4) to obtain the last equality. It follows that $S_X^* \mathbf{y} \approx T f^\dagger$. Hence, to recover $f^\dagger$ from $\mathbf{y}$, it would be natural to apply the inverse of T to $S_X^* \mathbf{y}$. However, T is not invertible whenever it has infinite rank, and regularization becomes necessary. We thus arrive at the chain of approximations which help motivate $\hat{f}_\lambda$:

$$\hat{f}_\lambda = g_\lambda(T_X) S_X^* \mathbf{y} \approx g_\lambda(T) T f^\dagger \approx f^\dagger,$$

where $g_\lambda(T)$ may be viewed as an approximate inverse for a suitably chosen regularization parameter λ. The rest of the paper is devoted to deriving bounds on the risk

$$\mathcal{R}_\rho(\hat{f}_\lambda) = \mathbb{E}\left(||f^\dagger - \hat{f}_\lambda||_{\rho_{\mathcal{X}}}^2 \right) = \mathbb{E}\left\{ ||f^\dagger - g_\lambda(T_X) S_X^* \mathbf{y}||_{\rho_{\mathcal{X}}}^2 \right\} \tag{11}$$

and investigating the effect of specific regularization families $\{g_\lambda\}_{\lambda > 0}$ on the accuracy of $\hat{f}_\lambda$.

4. Preliminary Simplifications

The main results in this paper involve bounding (11). This section is devoted to deriving some preliminary simplifications of (11).

By (4), the $L^2(\rho_{\mathcal{X}})$-norm in (11) can be converted into a norm on $\mathcal{H}$; indeed, we have $\mathcal{R}_\rho(\hat{f}_\lambda) = \mathbb{E}\{||T^{1/2}(f^\dagger - \hat{f}_\lambda)||_{\mathcal{H}}^2\}$. Next, define the residuals $\epsilon_i = y_i - f^\dagger(\mathbf{x}_i)$, $i = 1, \dots, n$, and let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$. Then $\hat{f}_\lambda = g_\lambda(T_X)T_X f^\dagger + g_\lambda(T_X)S_X^*\boldsymbol{\epsilon}$ and it follows that

$$\begin{aligned}\mathcal{R}_\rho(\hat{f}_\lambda) &= \mathbb{E}\left[||T^{1/2}\{I - g_\lambda(T_X)T_X\}f^\dagger - T^{1/2}g_\lambda(T_X)S_X^*\boldsymbol{\epsilon}||_{\mathcal{H}}^2\right] \\ &= \mathbb{E}\left[||T^{1/2}\{I - g_\lambda(T_X)T_X\}f^\dagger||_{\mathcal{H}}^2\right] + \mathbb{E}\left\{||T^{1/2}g_\lambda(T_X)S_X^*\boldsymbol{\epsilon}||_{\mathcal{H}}^2\right\}. \end{aligned} \tag{12}$$

This is a bias/variance decomposition of the risk $\mathcal{R}_\rho(\hat{f}_\lambda)$; the first term in (12) represents the bias of $\hat{f}_\lambda$ and the second term represents the variance.

In order to further simplify (12), we first note that the Hilbert space $\mathcal{H}$ is isometric to $\ell^2(\mathbb{N})$ via the isometry $\iota : \mathcal{H} \to \ell^2(\mathbb{N})$, given by

$$\iota : \sum_{j=1}^{\infty} \alpha_j \phi_j \mapsto (\alpha_1, \alpha_2, \dots)^\top \tag{13}$$

(if $\mathcal{H}$ is finite dimensional and $t_J^2 > t_{J+1}^2 = 0$, then take $0 = \alpha_{J+1} = \alpha_{J+2} = \cdots$; we take all elements of $\ell^2(\mathbb{N})$ to be infinite-dimensional column vectors). Using this equivalence, we can convert elements of $\mathcal{H}$ and linear operators on $\mathcal{H}$ appearing in (12) into (infinite-dimensional) vectors and matrices, respectively, which we find simpler to analyze in the sequel.

Before proceeding, we introduce some more notation. Define the infinite-dimensional diagonal matrix $\mathcal{T} = \mathrm{diag}(t_1^2, t_2^2, \dots)$ and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)^\top \in \ell^2(\mathbb{N})$. Next define the $n \times \infty$ matrices $\boldsymbol{\Psi} = (\psi_j(\mathbf{x}_i))_{1 \le i \le n; 1 \le j < \infty}$ and $\boldsymbol{\Phi} = \boldsymbol{\Psi}\mathcal{T} = (\phi_j(\mathbf{x}_i))_{1 \le i \le n; 1 \le j < \infty}$. Observe that

$$S_X = \boldsymbol{\Phi} \circ \iota, \tag{14}$$

$$S_X^* = \iota^{-1} \circ \left(\frac{1}{n}\boldsymbol{\Phi}^\top\right), \tag{15}$$

$$T = \iota^{-1} \circ \mathcal{T} \circ \iota. \tag{16}$$

Finally, let $||\cdot|| = ||\cdot||_{\ell^2(\mathbb{N})}$ denote the norm on $\ell^2(\mathbb{N})$.

Combining (12)–(16) yields the next proposition, which is the starting point for obtaining risk bounds in the sections that follow.

Proposition 1. *Suppose that $\hat{f}_\lambda$ is the estimator defined in (9). Then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = \mathrm{B}_\rho(g_\lambda) + \mathrm{V}_\rho(g_\lambda),$$

where

$$\mathrm{B}_\rho(g_\lambda) = \mathrm{B}_\rho^{(n)}(g_\lambda) = \mathbb{E}\left[\left|\left|\mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2\right],$$

$$\mathrm{V}_\rho(g_\lambda) = \mathrm{V}_\rho^{(n)}(g_\lambda) = \frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\right\}.$$

Proposition 1 is a further simplified version of the bias/variance decomposition (12); $\mathrm{B}_\rho(g_\lambda)$ and $\mathrm{V}_\rho(g_\lambda)$ are the bias and variance terms, respectively.

5. Main Result

In this section, we give separate bounds on the bias and variance terms from Proposition 1, which are then combined to yield our main result bounding $\mathcal{R}_\rho(\hat{f}_\lambda)$. Three major sources contribute to our upper bound on the bias in Proposition 2 below (and similarly to our bound on the variance in Proposition 3): Two types of approximation error, and what we refer to as the "intrinsic bias" (or "intrinsic variance").

The two types of approximation error correspond to finite- and infinite-dimensional components, which arise from the approximation (10). The intrinsic bias and variance are determined by the regularization family $\{g_\lambda\}_{\lambda>0}$ and the specific regularization parameter λ used to compute $\hat{f}_\lambda$. A schematic for our risk bound decomposition may be found in Figure 1. Note, however, that while the bias/variance decomposition of $\mathcal{R}_\rho(\hat{f}_\lambda)$ is an unambiguous additive identity, attribution of the various sources of error in Propositions 2–3 is somewhat more subtle. Still, we view this as a useful heuristic.

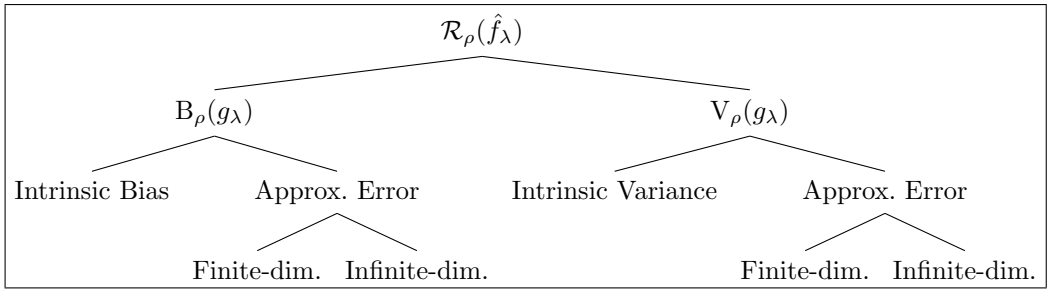

FIG 1. *Schematic for decomposition of risk bound.*

In vector-matrix notation (i.e., working in $\ell^2(\mathbb{N})$ under the isometry (13)) the approximation (10) can be rewritten as

$$\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi} \approx \mathcal{T}. \tag{17}$$

The approximation error in (17) is essentially the source of the approximation error terms in the schematic, Figure 1. Controlling this error is a key piece of our strategy for proving Propositions 2–3; it relies on large deviations results for random matrices and Hilbert-Schmidt operators, which have been developed in (Minsker, 2011; Tropp, 2015). The necessary technical lemmas for our bounds are derived in Appendix A.

Before proceeding, we aim to provide more intuition behind the finite-dimensional and infinite-dimensional approximation error cited in Figure 1. Note that the most direct approach to bounding the overall approximation error in (17) is to bound

$$\delta = \left|\left|\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}\right|\right|, \tag{18}$$

where $||\cdot||$ denotes the operator norm (this is straightforward using the results of Minsker (2011) and Tropp (2015)). However, in order to obtain improved bounds on $\mathrm{B}_\rho(g_\lambda)$ and $\mathrm{V}_\rho(g_\lambda)$, we decompose $n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}$ into a finite-dimensional block and an infinite-dimensional "tail," and consider these terms separately. More specifically, let $J \in \mathbb{N}$ be a fixed positive integer and define the block decompositions

$$\boldsymbol{\Phi} = (\boldsymbol{\Phi}_0 \;\; \boldsymbol{\Phi}_1), \; \mathcal{T} = \begin{pmatrix} \mathcal{T}_0 & 0 \\ 0 & \mathcal{T}_1 \end{pmatrix},$$

where $\boldsymbol{\Phi}_0 = (\phi_j(\mathbf{x}_i))_{1\le i\le n; 1\le j\le J}$, $\boldsymbol{\Phi}_1 = (\phi_j(\mathbf{x}_i))_{1\le i\le n; J+1\le j<\infty}$, $\mathcal{T}_0 = \mathrm{diag}(t_1^2, \dots, t_J^2)$, and $\mathcal{T}_1 = \mathrm{diag}(t_{J+1}^2, t_{J+2}^2, \dots)$. Additionally, define

$$\delta_0 = \left|\left|\frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 - \mathcal{T}_0\right|\right|, \quad \delta_1 = \left|\left|\begin{pmatrix} 0 & \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_1 \\ \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_0 & \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 - \mathcal{T}_1 \end{pmatrix}\right|\right|. \tag{19}$$

Observe that

$$\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T} = \begin{pmatrix} \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 - \mathcal{T}_0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_1 \\ \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_0 & \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 - \mathcal{T}_1 \end{pmatrix} \tag{20}$$

and $\delta \le \delta_0 + \delta_1$. The matrix $n^{-1}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 - \mathcal{T}_0$ has dimension $J \times J$ and contributes to the finite-dimensional approximation error in (17); δ_1 reflects the infinite-dimensional approximation error.

Explicit bounds on the bias and variance terms $\mathrm{B}_\rho(g_\lambda)$ and $\mathrm{V}_\rho(g_\lambda)$ are given in Propositions 2–3 below. While the connection between these bounds and δ_0, δ_1 may not be immediately transparent from the statement of the results, it is developed more fully in the corresponding proofs (found in Appendix B).

Proposition 2. *Let* $\mathrm{B}_\rho(g_\lambda)$ *be the bias term given in Proposition 1 and suppose that* $\{\tau_j^2\}_{j=1}^\infty$ *is a sequence of non-negative real numbers satisfying*

$$\sup_{\tilde{J}\in\mathbb{N},\ \mathbf{x}\in\mathcal{X}} \tau_{\tilde{J}}^2 \sum_{j=1}^{\tilde{J}} \psi_j(\mathbf{x})^2 \leq 1. \tag{21}$$

(Observe that $\tau_j^2 = t_j^2/\kappa^2$ *always satisfies this condition.) Then*

$$\mathrm{B}_\rho(g_\lambda) \leq \left\{ \frac{12}{t_J^2}\lambda^2 + \frac{408\kappa^2}{nt_J^2}\mathrm{tr}(\mathcal{T}_1) + \frac{180\kappa^4}{nt_J^2} + 13t_{J+1}^2 + \kappa^2 J\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}} \right\} ||f^\dagger||_{\mathcal{H}}^2.$$

Proposition 3. *Let* $\mathrm{V}_\rho(g_\lambda)$ *be the variance term given in Proposition 1 and suppose that* $\{\tau_j^2\}_{j=1}^\infty$ *is a sequence of non-negative real numbers satisfying (21). Then*

$$\mathrm{V}_\rho(g_\lambda) \leq \left[2J + \frac{2Jt_{J+1}^2}{\lambda} + \frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1) + \frac{32J}{\lambda}\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{1/2} + \frac{12\kappa^2 J}{\lambda n} + \frac{\kappa^2 J}{\lambda}\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}} \right] \frac{\sigma^2}{n}.$$

As mentioned above, it is a fairly subtle task to attribute each term in the upper bounds of Propositions 2-3 to a certain type of approximation error or the intrinsic bias or variance. However, a simple rule of thumb is that each appearance of λ in the upper bounds is related to the intrinsic bias or variance; t_J^2 and J are related to the finite-dimensional approximation error; and t_{J+1}^2 and $\mathrm{tr}(\mathcal{T}_1) = \sum_{j>J} t_j^2$ are related to the infinite-dimensional approximation error. Propositions 1–3 are easily combined to obtain our main theorem.

Theorem 1. *Suppose that* $\hat{f}_\lambda$ *is the estimator defined in (9) and suppose that* $\{\tau_j^2\}_{j=1}^\infty$ *is a sequence of non-negative real numbers satisfying (21). Then for all positive integers* $J \in \mathbb{N}$,

$$\begin{aligned}\mathcal{R}_\rho(\hat{f}_\lambda) \leq & \left[2J + \frac{2Jt_{J+1}^2}{\lambda} + \frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1) + \frac{32J}{\lambda}\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{1/2} + \frac{12\kappa^2 J}{\lambda n} + \frac{\kappa^2 J}{\lambda}\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}} \right] \frac{\sigma^2}{n} \\ & + \left\{ \frac{12}{t_J^2}\lambda^2 + \frac{408\kappa^2}{nt_J^2}\mathrm{tr}(\mathcal{T}_1) + \frac{180\kappa^4}{nt_J^2} + 13t_{J+1}^2 + \kappa^2 J\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}} \right\} ||f^\dagger||_{\mathcal{H}}^2.\end{aligned}$$

The constants in the bounds from Propositions 2–3 (and hence Theorem 1) have not been optimized and can likely be improved. However, we will see in the next section that Theorem 1 implies the risk of $\hat{f}_\lambda$ achieves the minimax rate in a variety of important examples. Bauer

et al. (2007) have previously obtained risk bounds on general regularization estimators similar to $\hat{f}_\lambda$. However, their bounds (e.g. Theorem 10 in (Bauer et al., 2007)) are independent of the ambient RKHS $\mathcal{H}$, i.e. they do not depend on the eigenvalues $\{t_j^2\}$. Our bounds are tighter than those in (Bauer et al., 2007) because we take advantage of the structure of $\mathcal{H}$. In contrast with our Theorem 1, Bauer et al.'s (2007) results do not give minimax bounds (not easily, at least), because minimax rates must depend on the t_j^2 (see Corollaries 1–4 below).

6. Consequences

In this section, we use Theorem 1 to derive minimax optimal rates for all regularization families, and to demonstrate the adaptation properties for regularization families with sufficient qualification.

6.1. Minimax Optimality

For $0 < r < \infty$, define

$$B_{\mathcal{H}}(r) = \left\{ f = \sum_{j=1}^{\infty} \alpha_j \psi_j \in L^2(\rho_{\mathcal{X}});\ \sum_{j=1}^{\infty} \frac{\alpha_j^2}{t_j^2} \leq r^2 \right\} = \{ f \in \mathcal{H};\ ||f||_{\mathcal{H}}^2 \leq r^2 \} \subseteq \mathcal{H}.$$

The set $B_{\mathcal{H}}(r)$ can be interpreted as an ellipsoid in $L^2(\rho_{\mathcal{X}})$ or a ball in $\mathcal{H}$. There is an extensive literature in statistics on nonparametric regression problems and minimax rates for estimating $f^\dagger$ over ellipsoids of the form $B_{\mathcal{H}}(r)$ (e.g., Nemirovskii, 1985; Nemirovskii et al., 1983; Tsybakov, 2004). These rates are typically determined by some condition on the rate of decay of $t_1^2 \geq t_2^2 \geq \cdots \geq 0$. In the classical statistics literature, the basis $\{\psi_j\}_{j=1}^{\infty}$ and sequence $\{t_j\}_{j=1}^{\infty}$ are given directly, while in the present setting both are derived from the interaction between the kernel K and the distribution $\rho_{\mathcal{X}}$. In this section, we show that for several commonly studied decay conditions on $\{t_j^2\}_{j=1}^{\infty}$ and appropriate choice of the regularization parameter λ, the estimator $\hat{f}_\lambda$ achieves the minimax rate for estimating $f^\dagger \in B_{\mathcal{H}}(r)$, as $n \to \infty$. Our results are summarized in the four corollaries to Theorem 1 given below; proofs may be found in Appendix B.

Corollary 1 (Polynomial-decay kernels)**.** *Suppose there are constants $C > 0$ and $\nu > 1/2$ such that $0 < t_j^2 \leq C j^{-2\nu}$ for all $j = 1, 2, \ldots$. Let $\lambda = n^{-\frac{2\nu}{2\nu+1}}$. Then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{ \left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2 \right) n^{-\frac{2\nu}{2\nu+1}} \right\}.$$

Corollary 2 (Exponential-decay kernels)**.** *Suppose there are constants* $C, \alpha > 0$ *such that* $0 < t_j^2 \leq Ce^{-\alpha j}$ *for all* $j = 1, 2, \ldots$. *Let* $\lambda = n^{-1}\log(n)$. *Then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{\left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2\right)\frac{\log(n)}{n}\right\}.$$

Corollary 3 (Gaussian-decay kernels)**.** *Suppose there are constants* $C, \alpha > 0$ *such that* $0 < t_j^2 \leq Ce^{-\alpha j^2}$ *for all* $j = 1, 2, \ldots$. *Additionally, assume that*

$$\sup_{\tilde{J}\in\mathbb{N},\ \mathbf{x}\in\mathcal{X}} C^{-1}e^{-\alpha\tilde{J}}\sum_{j=1}^{\tilde{J}}\psi_j(\mathbf{x})^2 \leq 1. \tag{22}$$

Let $\lambda = n^{-1}\log(n)^{1/2}$. *Then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{\left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2\right)\frac{\log(n)^{1/2}}{n}\right\}.$$

Corollary 4 (Finite rank kernels)**.** *Suppose that* $0 = t_{J+1}^2 = t_{J+2}^2 = \cdots$. *Let* $\lambda = n^{-1}$. *Then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{\left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2\right)\frac{J}{n}\right\}.$$

In Corollaries 1–4, the implicit constants in the big-O bound may depend on K and $\rho_{\mathcal{X}}$, but not on $f^\dagger$, σ^2, and n. In fact, Theorem 1 can be used to derive explicit bounds on the risk $\mathcal{R}_\rho(\hat{f}_\lambda)$ in Corollaries 1–4 that elucidate the dependence on the kernel; however, such results are not reported here for the sake of simplicity. The upper bounds in Corollaries 1–4 immediately yield minimax optimality results over ellipsoids $B_{\mathcal{H}}(r) \subseteq L^2(\rho_{\mathcal{X}})$ for the corresponding kernels and eigenvalue-decay conditions (Belitser and Levit, 1995; Caponnetto and De Vito, 2007; Golubev et al., 1996; Ibragimov and Khas'minskii, 1983). It is noteworthy that the polynomial decay kernels and corresponding RKHSs in Corollary 1 correspond to Sobolev spaces, which are of fundamental importance in nonparametric statistics and applied mathematics (Tsybakov, 2004). Additionally, Corollaries 2 and 3 are particularly relevant for the popular Gaussian kernels of the form $K(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\alpha^2||\mathbf{x} - \tilde{\mathbf{x}}||^2)$, $\alpha > 0$ (Fasshauer and McCourt, 2012; Guo et al., 2002). Note that the additional condition (22) in Corollary 3 is required to ensure that the finite-dimensional approximation error (discussed in Section 5) decays fast enough. Condition (22) holds if, for instance, the eigenfunctions $\{\psi_j\}$ are uniformly bounded. If (22) does not hold, one may obtain slightly weaker optimality results for Gaussian-decay kernels where $||f^\dagger - \hat{f}_\lambda||_{\rho_{\mathcal{X}}}^2$ converges at the minimax rate in probability, rather than expectation (Caponnetto and De Vito, 2007).

6.2. Adaptation

Our results in the previous sections apply to all regularization families satisfying conditions (R1)–(R3). In this section, we derive results that depend on the qualification of the regularization (7) and the specific regularization family.

A family of estimators $\{\hat{f}_\lambda\}_{j=1}^\infty$ is adaptive if it can achieve the minimax rate for estimating $f^\dagger$ over a wide range of regions or subsets of $\mathcal{H}$. The qualification of the regularization is related how much a given regularization family can adapt to the regularity of the signal $f^\dagger$; generally, higher qualification corresponds to better adaptation (recall that ridge regularization has qualification 1 and principal component regularization has qualification ∞). In Section 6.2.1, we show that regularization families with sufficiently high qualification are minimax over $B_{\mathcal{H}_0}(r) \subseteq \mathcal{H}_0$ for Hilbert spaces $\mathcal{H}_0 \subseteq \mathcal{H}$, which consist of functions that are "even more regular" (e.g., more smooth) than those in $\mathcal{H}$. In Section 6.2.2, we restrict our attention to principal component regularization and show that KPCR can effectively adapt to finite-rank signals. On the other hand, regularization families with lower qualification (e.g., ridge regularization) are known to saturate and the corresponding estimators are sub-optimal over subsets corresponding to highly-regular signals in many instances (Dhillon et al., 2013; Dicker, 2015; Mathé, 2005).

6.2.1. Adaptation in Sobolev Spaces (Polynomial-Decay Eigenvalues)

In this section, we assume that

$$t_j^2 = j^{-2\nu}, \quad j = 1, 2, \dots \tag{23}$$

for some $\nu > 1/2$. Let $f = \sum_{j=1}^\infty \theta_j \psi_j \in L^2(\rho_\mathcal{X})$. Then $f \in \mathcal{H}$ if and only if $\sum_{j=1}^\infty j^{2\nu}\theta_j^2 < \infty$. For $\mu \geq 0$, define

$$\mathcal{H}_\mu = \left\{ f = \sum_{j=1}^\infty \theta_j \psi_j \in L^2(\rho_\mathcal{X});\ \sum_{j=1}^\infty j^{2\nu(\mu+1)}\theta_j^2 < \infty \right\} \subseteq \mathcal{H}.$$

The $\mathcal{H}_\mu$ is itself a Hilbert space with norm $||f||_{\mathcal{H}_\mu}^2 = \sum_{j=1}^\infty j^{2\nu(\mu+1)}\theta_j^2$. Observe that $\mathcal{H} = \mathcal{H}_0$ and $||\cdot||_\mathcal{H} = ||\cdot||_{\mathcal{H}_0}$. Corollary 1 implies that $\hat{f}_\lambda$ is minimax over $B_\mathcal{H}(r)$ for any regularization family $\{g_\lambda\}_{\lambda>0}$. The next result, which is proved in Appendix B, implies that $\hat{f}_\lambda$ is minimax over $B_{\mathcal{H}_\mu}(r) \subseteq \mathcal{H}_\mu \subseteq \mathcal{H}$, provided the qualification of $\{g_\lambda\}_{\lambda>0}$ is at least $\lceil \mu/2 + 1 \rceil$.

Proposition 4. *Assume that (23) holds and that the qualification of $\{g_\lambda\}_{\lambda\geq 0}$ is at least $\lceil \mu/2 + 1 \rceil$. Let $\mu > 0$. If $f^\dagger \in \mathcal{H}_\mu$ and $\lambda = n^{-\frac{2\nu}{2\nu(\mu+1)+1}}$, then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{ \left(||f^\dagger||_{\mathcal{H}_\mu}^2 + \sigma^2 \right) n^{-\frac{2\nu(\mu+1)}{2\nu(\mu+1)+1}} \right\}.$$

As in Corollaries 1–4, the constants in the big-O bounds in this proposition and the next may depend on the kernel K. Since $\mathcal{H}_{\mu_2} \subseteq \mathcal{H}_{\mu_1}$ for $\mu_1 \leq \mu_2$, we see that greater adaptation is possible via Proposition 4 for regularization families with higher qualification. Observe that Proposition 4 does not apply to KRR for any positive $\mu > 0$, while it applies to KPCR for all $\mu \geq 0$.

6.2.2. Adaptation to Finite-Dimensional Signals

In this section, we focus on KPCR and signals (6) with $0 = \beta_{J+1} = \beta_{J+2} = \cdots$.

Proposition 5. *Let $\hat{f}_\lambda$ be the KPCR estimator with principal component regularization (8) and assume that $f^\dagger = \sum_{j=1}^J \beta_j \phi_j \in \mathcal{H}$. Fix $0 < r < 1$. If $t_J^{-2} = O(n^{1/2})$, $(Jt_J^4)^{-1} = O(1)$, and $\lambda = (1-r)t_J^2$, then*

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left\{\left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2\right)\frac{J}{n}\right\}.$$

The upper bound in Proposition 5 matches the bound in Corollary 4. However, Corollary 4 relies on the structure of the kernel (i.e., the decay of the eigenvalues $\{t_j^2\}_{j=1}^\infty$), while Proposition 5 takes advantage of the structure of the signal $f^\dagger$. It follows from Proposition 5 that the KPCR estimator is minimax rate-optimal over J-dimensional signals in $\mathcal{H}$ for a very broad class of kernels. On the other hand, it is known that KRR may perform dramatically worse than KPCR in problems where $f^\dagger$ lies in low-dimensional subspaces of the ambient space $\mathcal{H}$ due to the saturation effect (see, for example, (Dhillon et al., 2013)).

7. Discussion

Our unified analysis for a general class of regularization families in nonparametric regression highlights two important statistical properties. First, the results show minimax optimality for this general class in several commonly studied settings, which was only previously established for specific regularization methods. Second, the results demonstrate the adaptivity of certain regularization families to sub-ellipsoids of the RKHS, showing that these techniques may take advantage of additional smoothness properties that the signal may possess. It is notable that the most well-studied family, KRR/Tikhonov regularization, does not possess this adaptation property.

The minimax optimality results were obtained using *a priori* settings of the regularization parameter λ that depend on the rate of decay of the eigenvalues corresponding to the kernel K and the distribution $\rho_{\mathcal{X}}$. A data-driven estimator that chooses λ based on cross validation was proposed for a different class of regularized estimators by Caponnetto and Yao (2010) in a semi-supervised setting. Adapting a similar technique to the estimator $\hat{f}_\lambda$ studied in our work

may be needed to fully take advantage of the adaptivity of sufficiently qualified regularization families.

Kernel methods were popularized as a computational "trick" for coping with high- and even infinite-dimensional feature spaces. However, they are now often regarded as computationally prohibitive for many very large data problems on account of their standard implementations' worst-case $O(n^3)$ time and $O(n^2)$ space complexity. Approximation methods for coping with this computational complexity include Nyström approximation (Drineas and Mahoney, 2005; Williams and Seeger, 2001), matrix sketching (Gittens and Mahoney, 2013), and random features (Lopez-Paz et al., 2014; Rahimi and Recht, 2008). Recent work has considered the effect of these approximations on the statistical behavior of KRR (Bach, 2013; Yang et al., 2015). It is natural to also consider these approximations in the context of other regularization families—especially KPCR, given its connection to low-rank approximations.

Appendix A: Lemmas Required for Results in the Main Text

As in Propositions 2–3, let $J \in \mathbb{N}$ be a fixed positive integer. Define the $n \times J$ matrix $\boldsymbol{\Psi}_0 = (\psi_j(\mathbf{x}_i))_{1\leq i\leq n; 1\leq j\leq J}$ and let $\lambda_{\min}(n^{-1}\boldsymbol{\Psi}_0^\top\boldsymbol{\Psi}_0)$ be the smallest eigenvalue of $n^{-1}\boldsymbol{\Psi}_0^\top\boldsymbol{\Psi}_0$. Additionally define the event

$$\mathcal{A}(\delta) = \left\{\lambda_{\min}\left(\frac{1}{n}\boldsymbol{\Psi}_0^\top\boldsymbol{\Psi}_0\right) \geq \delta\right\}.$$

The next two lemmas are key to proving Propositions 2 and 3, respectively.

Lemma 1. *On the event* $\mathcal{A}(1/2)$,

$$\left|\left|\mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2 \leq \left(\frac{12}{t_J^2}\lambda^2 + \frac{12}{t_J^2}\delta_1^2 + 13t_{J+1}^2\right)||f^\dagger||_{\mathcal{H}}^2,$$

where δ_1 *is defined in (19)*

Proof. Let

$$I_{\mathrm{B}} = \left|\left|\mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2.$$

Then

$$\begin{aligned} I_{\mathrm{B}} &= \left|\left|\begin{pmatrix} \mathcal{T}_0^{1/2} & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\quad + \left|\left|\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1^{1/2} \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq I_{\mathrm{B}0} + t_{J+1}^2\left|\left|g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right|\right|^2 ||f^\dagger||_{\mathcal{H}}^2 \\ &\leq I_{\mathrm{B}0} + t_{J+1}^2||f^\dagger||_{\mathcal{H}}^2, \end{aligned} \tag{24}$$

where the second-to-last inequality uses the fact that $||\boldsymbol{\beta}||^2 = ||f^\dagger||_{\mathcal{H}}^2$, the last bound follows from (R2), and

$$I_{\mathrm{B}0} = \left|\left|\begin{pmatrix} \mathcal{T}_0^{1/2} & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2.$$

Since we are on the event $\mathcal{A}(1/2)$,

$$\begin{aligned} I_{\mathrm{B}0} &= \left|\left|\begin{pmatrix} \mathcal{T}_0^{1/2}\left(\frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq \left|\left|\mathcal{T}_0^{1/2}\left(\frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0\right)^{-1}\right|\right|^2\left|\left|\begin{pmatrix} \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq \frac{4}{t_J^2}\left|\left|\begin{pmatrix} \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq \frac{12}{t_J^2}\left|\left|\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\quad + \frac{12}{t_J^2}\left|\left|\begin{pmatrix} 0 & \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_1 \\ \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_0 & \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 - \mathcal{T}_1 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\quad + \frac{12}{t_J^2}\left|\left|\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq \left\{\frac{12}{t_J^2}(\lambda^2 + \delta_1^2) + 12t_{J+1}^2\right\}||f^\dagger||_{\mathcal{H}}^2 \end{aligned}$$

The lemma follows by combining this bound on $I_{\mathrm{B}0}$ with (24). □

Lemma 2. *On the event* $\mathcal{A}(1/2)$,

$$\frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\middle| X\right\} \leq \left\{2J + \frac{2J}{\lambda}(\delta_1 + t_{J+1}^2) + \frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1)\right\}\frac{\sigma^2}{n}.$$

Proof. Suppose we are on the event $\mathcal{A}(1/2)$ and let

$$\frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\middle| X\right\} = I_{\mathrm{V}0} + I_{\mathrm{V}1}, \tag{25}$$

where

$$I_{\mathrm{V}0} = \frac{1}{n^2}\mathbb{E}\left\{\left|\left|\begin{pmatrix}\mathcal{T}_0^{1/2} & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\middle| X\right\},$$
$$I_{\mathrm{V}1} = \frac{1}{n^2}\mathbb{E}\left\{\left|\left|\begin{pmatrix}0 & 0\\ 0 & \mathcal{T}_1^{1/2}\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\middle| X\right\}.$$

We bound $I_{\mathrm{V}0}$ and $I_{\mathrm{V}1}$ separately.

To bound $I_{\mathrm{V}0}$, first we use (2) to obtain

$$I_{\mathrm{V}0} \leq \frac{\sigma^2}{n^2}\mathrm{tr}\left\{\mathbf{\Phi}g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\mathcal{T}_0 & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\right\}. \tag{26}$$

If A, B are trace class operators, then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$; if, furthermore, they are positive and self-adjoint, then $\mathrm{tr}(AB) \leq ||A||\mathrm{tr}(B)$ (see, for example, Theorem 18.11 of (Conway, 1999)).

Thus,

$$\begin{aligned}
&\frac{\sigma^2}{n^2}\mathrm{tr}\left\{\mathbf{\Phi} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\mathcal{T}_0 & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\right\}\\
&\quad=\frac{\sigma^2}{n}\mathrm{tr}\left\{g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^2\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\mathcal{T}_0 & 0\\ 0 & 0\end{pmatrix}\right\}\\
&\quad=\frac{\sigma^2}{n}\mathrm{tr}\left\{\begin{pmatrix}\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{1/2} & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^2\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{1/2} & 0\\ 0 & 0\end{pmatrix}\right.\\
&\qquad\left.\cdot\begin{pmatrix}\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{-1/2}\mathcal{T}_0\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{-1/2} & 0\\ 0 & 0\end{pmatrix}\right\}\\
&\quad\le\frac{\sigma^2}{n}\left\|\begin{pmatrix}\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{1/2} & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^2\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{1/2} & 0\\ 0 & 0\end{pmatrix}\right\|\\
&\qquad\cdot\mathrm{tr}\left\{\mathcal{T}_0\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{-1}\right\}\\
&\quad=\frac{\sigma^2}{n}\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0 & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\frac{1}{\sqrt{n}}\mathbf{\Phi}^\top\right\|\\
&\qquad\cdot\mathrm{tr}\left\{\mathcal{T}_0\left(\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\right)^{-1}\right\}\\
&\quad\le\frac{2J\sigma^2}{n}\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0 & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\frac{1}{\sqrt{n}}\mathbf{\Phi}^\top\right\|,
\end{aligned}$$

where the last inequality follows from the fact that we are on the event $\mathcal{A}(1/2)$. Combining this with (26) yields

$$I_{\mathrm{V0}}\le\frac{2J\sigma^2}{n}\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\begin{pmatrix}\frac{1}{n}\mathbf{\Phi}_0^\top\mathbf{\Phi}_0 & 0\\ 0 & 0\end{pmatrix} g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\frac{1}{\sqrt{n}}\mathbf{\Phi}^\top\right\|.$$

Next, we use the decomposition (20) and the regularization conditions (R1) and (R3),

$$\begin{aligned} I_{\mathrm{V}0} &\leq \frac{2J\sigma^2}{n}\left|\left|g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^2\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^2\right|\right| \\ &\quad + \frac{2J\sigma^2}{n}\left|\left|\frac{1}{\sqrt{n}}\boldsymbol{\Phi}g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\begin{pmatrix} 0 & \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_1 \\ \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_0 & \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 - \mathcal{T}_1 \end{pmatrix} g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{\sqrt{n}}\boldsymbol{\Phi}^\top\right|\right| \\ &\quad + \frac{2J\sigma^2}{n}\left|\left|\frac{1}{\sqrt{n}}\boldsymbol{\Phi}g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1 \end{pmatrix} g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{\sqrt{n}}\boldsymbol{\Phi}^\top\right|\right| \\ &\leq \left\{2J + \frac{2J}{\lambda}(\delta_1 + t_{J+1}^2)\right\}\frac{\sigma^2}{n} \end{aligned} \tag{27}$$

Finally, to bound $I_{\mathrm{V}1}$,

$$\begin{aligned} I_{\mathrm{V}1} &\leq \frac{\sigma^2}{n}\mathrm{tr}\left\{\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1 \end{pmatrix} g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^2\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\right\} \\ &\leq \frac{\sigma^2}{n}\mathrm{tr}\left\{\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1 \end{pmatrix}\right\}\left|\left|g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^2\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\right|\right| \\ &\leq \frac{\sigma^2}{\lambda n}\mathrm{tr}(\mathcal{T}_1). \end{aligned} \tag{28}$$

The results follows by combining (25) and (27)–(28). □

In order to prove Theorem 1 and other results in the main text we require large deviation and moment bounds on quantities related to $n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}$. These bounds are collected in Lemmas 3–5 and Corollary 5 below. Lemma 3 is a general large deviation bound for sums of self-adjoint Hilbert-Schmidt operators; the other results are more tailored to the setting studied in the main text.

Lemma 3. *Let $R > 0$ be a positive real constant and consider a finite sequence of self-adjoint Hilbert-Schmidt operators $\{\mathbf{X}_i\}_{i=1}^n$ satisfying $\mathbb{E}(\mathbf{X}_i) = 0$ and $||\mathbf{X}_i|| \leq R$ almost surely. Define $\mathbf{Y} = \sum_{i=1}^n \mathbf{X}_i$ and suppose there are constants $C, d > 0$ satisfying $||\mathbb{E}(\mathbf{Y}^2)|| \leq C$ and $\mathrm{tr}\{\mathbb{E}(Y^2)\} \leq dC$. For all $t \geq C^{1/2} + R/3$,*

$$P(||\mathbf{Y}|| \geq t) \leq 4d\exp\left(\frac{-t^2/2}{C + Rt/3}\right).$$

Proof. This is a straightforward generalization of Theorem 7.7.1 of (Tropp, 2015), using the arguments from Section 4 from (Minsker, 2011) to extend from self-adjoint matrices to self-adjoint Hilbert-Schmidt operators. □

Corollary 5. *Suppose the conditions of Lemma 3 are satisfied and that $q > 0$ is a constant. Then*

$$E(||\mathbf{Y}||) \leq (1+15d)C^{1/2} + \left(\frac{1+16d}{3}\right) R,$$
$$E(||\mathbf{Y}||^2) \leq (2+32d)C + \left(\frac{2+128d}{9}\right) R^2,$$
$$E(||\mathbf{Y}||^q) = O\left\{(1+d)(C^{q/2}+R^q)\right\},$$

where the implicit constant in the big-O bound may depend on q, but not d, C, or R.

Proof. We bound $\mathbb{E}(||\mathbf{Y}||)$ first. Indeed,

$$\begin{aligned}
\mathbb{E}(||\mathbf{Y}||) &= \int_0^\infty \mathcal{P}(||\mathbf{Y}|| \geq t)\, dt \\
&\leq C^{1/2} + \frac{R}{3} + \int_{C^{1/2}+R/3}^\infty 4d \exp\left(\frac{-t^2/2}{C+Rt/3}\right)\, dt \\
&\leq C^{1/2} + \frac{R}{3} + \int_0^{3C/R} 4d\exp\left(-\frac{t^2}{4C}\right)\, dt + \int_{3C/R}^\infty 4d\exp\left(-\frac{3t}{4R}\right)\, dt \\
&= C^{1/2} + \left(\frac{1}{3} + \frac{16}{3}de^{-\frac{9C}{4R^2}}\right) R + \int_0^{3C/R} 4d\exp\left(-\frac{t^2}{4C}\right)\, dt \\
&\leq \left(1+8d\pi^{1/2}\right) C^{1/2} + \left(1+16de^{-\frac{9C}{4R^2}}\right)\frac{R}{3} \\
&\leq (1+15d)C^{1/2} + \left(\frac{1+16d}{3}\right) R.
\end{aligned}$$

For the second moment bound,

$$\begin{aligned}
\mathbb{E}(||\mathbf{Y}||^2) &= \int_0^\infty \mathcal{P}\left(||\mathbf{Y}|| \geq t^{1/2}\right)\, dt \\
&\leq \left(C^{1/2} + \frac{R}{3}\right)^2 + \int_{\left(C^{1/2}+\frac{R}{3}\right)^2}^\infty 4d\exp\left(\frac{-t/2}{C + Rt^{1/2}/3}\right)\, dt \\
&\leq \left(C^{1/2} + \frac{R}{3}\right)^2 + \int_0^{9C^2/R^2} 4d\exp\left(-\frac{t}{4C}\right)\, dt + \int_{9C^2/R^2}^\infty 4d\exp\left(-\frac{3t^{1/2}}{4R}\right)\, dt \\
&= \left(C^{1/2} + \frac{R}{3}\right)^2 + 16Cd\left\{1 - \exp\left(-\frac{9C}{4R^2}\right)\right\} + \frac{128R^2d}{9}\left\{\frac{9C}{4R^2} + 1\right\}\exp\left(-\frac{9C}{4R^2}\right) \\
&= \left(C^{1/2} + \frac{R}{3}\right)^2 + 16Cd\left\{1 + \exp\left(-\frac{9C}{4R^2}\right)\right\} + \frac{128R^2d}{9}\exp\left(-\frac{9C}{4R^2}\right) \\
&\leq \left(C^{1/2} + \frac{R}{3}\right)^2 + 32Cd + \frac{128R^2d}{9} \\
&\leq (2 + 32d)C + \left(\frac{2 + 128d}{9}\right)R^2.
\end{aligned}$$

Finally, to bound $\mathbb{E}(||\mathbf{Y}||^q)$,

$$\begin{aligned}
\mathbb{E}(||\mathbf{Y}||^q) &= \int_0^\infty \mathcal{P}\left(||\mathbf{Y}|| \geq t^{1/q}\right)\, dt \\
&\leq \left(C^{1/2} + \frac{R}{3}\right)^q + \int_{\left(C^{1/2}+\frac{R}{3}\right)^q}^\infty 4d\exp\left(\frac{-t^{2/q}/2}{C + Rt^{1/q}/3}\right)\, dt \\
&\leq \left(C^{1/2} + \frac{R}{3}\right)^q + \int_0^{(3C/R)^q} 4d\exp\left(-\frac{t^{2/q}}{4C}\right)\, dt + \int_{(3C/R)^q}^\infty 4d\exp\left(-\frac{3t^{1/q}}{4R}\right)\, dt \\
&= O\left\{(1+d)(C^{q/2} + R^q)\right\},
\end{aligned}$$

as was to be shown. □

Lemma 4. *Let δ, δ_1 be as defined in (18)–(19) and let $q > 0$ be a constant. Then*

$$\mathbb{E}(\delta) \leq \frac{16\kappa^2}{n^{1/2}} + \frac{6\kappa^2}{n}, \tag{29}$$

$$\mathbb{E}(\delta^2) \leq \frac{34\kappa^4}{n} + \frac{15\kappa^4}{n^2}, \tag{30}$$

$$\mathbb{E}(\delta^q) = O\left\{\left(\frac{\kappa^4}{n}\right)^{q/2}\right\} \tag{31}$$

and

$$\mathbb{E}(\delta_1) \leq 16\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{1/2} + \frac{6\kappa^2}{n}, \tag{32}$$

$$\mathbb{E}(\delta_1^2) \leq \frac{34\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1) + \frac{15\kappa^4}{n^2}, \tag{33}$$

$$\mathbb{E}(\delta_1^q) = O\left[\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{q/2} + \left(\frac{\kappa^2}{n}\right)^q\right]. \tag{34}$$

Proof. These bounds follow from Corollary 5. To prove the bounds on δ, we take $\mathbf{Y} = n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T} = \sum_{i=1}^n \mathbf{X}_i$ and $\mathbf{X}_i = n^{-1}(\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top - \mathcal{T})$, where $\boldsymbol{\phi}_i = (\phi_1(\mathbf{x}_i), \phi_2(\mathbf{x}_i), ...)^\top$. Clearly $\mathbb{E}(\mathbf{X}_i) = 0$ and $||\mathbf{X}_i|| \leq \kappa^2/n$. Additionally, since

$$\mathbb{E}(\mathbf{Y}^2) = \frac{1}{n}\mathbb{E}\left\{(\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top - \mathcal{T})^2\right\} = \frac{1}{n}\mathbb{E}(||\boldsymbol{\phi}_i||^2\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top) - \frac{1}{n}\mathcal{T}^2,$$

it follows that $||\mathbb{E}(\mathbf{Y}^2)|| \leq \kappa^4/n$ and $\mathrm{tr}\{\mathbb{E}(\mathbf{Y}^2)\} \leq \kappa^4/n$. The bounds (29)–(31) now follow from a direct application of Corollary 5 with $C = \kappa^4/n$, $d = 1$, and $R = \kappa^2/n$.

To prove (32)–(34), take

$$\mathbf{Y} = \begin{pmatrix} 0 & \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_1 \\ \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_0 & \frac{1}{n}\boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 - \mathcal{T}_1 \end{pmatrix} = \sum_{i=1}^n \mathbf{X}_i,$$

where

$$\mathbf{X}_i = \frac{1}{n}\begin{pmatrix} 0 & \boldsymbol{\phi}_{0,i}\boldsymbol{\phi}_{1,i}^\top \\ \boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{0,i}^\top & \boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{1,i}^\top - \mathcal{T}_1 \end{pmatrix}$$

and

$$\boldsymbol{\phi}_{0,i} = (\phi_1(\mathbf{x}_i), ..., \phi_J(\mathbf{x}_i))^\top, \quad \boldsymbol{\phi}_{1,i} = (\phi_{J+1}(\mathbf{x}_i), \phi_{J+2}(\mathbf{x}_i), ...)^\top.$$

Observe that $\mathbb{E}(\mathbf{X}_i) = 0$ and $||\mathbf{X}_i|| \leq \kappa^2/n$. Additionally, since

$$\begin{aligned}\mathbb{E}(\mathbf{Y}^2) &= n\mathbb{E}(\mathbf{X}_i^2) \\ &= \frac{1}{n}\mathbb{E}\left\{\begin{pmatrix} ||\boldsymbol{\phi}_{1,i}||^2\boldsymbol{\phi}_{0,i}\boldsymbol{\phi}_{0,i}^\top & \boldsymbol{\phi}_{0,i}\boldsymbol{\phi}_{1,i}^\top(\boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{1,i}^\top - \mathcal{T}_1) \\ (\boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{1,i}^\top - \mathcal{T}_1)\boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{0,i}^\top & (\boldsymbol{\phi}_{1,i}\boldsymbol{\phi}_{1,i}^\top - \mathcal{T}_1)^2 \end{pmatrix}\right\} \\ &= \frac{1}{n}\mathbb{E}\left(||\boldsymbol{\phi}_{1,i}||^2\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right) - \frac{1}{n}\begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1^2 \end{pmatrix}.\end{aligned}$$

it follows that $||\mathbb{E}(\mathbf{Y}^2)|| \leq \kappa^2 t_{J+1}^2/n \leq \kappa^2\mathrm{tr}(\mathcal{T}_1)/n$ and $\mathrm{tr}\{\mathbb{E}(\mathbf{Y}^2)\} \leq \kappa^2\mathrm{tr}(\mathcal{T}_1)/n$. Bounds (32)–(34) follow from Corollary 5, with $C = \kappa^2\mathrm{tr}(\mathcal{T}_1)/n$, $d = 1$, and $R = \kappa^2/n$. □

Lemma 5. *Suppose that* $\tau_J^2 \geq 0$ *and*

$$\sup_{\mathbf{x}\in X} \sum_{j=1}^{J} \psi_j(\mathbf{x})^2 \leq \frac{1}{\tau_J^2}.$$

If $0 \leq \delta < 1$*, then*

$$\Pr\{\mathcal{A}(\delta)\} \geq 1 - J\left\{\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right\}^{\tau_J^2 n}.$$

Proof. If $\tau_J^2 = 0$, the result is trivial. For $\tau_J^2 > 0$, this is a direct application of Theorem 5.1.1 from (Tropp, 2015). Indeed, consider the decomposition

$$\frac{1}{n}\boldsymbol{\Psi}_0^\top \boldsymbol{\Psi}_0 = \sum_{i=1}^{n} \mathbf{X}_i,$$

where

$$\mathbf{X}_i = \frac{1}{n}\boldsymbol{\psi}_{0,i}\boldsymbol{\psi}_{0,i}^\top$$

and $\boldsymbol{\psi}_{0,i} = (\psi_1(\mathbf{x}_i), ..., \psi_J(\mathbf{x}_i))^\top$. The lemma follows by Theorem 5.1.1 of (Tropp, 2015), upon noticing that

$$||\mathbf{X}_i|| = \left|\left|\frac{1}{n}\boldsymbol{\psi}_{0,i}\boldsymbol{\psi}_{0,i}^\top\right|\right| = \frac{1}{n}\sum_{j=1}^{J}\psi_j(\mathbf{x}_i)^2 \leq \frac{1}{\tau_J^2 n}$$

and

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{\Psi}_0^\top\boldsymbol{\Psi}_0\right) = I.$$

□

Appendix B: Proofs of Results from the Main Text

Proof of Proposition 2 We decompose $\mathrm{B}_\rho(g_\lambda)$ according to whether we are on the event $\mathcal{A}(1/2)$ or not. Let

$$\mathrm{B}_\rho(g_\lambda) = \mathrm{B}_\rho\{g_\lambda; \mathcal{A}(1/2)\} + \mathrm{B}_\rho\{g_\lambda; \mathcal{A}(1/2)^c\}, \tag{35}$$

where

$$\mathrm{B}_\rho\{g_\lambda; \mathcal{A}(1/2)\} = \mathbb{E}\left[\left|\left|\mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2; \mathcal{A}(1/2)\right],$$

$$\mathrm{B}_\rho\{g_\lambda; \mathcal{A}(1/2)\} = \mathbb{E}\left[\left|\left|\mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2; \mathcal{A}(1/2)\right].$$

By Lemmas 1 and 4,

$$\begin{aligned}\mathrm{B}_\rho\{g_\lambda;\mathcal{A}(1/2)\} &\leq \mathbb{E}\left\{\left(\frac{12}{t_J^2}\lambda^2+\frac{12}{t_J^2}\delta_1^2+13t_{J+1}^2\right)||f^\dagger||_{\mathcal{H}}^2;\mathcal{A}(1/2)\right\}\\ &\leq \left\{\frac{12}{t_J^2}\lambda^2+\frac{12}{t_J^2}\mathbb{E}(\delta_1^2)+13t_{J+1}^2\right\}||f^\dagger||_{\mathcal{H}}^2\\ &\leq \left\{\frac{12}{t_J^2}\lambda^2+\frac{408\kappa^4}{nt_J^2}\mathrm{tr}(\mathcal{T}_1)+\frac{180\kappa^4}{n^2t_J^2}+13t_{J+1}^2\right\}||f^\dagger||_{\mathcal{H}}^2. \end{aligned} \tag{36}$$

On the other hand, since

$$\left|\left|\mathcal{T}^{1/2}\left\{I-g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right\}\boldsymbol{\beta}\right|\right|^2 \leq \kappa^2||f^\dagger||_{\mathcal{H}}^2,$$

it follows from Lemma 5 that

$$\mathrm{B}_\rho\{g_\lambda;\mathcal{A}(1/2)^c\} \leq \kappa^2||f^\dagger||_{\mathcal{H}}^2\Pr\{\mathcal{A}(1/2)^c\} \leq \kappa^2||f^\dagger||_{\mathcal{H}}^2 J\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}. \tag{37}$$

The proposition follows by combining (35)–(37). ■

Proof of Proposition 3 The steps of the proof are similar to those in the proof of Proposition 2. Let

$$\mathrm{V}_\rho(g_\lambda) = \mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)\}+\mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)^c\}, \tag{38}$$

where

$$\begin{aligned}\mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)\} &= \frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\boldsymbol{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2;\mathcal{A}(1/2)\right\},\\ \mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)^c\} &= \frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\boldsymbol{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2;\mathcal{A}(1/2)^c\right\}.\end{aligned}$$

By Lemmas 1 and 4,

$$\begin{aligned}\mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)\} &\leq \mathbb{E}\left[\left\{2J+\frac{2J}{\lambda}(\delta_1+t_{J+1}^2)+\frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1)\right\}\frac{\sigma^2}{n};\mathcal{A}(1/2)\right]\\ &\leq \left[2J+\frac{2J}{\lambda}\{\mathbb{E}(\delta_1)+t_{J+1}^2\}+\frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1)\right]\frac{\sigma^2}{n}\\ &\leq \left[2J+\frac{32J}{\lambda}\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{1/2}+\frac{12\kappa^2J}{\lambda n}+\frac{2Jt_{J+1}^2}{\lambda}+\frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1)\right]\frac{\sigma^2}{n}. \end{aligned} \tag{39}$$

On the other hand, since

$$\frac{1}{n^2}\mathbb{E}\left\{\left|\left|\mathcal{T}^{1/2}g_\lambda\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\mathbf{\Phi}^\top\boldsymbol{\epsilon}\right|\right|^2\middle|\mathbf{X}\right\} \leq \frac{\kappa^2\sigma^2}{\lambda n},$$

it follows from Lemma 5 that

$$\mathrm{V}_\rho\{g_\lambda;\mathcal{A}(1/2)^c\} \leq \frac{\kappa^2\sigma^2}{\lambda n}\Pr\{\mathcal{A}(1/2)^c\} \leq \frac{\kappa^2\sigma^2 J}{\lambda n}\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}. \tag{40}$$

The proposition follows by combining (38)–(40). ■

Proof of Corollary 1 Choose J in Theorem 1 so that $t_{J+1}^2 < \lambda \leq t_J^2$. Then, $J \leq C^{\frac{1}{2\nu}}\lambda^{-\frac{1}{2\nu}}$, $n = \lambda^{-\left(1+\frac{1}{2\nu}\right)}$ and, applying Theorem 1 with $\tau_j^2 = t_j^2/\kappa^2$,

$$\begin{aligned}\mathcal{R}_\rho(\hat{f}_\lambda) \leq &\left\{20C^{\frac{1}{2\nu}}\lambda + (1+16C^{\frac{1}{2\nu}}\kappa^2)\lambda^{\frac{1}{2\nu}}\sum_{j=J+1}^{\infty}t_j^2\right\}\sigma^2 \\ &+\left\{25\lambda + 408\kappa^2\lambda^{\frac{1}{2\nu}}\sum_{j=J+1}^{\infty}t_j^2\right\}||f^\dagger||_{\mathcal{H}}^2 + o\left\{\left(\sigma^2+||f^\dagger||_{\mathcal{H}}^2\right)\lambda\right\}.\end{aligned}$$

Since $\lambda = n^{-\frac{2\nu}{2\nu+1}}$, the corollary will follow if we can show that

$$\sum_{j=J+1}^{\infty}t_j^2 = O(\lambda^{1-\frac{1}{2\nu}}). \tag{41}$$

Let $J_0 = \left\lfloor C^{\frac{1}{2\nu}}\lambda^{-\frac{1}{2\nu}}\right\rfloor + 1$. Then

$$\begin{aligned}\sum_{j=J+1}^{\infty}t_j^2 &\leq \sum_{j=J+1}^{J_0}t_j^2 + \sum_{j=J_0+1}^{\infty}t_j^2 \\ &\leq J_0\lambda + C\int_{J_0}^{\infty}t^{-2\nu}\,dt \\ &\leq J_0\lambda + \frac{C}{2\nu-1}J_0^{1-2\nu} \\ &\leq \left(1+\frac{2\nu}{2\nu-1}C^{\frac{1}{2\nu}}\right)\lambda^{1-\frac{1}{2\nu}},\end{aligned} \tag{42}$$

which yields (41). This completes the proof of Corollary 1. ■

Proof of Corollary 2 Choose J in Theorem 1 so that

$$t_{J+1}^2 < \frac{2\kappa^2}{\log(e/2)}\lambda \leq t_J^2.$$

Then $J = O\{\log(n)\}$ and it follows from Theorem 1 (with $\tau_j^2 = t_j^2/\kappa^2$) that

$$\mathcal{R}_\rho(\hat{f}_\lambda) \leq O\left[\left(\sigma^2 + ||f^\dagger||_{\mathcal{H}}^2\right)\left\{\lambda + \frac{1}{\log(n)}\sum_{j=J+1}^{\infty} t_j^2\right\}\right].$$

The corollary will follow if we can show that

$$\sum_{j=J+1}^{\infty} t_j^2 = O\left\{\frac{\log(n)^2}{n}\right\}.$$

Let $J_0 = \lfloor \alpha^{-1}\log(n)\rfloor + 1$. Then

$$\sum_{j=J+1}^{\infty} t_j^2 \leq J_0 t_{J+1}^2 + \sum_{j=J_0+1}^{\infty} t_j^2 \leq O\left\{\frac{\log(n)^2}{n}\right\} + C\sum_{j=J_0+1} e^{-\alpha j} = O\left\{\frac{\log(n)^2}{n}\right\},$$

as was to be shown. ■

Proof of Corollary 3 Choose J in Theorem 1 so that $t_{J+1}^2 < \lambda \leq t_J^2$. Then

$$J = O\left\{\log(n)^{1/2}\right\}$$

and, by Theorem 1 with $\tau_j^2 = C^{-1}e^{-\alpha j}$,

$$\mathcal{R}_\rho(\hat{f}_\lambda) = O\left[\left(\sigma^2 + ||f^\dagger||_{\mathcal{H}}^2\right)\left\{\lambda + \frac{1}{\log(n)^{1/2}}\sum_{j=J+1}^{\infty} t_j^2\right\}\right].$$

The corollary will follow if we can show that

$$\sum_{j=J+1}^{\infty} t_j^2 = O\left\{\frac{\log(n)}{n}\right\}.$$

Let

$$J_0 = \left\lceil \left\{\frac{1}{\alpha}\log\left(\frac{C}{2\alpha\lambda}\right)\right\}^{1/2}\right\rceil.$$

Then

$$\sum_{j=J+1}^{\infty} t_j^2 \leq J_0\lambda + C\int_{J_0}^{\infty} e^{-\alpha t^2}\, dt \leq J_0\lambda + \frac{C}{2\alpha J_0}e^{-\alpha J_0^2} = O\left\{\frac{\log(n)}{n}\right\},$$

as was to be shown. ■

Proof of Corollary 4 To prove this corollary, the strategy is the same as in Corollaries 1–4. However, we must modify the bounds in Lemmas 1-2. First note that $\mathcal{T}_1 = 0$. Then following the notation in the proof of Lemmas 1-2, on the event $\mathcal{A}(1/2)$,

$$\begin{aligned} I_{\mathrm{B}} &= I_{\mathrm{B}0} \\ &\leq \left|\left|\mathcal{T}_0^{1/2}\left(\frac{1}{n}\boldsymbol{\Phi}_0^{\top}\boldsymbol{\Phi}_0\right)^{-1/2}\right|\right|^2 \left|\left|\begin{pmatrix} \left(\frac{1}{n}\boldsymbol{\Phi}_0^{\top}\boldsymbol{\Phi}_0\right)^{1/2} & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq \mathrm{tr}\left\{\mathcal{T}_0\left(\frac{1}{n}\boldsymbol{\Phi}_0^{\top}\boldsymbol{\Phi}_0\right)^{-1}\right\} \left|\left|\begin{pmatrix} \left(\frac{1}{n}\boldsymbol{\Phi}_0^{\top}\boldsymbol{\Phi}_0\right)^{1/2} & 0 \\ 0 & 0 \end{pmatrix}\left\{g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} - I\right\}\boldsymbol{\beta}\right|\right|^2 \\ &\leq 2J\lambda||f^{\dagger}||_{\mathcal{H}}^2. \end{aligned}$$

It follows that

$$\mathrm{B}_\rho(g_\lambda) \leq \left\{2J\lambda + \kappa^2 J\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}\right\}||f^{\dagger}||_{\mathcal{H}}^2. \tag{43}$$

Additionally, on the event $\mathcal{A}(1/2)$,

$$I_{\mathrm{V}} = I_{\mathrm{V}0} \leq \frac{\sigma^2}{n}\mathrm{tr}\left\{\mathcal{T}_0\left(\frac{1}{n}\boldsymbol{\Phi}_0^{\top}\boldsymbol{\Phi}_0\right)^{-1}\right\}\left|\left|g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)\left(\frac{1}{n}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)\right|\right|^2 \leq \frac{2J\sigma^2}{n}.$$

Thus,

$$\mathrm{V}_\rho(g_\lambda) \leq \left\{2J + \frac{\kappa^2 J}{\lambda}\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}\right\}\frac{\sigma^2}{n}. \tag{44}$$

Combining Proposition 1 and (43)–(44), we obtain

$$\mathcal{R}_\rho(\hat{f}_\lambda) \leq \left\{2J\lambda + \kappa^2 J\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}\right\}||f^{\dagger}||_{\mathcal{H}}^2 + \left\{2J + \frac{\kappa^2 J}{\lambda}\left(\frac{2}{e}\right)^{\frac{\tau_J^2 n}{2}}\right\}\frac{\sigma^2}{n}.$$

Corollary 4 follows immediately. ■

Proof of Proposition 4 Pick J so that $t_{J+1}^2 < \lambda \leq t_J^2$. For real numbers $\mu_1, \mu_2 \geq 0$ and functions $h : [0, \infty) \to [0, \infty)$, define

$$K_{\mu_1,\mu_2}(h) = \left|\left| \mathcal{T}^{\mu_1} h\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\mathcal{T}^{\mu_2}\right|\right|.$$

Also define the operator S by $Sh(t) = th(t)$. Let $h_\lambda(t) = 1 - g_\lambda(t)t$. Finally, suppose that

$$f^\dagger = \sum_{j=1}^\infty \beta_j\phi_j = \sum_{j=1}^\infty \alpha_j t_j^\mu \phi_j \in H_\mu.$$

Then $||f^\dagger||_{\mathcal{H}_\mu}^2 = ||\boldsymbol{\alpha}||^2$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ...)^\top$.

To prove the proposition, we follow the same strategy as in the proofs of Propositions 2–3; however, several modifications are required along the way. Following the notation of Lemmas 1–2, our first task is to bound I_{B}. We have

$$\begin{aligned} I_{\mathrm{B}} &= \left|\left| \mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\right\}\boldsymbol{\beta}\right|\right|^2 \\ &= \left|\left| \mathcal{T}^{1/2}\left\{I - g_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\right\}\mathcal{T}^{\mu/2}\boldsymbol{\alpha}\right|\right|^2 \\ &\leq K_{1/2,\mu/2}(h_\lambda)^2 ||f^\dagger||_{\mathcal{H}_\mu}^2. \end{aligned} \tag{45}$$

Furthermore, on $\mathcal{A}(1/2)$,

$$\begin{aligned} K_{1/2,\mu/2}(h_\lambda) &= \left|\left| \begin{pmatrix} \mathcal{T}_0^{1/2} & 0 \\ 0 & 0 \end{pmatrix} h_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\mathcal{T}^{\mu/2}\right|\right| + \left|\left| \begin{pmatrix} 0 & 0 \\ 0 & \mathcal{T}_1^{1/2} \end{pmatrix} h_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\mathcal{T}^{\mu/2}\right|\right| \\ &\leq \left|\left| \mathcal{T}_0^{1/2}\left(\frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0\right)^{-1}\right|\right| \left|\left| \begin{pmatrix} \frac{1}{n}\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 & 0 \\ 0 & 0 \end{pmatrix} h_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\mathcal{T}^{\mu/2}\right|\right| \\ &\quad + t_{J+1}K_{0,\mu/2}(h_\lambda) \\ &\leq \frac{2}{t_J}\left\{K_{0,\mu/2}(Sh_\lambda) + (\delta_1 + t_{J+1}^2)K_{0,\mu/2}(h_\lambda)\right\} + t_{J+1}K_{0,\mu/2}(h_\lambda) \\ &\leq \frac{2}{t_J}K_{0,\mu/2}(Sh_\lambda) + \left(\frac{2\delta_1}{t_J} + 3t_{J+1}\right)K_{0,\mu/2}(h_\lambda). \end{aligned} \tag{46}$$

Next we bound $K_{0,\mu/2}(h_\lambda)$ and $K_{0,\mu/2}(Sh_\lambda)$. Suppose that $\mu/2 = m + \gamma$, where $m \geq 0$ is an integer and $0 \leq g < 1$. Recall the definition of δ from (18). By repeated application of the

triangle inequality,

$$\begin{aligned}
K_{0,\mu/2}(h) &= K_{m+\gamma,0}(h) \\
&= \left|\left|\mathcal{T}^{m+\gamma} h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| \\
&\leq \left|\left|\mathcal{T}^{m-1+\gamma} Sh\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| + \delta\kappa^{2(m-1+\gamma)}\left|\left|h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| \\
&\vdots \\
&\leq \left|\left|\mathcal{T}^{\gamma} S^m h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| + \delta\sum_{j=1}^{m}\kappa^{2(m-j+\gamma)}\left|\left|S^{j-1} h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| \\
&= K_{0,\gamma}(S^m h) + \delta\sum_{j=1}^{m}\kappa^{2(m-j+\gamma)}\left|\left|S^{j-1} h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right|. \qquad (47)
\end{aligned}$$

Bounding $K_{0,\gamma}(S^m h)$ will require some additional manipulations. However, before proceeding in this direction, we gather our bounds so far. By (46)–(47) and because $\{g_\lambda\}$ has qualification at least $\lceil \mu/2 \rceil$,

$$K_{1/2,\mu/2}(h_\lambda) = O\left\{\frac{1}{\lambda^{1/2}}K_{0,\gamma}(S^{m+1}h_\lambda) + \left(\frac{\delta_1}{\lambda^{1/2}} + \lambda^{1/2}\right)K_{0,\gamma}(S^m h_\lambda) + \delta\lambda^{1/2} + \frac{\delta_1\delta}{\lambda^{1/2}}\right\}$$

on $\mathcal{A}$. Combining this with (45) yields

$$\begin{aligned}
\mathrm{B}_\rho(g_\lambda) &= \mathbb{E}\{I_\mathrm{B}; \mathcal{A}(1/2)\} + \mathbb{E}\{I_\mathrm{B}; \mathcal{A}(1/2)^c\} \\
&\leq O\left[\mathbb{E}\left\{\frac{1}{\lambda}K_{0,\gamma}(S^{m+1}h_\lambda)^2 + \left(\frac{\delta_1^2}{\lambda} + \lambda\right)K_{0,\gamma}(S^m h_\lambda)^2 + \delta^2\lambda + \frac{\delta_1^2\delta^2}{\lambda}\right\}||f^\dagger||_{\mathcal{H}_\mu}^2\right] \\
&\quad + \kappa^{2(\mu+1)}||f^\dagger||_{\mathcal{H}_\mu}^2 \Pr\{\mathcal{A}(1/2)^c\}.
\end{aligned}$$

By Lemma 4,

$$\begin{aligned}
\mathbb{E}(\delta^2) &= O\left(\frac{1}{n}\right), \\
E(\delta_1^2\delta^2) &= O\left\{\frac{\mathrm{tr}(\mathcal{T}_1)}{n^2} + \frac{1}{n^3}\right\}.
\end{aligned}$$

(Note that in the current proof, we allow the big-O constants to depend on the kernel K, while this is not allowed in Lemma 4; see the comment following the statement of Proposition

4 and the statement of Corollary 5.) Thus,

$$\begin{aligned}
\mathrm{B}_\rho(g_\lambda) &\leq O\left[\mathbb{E}\left\{\frac{1}{\lambda}K_{0,\gamma}(S^{m+1}h_\lambda)^2 + \left(\frac{\delta_1^2}{\lambda}+\lambda\right)K_{0,\gamma}(S^m h_\lambda)^2\right\}||f^\dagger||_{\mathcal{H}_\mu}^2\right] \\
&\quad + O\left[\frac{\lambda}{n}||f^\dagger||_{\mathcal{H}_\mu}^2 + \frac{1}{\lambda}\left\{\frac{\mathrm{tr}(\mathcal{T}_1)}{n^2}+\frac{1}{n^3}\right\}||f^\dagger||_{\mathcal{H}_\mu}^2\right] + \kappa^{2(\mu+1)}||f^\dagger||_{\mathcal{H}_\mu}^2 \Pr\left\{\mathcal{A}(1/2)^c\right\} \\
&= O\left[\mathbb{E}\left\{\frac{1}{\lambda}K_{0,\gamma}(S^{m+1}h_\lambda)^2 + \left(\frac{\delta_1^2}{\lambda}+\lambda\right)K_{0,\gamma}(S^m h_\lambda)^2\right\}||f^\dagger||_{\mathcal{H}_\mu}^2\right] \\
&\quad + o\left(||f^\dagger||_{\mathcal{H}}^2\lambda^{\mu+1}\right), \qquad (48)
\end{aligned}$$

where we have also made use of Lemma 5 and (42) to bound $\Pr\{\mathcal{A}(1/2)^c\}$ and $\mathrm{tr}(\mathcal{T}_1)$, respectively.

To complete our analysis of the bias term $\mathrm{B}_\rho(g_\lambda)$, it remains to bound the expectations involving $K_{0,\gamma}$ above. We consider separately the cases where $0 \leq \gamma \leq 1/2$ and $1/2 < \gamma < 1$. If $1/2 < \gamma < 1$ and $h : [0,\infty) \to [0,\infty)$ is a nonnegative function, then, since $z \mapsto z^\gamma$ is operator monotone (Mathé and Pereverzev, 2002),

$$\begin{aligned}
K_{0,\gamma}(h) &= \left|\left|\mathcal{T}^\gamma h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| \\
&\leq \left|\left|\mathcal{T}^\gamma - \left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^\gamma\right|\right| K_{0,0}(h) + \left|\left|\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^\gamma h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right| \\
&\leq \delta^\gamma K_{0,0}(h) + \left|\left|\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^\gamma h\left(\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi}\right)\right|\right|.
\end{aligned}$$

Hence, taking $h = S^m h_\lambda$ and $h = S^{m+1}h_\lambda$, we conclude that if $1/2 < \gamma < 1$, then

$$\begin{aligned}
K_{0,\gamma}(S^m h_\lambda) &\leq \delta^\gamma\lambda^m + \lambda^{\mu/2}, \\
K_{0,\gamma}(S^{m+1}h_\lambda) &\leq \delta^\gamma\lambda^{m+1} + \lambda^{\mu/2+1}.
\end{aligned}$$

Thus, by Lemmas 4 and , if $q > 0$ is constant and $1/2 < \gamma < 1$, then

$$\mathbb{E}\left\{K_{0,\gamma}(S^m h_\lambda)^q\right\} = O\left\{n^{-q\gamma/2}\lambda^{qm} + \lambda^{q\mu/2}\right\} = O(\lambda^{q\mu/2}), \qquad (49)$$

$$\mathbb{E}\left\{K_{0,\gamma}(S^{m+1}h_\lambda)^q\right\} = O\left\{n^{-q\gamma/2}\lambda^{q(m+1)} + \lambda^{q(\mu/2+1)}\right\} = O\{\lambda^{q(\mu/2+1)}\}. \qquad (50)$$

On the other hand, if $0 \leq \gamma \leq 1/2$, then, on $\mathcal{A}(1/2)$,

$$\begin{aligned}
K_{0,\gamma}(h) &= \left|\left|\mathcal{T}^{\gamma} h\left(\frac{1}{n}\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)\right|\right| \\
&\leq \left|\left|\begin{pmatrix} \mathcal{T}_0^{\gamma} & 0 \\ 0 & 0 \end{pmatrix} h\left(\frac{1}{n}\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)\right|\right| + t_{J+1}^{2\gamma} K_{0,0}(h) \\
&\leq \left|\left|\mathcal{T}_0^{\gamma}\left(\frac{1}{n}\mathbf{\Phi}_0^{\top}\mathbf{\Phi}_0\right)^{-1}\right|\right| \left|\left|\begin{pmatrix} \frac{1}{n}\mathbf{\Phi}_0^{\top}\mathbf{\Phi}_0 & 0 \\ 0 & 0 \end{pmatrix} h\left(\frac{1}{n}\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)\right|\right| + t_{J+1}^{2\gamma} K_{0,0}(h) \\
&\leq 2t_J^{2(\gamma-1)}\left\{K_{0,0}(Sh) + (\delta_1 + t_{J+1}^2) K_{0,0}(h)\right\} + t_{J+1}^{2\gamma} K_{0,0}(h) \\
&\leq 2\lambda^{\gamma-1} K_{0,0}(Sh) + 2\lambda^{\gamma-1}\delta_1 K_{0,0}(h) + 3\lambda^{\gamma} K_{0,0}(h).
\end{aligned}$$

Again taking $h = S^m h_\lambda$ and $h = S^{m+1} h_\lambda$, if $0 \leq \gamma \leq 1/2$ and we are on $\mathcal{A}(1/2)$, then

$$\begin{aligned}
K_{0,\gamma}(S^m h_\lambda) &\leq 2\delta_1 \lambda^{\mu/2-1} + 5\lambda^{\mu/2}, \\
K_{0,\gamma}(S^{m+1} h_\lambda) &\leq 2\delta_1 \lambda^{\mu/2} + 5\lambda^{\mu/2+1}.
\end{aligned}$$

Taking expectations, Lemma 4–5 imply that if $0 \leq \gamma \leq 1$, then

$$\mathbb{E}\left\{K_{0,\gamma}(S^m h_\lambda)^q\right\} = O\left[\left\{\frac{\mathrm{tr}(\mathcal{T}_1)}{n}\right\}^{q/2} \lambda^{q(\mu/2-1)} + \frac{1}{n^q}\lambda^{q(\mu/2-1)} + \lambda^{q\mu/2}\right] = O(\lambda^{q\mu/2}), \tag{51}$$

$$\mathbb{E}\left\{K_{0,\gamma}(S^{m+1} h_\lambda)^q\right\} = O\left[\left\{\frac{\mathrm{tr}(\mathcal{T}_1)}{n}\right\}^{q/2} \lambda^{q\mu/2} + \frac{1}{n^r}\lambda^{q\mu/2} + \lambda^{q(\mu/2+1)}\right] = O\{\lambda^{q(\mu/2+1)}\}. \tag{52}$$

Thus, combining (49)–(52), we see that

$$\mathbb{E}\left\{K_{0,\gamma}(S^m h_\lambda)^q\right\} = O(\lambda^{q\mu/2}), \tag{53}$$

$$\mathbb{E}\left\{K_{0,\gamma}(S^{m+1} h_\lambda)^q\right\} = O\{\lambda^{q(\mu/2+1)}\} \tag{54}$$

for all $0 \leq \gamma < 1$. Finally, using (53)–(54) in (48), along with Lemma 4 yields

$$\begin{aligned}
\mathrm{B}_\rho(g_\lambda) &= O\left[\mathbb{E}\left\{\frac{1}{\lambda}K_{0,\gamma}(S^{m+1}h_\lambda)^2 + \left(\frac{\delta_1^2}{\lambda} + \lambda\right) K_{0,\gamma}(S^m h_\lambda)^2\right\} ||f^\dagger||_{\mathcal{H}_\mu}^2\right] \\
&\quad + o\left(||f^\dagger||_{\mathcal{H}}^2 \lambda^{\mu+1}\right) \\
&= O\left(\lambda^{\mu+1}||f^\dagger||_{\mathcal{H}}^2\right).
\end{aligned} \tag{55}$$

To complete the proof, we must bound $\mathrm{V}_\rho(g_\lambda)$, but this is a direct application of Proposition 3: One easily checks that

$$\mathrm{V}_\rho(g_\lambda) = O(\lambda^{\mu+1}\sigma^2). \tag{56}$$

The proposition follows by combining (55)–(56) with Proposition 1.

■

Proof of Proposition 5 Let $\hat{t}_1^2 \geq \hat{t}_2^2 \geq \cdots \geq 0$ denote the eigenvalues of $n^{-1}\boldsymbol{\Phi}^{\mathcal{T}}\boldsymbol{\Phi}$ and define $\hat{\mathcal{T}} = \text{diag}(\hat{t}_1^2, \hat{t}_2^2, ...)$. Let $\hat{U}$ be an orthogonal transformation satisfying $n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \hat{U}\hat{\mathcal{T}}\hat{U}^\top$. Additionally, let $h_\lambda(t) = I\{t \leq \lambda\}$, define $\hat{J} = \hat{J}_\lambda = \inf\{j; \hat{t}_j^2 > \lambda\}$, and write $\hat{U} = (\hat{U}_{\hat{J}} \;\; \hat{U}_{\hat{J}^c})$, where $\hat{U}_{\hat{J}}$ is the $\infty \times \hat{J}$ matrix comprised of the first $\hat{J}$ columns of $\hat{U}$ and $\hat{U}_{\hat{J}^c}$ consists of the remaining columns of $\hat{U}$. Finally, define $\hat{\mathcal{T}}_0 = \text{diag}(\hat{t}_{\hat{J}+1}^2, \hat{t}_{\hat{J}+2}^2, ...)$ and define the $\infty \times J$ matrix $U_J = (I_J \;\; 0)^\top$.

Following the notation from the proof of Lemma 1, consider the following bound on I_{B},

$$I_{\text{B}} = \left|\left|\mathcal{T}^{1/2} h_\lambda\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\boldsymbol{\beta}\right|\right|^2 = \left|\left|\mathcal{T}^{1/2}\hat{U}_{\hat{J}^c}\hat{U}_{\hat{J}^c}^\top U_J U_J^\top \boldsymbol{\beta}\right|\right|^2 \leq \kappa^2 \left|\left|\hat{U}_{\hat{J}^c}^{|top} U_J\right|\right|^2 ||f^\dagger||_{\mathcal{H}}^2.$$

It follows that

$$\text{B}_\rho(g_\lambda) = \mathbb{E}(I_{\text{B}}) \leq \kappa^2 ||f^\dagger||_{\mathcal{H}}^2 \mathbb{E}\left(\left|\left|\hat{U}_{\hat{J}^c}^\top U_J\right|\right|^2 ; \hat{J} \geq J\right) + \kappa^2 ||f^\dagger||_{\mathcal{H}}^2 \Pr(\hat{J} < J). \tag{57}$$

Now we bound $||\hat{U}_{\hat{J}^c}^\top U_J||$ on the event $\{\hat{J} \geq J\}$. We derive this bound from basic principles, but it is essentially the Davis-Kahan inequality (Davis and Kahan, 1970). Let $D = n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}$. Then

$$DU_J = \frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}U_J - \mathcal{T}U_J = \frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}U_J - U_J\mathcal{T}_0$$

and

$$U_J^\top D\hat{U}_{\hat{J}^c} = U_J^\top\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)\hat{U}_{\hat{J}^c} - \mathcal{T}_0 U_J^\top\hat{U}_{\hat{J}^c} = U_J^\top\hat{U}_{\hat{J}^c}\hat{\mathcal{T}}_1 - \mathcal{T}_0 U_J^\top\hat{U}_{\hat{J}^c}.$$

Now observe that

$$\begin{aligned}\left|\left|\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}\right|\right| &\geq ||U_J^\top\Delta\hat{U}_{\hat{J}^c}|| \\ &\geq ||\mathcal{T}_0 U_J^\top\hat{U}_{\hat{J}^c}|| - ||U_J^\top\hat{U}_{\hat{J}^c}\hat{\mathcal{T}}_1|| \\ &\geq t_J^2||U_J^\top\hat{U}_{\hat{J}^c}|| - (1-r)t_J^2||U_J^\top\hat{U}_{\hat{J}^c}|| \\ &= rt_J^2||U_J^\top\hat{U}_{\hat{J}^c}||.\end{aligned}$$

Thus,

$$||U_J^\top\hat{U}_{\hat{J}^c}|| \leq \frac{1}{rt_J^2}\left|\left|\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathcal{T}\right|\right| \tag{58}$$

on $\{\hat{J} > J\}$.

Next we bound

$$\Pr(\hat{J} < J) = \Pr(\hat{t}_J^2 \leq \lambda) = \Pr\left\{\hat{t}_J^2 \leq (1-r)t_J^2\right\}.$$

By Weyl's inequality,

$$|\hat{t}_J^2 - t_J^2| \leq \delta = \left|\left|\frac{1}{n}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathcal{T}\right|\right|.$$

Additionally, by Lemma 3,

$$\Pr(\delta \geq rt_J^2) \leq 4\exp\left(-\frac{nr^2t_J^4}{2\kappa^4 + 2\kappa^2 rt_J^2/3}\right)$$

Thus,

$$\Pr(\hat{J} < J) = \Pr\left\{\hat{t}_J^2 \leq (1-r)t_J^2\right\} \leq \Pr(\delta \geq rt_J^2) \leq 4\exp\left(-\frac{nr^2t_J^4}{2\kappa^4 + 2\kappa^2 rt_J^2/3}\right). \tag{59}$$

Combining (57)–(59) and using Lemma 4 gives

$$\begin{aligned}
\mathrm{B}_\rho(g_\lambda) &\leq \frac{\kappa^2}{r^2t_J^4}||f^\dagger||_{\mathcal{H}}^2\mathbb{E}(\delta^2) + 4\kappa^2||f^\dagger||_{\mathcal{H}}^2\exp\left(-\frac{nr^2t_J^4}{2\kappa^4 + 2\kappa^2 rt_J^2/3}\right) \\
&\leq \frac{\kappa^2}{r^2t_J^4}||f^\dagger||_{\mathcal{H}}^2\left(\frac{34\kappa^4}{n} + \frac{15\kappa^4}{n^2}\right) + 4\kappa^2||f^\dagger||_{\mathcal{H}}^2\exp\left(-\frac{nr^2t_J^4}{2\kappa^4 + 2\kappa^2 rt_J^2/3}\right).
\end{aligned}$$

Next, we combine this bound on $\mathrm{B}_\rho(g_\lambda)$ with Proposition 3 to obtain

$$\begin{aligned}
\mathcal{R}_\rho(\hat{f}_\lambda) &= \mathrm{B}_\rho(g_\lambda) + \mathrm{V}_\rho(g_\lambda) \\
&\leq \frac{\kappa^2}{r^2t_J^4}||f^\dagger||_{\mathcal{H}}^2\left(\frac{34\kappa^4}{n} + \frac{15\kappa^4}{n^2}\right) + 4\kappa^2||f^\dagger||_{\mathcal{H}}^2\exp\left(-\frac{nr^2t_J^4}{2\kappa^4 + 2\kappa^2 rt_J^2/3}\right) \\
&\quad + \left[2J + \frac{2Jt_{J+1}^2}{\lambda} + \frac{1}{\lambda}\mathrm{tr}(\mathcal{T}_1) + \frac{32J}{\lambda}\left\{\frac{\kappa^2}{n}\mathrm{tr}(\mathcal{T}_1)\right\}^{1/2} + \frac{12\kappa^2 J}{\lambda n} + \frac{\kappa^2 J}{\lambda}\left(\frac{2}{e}\right)^{\frac{t_J^2 n}{2\kappa^2}}\right]\frac{\sigma^2}{n} \\
&= O\left\{\left(||f^\dagger||_{\mathcal{H}}^2 + \sigma^2\right)\frac{J}{n}\right\},
\end{aligned}$$

as was to be shown. ■

References

Aronszajn, N. (1950). Theory of reproducing kernels. *T. A. Math. Soc.* **68** 337–404.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*.

Bauer, F., Pereverzev, S. and Rosasco, L. (2007). On regularization algorithms in learning theory. *J. Complexity* **23** 52–72.

Belitser, E. and Levit, B. (1995). On minimax filtering over ellipsoids. *Math. Meth. Statist* **4** 259–273.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368.

Caponnetto, A. and Yao, Y. (2010). Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.* **8** 161–183.

Carmeli, C., De Vito, E. and Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl.* **4** 377–408.

Conway, J. (1999). *A Course in Operator Theory*. American Mathematical Society.

Davis, C. and Kahan, W. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.

Dhillon, P., Foster, D., Kakade, S. and Ungar, L. (2013). A risk comparison of ordinary least squares vs ridge regression. *J. Mach. Learn. Res.* **14** 1505–1511.

Dicker, L. (2015). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Benoulli* To appear.

Drineas, P. and Mahoney, M. (2005). On the Nyström method for approximating a Gram matrix for improving kernel-based learning. *J. Mach. Learn. Res.* **6** 2153–2175.

Engl, H., Hanke, M. and Neubauer, A. (1996). *Regularization of Inverse Problems*. Mathematics and Its Applications, Vol. 375, Springer.

Fasshauer, G. and McCourt, M. (2012). Stable evaluation of Gaussian radial basis function interpolants. *SIAM J. Sci. Comput.* **34** A737–A762.

Gittens, A. and Mahoney, M. (2013). Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning*.

Golubev, Y., Levit, B. and Tsybakov, A. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2** 167–181.

Guo, Y., Bartlett, P., Shawe-Taylor, J. and Williamson, R. (2002). Covering numbers for support vector machines. *IEEE T. Inform. Theory* **48** 239–250.

Gyorfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A Distribution Free Theory of Nonparametric Regression*. Springer.

Hsu, D., Kakade, S. and Zhang, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600.

Ibragimov, I. and Khas'minskii, R. (1983). Estimation of distribution density belonging to a class of entire functions. *Theor. Probab. Appl+* **27** 551–562.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk

minimization. *Ann. Stat.* **34** 2593–2656.

Lo Gerfo, L., Rosasco, L., Odone, F., De Vito, E. and Verri, A. (2008). Spectral algorithms for supervised learning. *Neural Comput.* **7** 1873–1897.

Lopez-Paz, D., Sra, S., Smola, A., Ghahramani, Z. and Schölkopf, B. (2014). Randomized nonlinear component analysis. In *International Conference on Machine Learning*.

Mathé, P. (2005). Saturation of regularization methods for linear ill-posed problems in Hilbert spaces. *SIAM J. Numer. Anal.* **42** 968–973.

Mathé, P. and Pereverzev, S. (2002). Moduli of continuity for operator valued functions. *Numer. Func. Anal. Opt.* **23** 623–631.

Minsker, S. (2011). On Some Extensions of Bernstein's Inequality for Self-adjoint Operators. *ArXiv e-prints* .

Nemirovskii, A. (1985). Nonparametric estimation of smooth regression functions. *Sov. J. Comput. Syst. S+* **23** 1–11.

Nemirovskii, A., Polyak, B. and Tsybakov, A. (1983). Estimators of maximum likelihood type for nonparametric regression. *Dokl. Math.* **28** 788–792.

Neubauer, A. (1997). On converse and saturation results for Tikhonov regularization of linear ill-posed problems. *SIAM J. Numer. Anal.* **34** 517–527.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*.

Rosasco, L., De Vito, E. and Verri, A. (2005). Spectral methods for regularization in learning theory. Tech. Rep. DISI-TR-05-18, Universitá degli Studi di Genova, Italy.

Steinwart, I., Hush, D. and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Conference on Learning Theory*.

Tropp, J. (2015). An Introduction to Matrix Concentration Inequalities. *ArXiv e-prints* .

Tsybakov, A. (2004). *Introduction a l'estimation non-paramétrique*. Springer.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

Williams, C. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*.

Yang, Y., Pilanci, M. and Wainwright, M. (2015). Randomized sketches for kernels: Fast and optimal non-parametric regression. *ArXiv e-prints* .

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17** 2077–2098.

Zhang, Y., Duchi, J. and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*.