Journal of Machine Learning Research 25 (2024) 1-51 Submitted 12/23; Revised 5/24; Published 6/24

Towards Optimal Sobolev Norm Rates for the Vector-Valued Regularized Least-Squares Algorithm

Zhu Li* ZHU.LI@UCL.AC.UK
Gatsby Computational Neuroscience Unit
University College London
London, W1T 4JG, UK

Dimitri Meunier* DIMITRI.MEUNIER.21@UCL.AC.UK
Gatsby Computational Neuroscience Unit
University College London
London, W1T 4JG, UK

Mattes Mollenhauer MATTES.MOLLENHAUER@MERANTIX-MOMENTUM.COM
Merantix Momentum
Merantix AI Campus
Max–Urich–Straße 3, 13355 Berlin, Germany

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit
University College London
London, W1T 4JG, UK

Editor: Daniel Hsu

Abstract

We present the first optimal rates for infinite-dimensional vector-valued ridge regression on a continuous scale of norms that interpolate between L_2 and the hypothesis space, which we consider as a vector-valued reproducing kernel Hilbert space. These rates allow to treat the misspecified case in which the true regression function is not contained in the hypothesis space. We combine standard assumptions on the capacity of the hypothesis space with a novel tensor product construction of vector-valued interpolation spaces in order to characterize the smoothness of the regression function. Our upper bound not only attains the same rate as real-valued kernel ridge regression, but also removes the assumption that the target regression function is bounded. For the lower bound, we reduce the problem to the scalar setting using a projection argument. We show that these rates are optimal in most cases and independent of the dimension of the output space. We illustrate our results for the special case of vector-valued Sobolev spaces.

Keywords: Statistical learning, regularized least squares, optimal rates, interpolation norms.

1. Introduction

Optimal learning rates for least-squares regression with scalar outputs have been studied extensively in the context of reproducing kernel Hilbert spaces (RKHS) over the last two

*Equal Contribution

decades. While some analyses considered vector-valued outputs, the optimality of vector-valued kernel-based algorithms remained an open question in important settings which include model misspecification or infinite-dimensional response variables. We consider a data set $D = \{(x_i, y_i)\}_{i=1}^n$ of observations independently sampled from a joint unknown distribution P on $E_X \times \mathcal{Y}$, where $\mathcal{Y}$ is a potentially infinite-dimensional output space and E_X is the covariate space. Let (X, Y) be a random variable taking values in $E_X \times \mathcal{Y}$ distributed according to P. The objective is to estimate the *regression function* or *conditional mean function* $F_* : E_X \to \mathcal{Y}$ given by $F_*(x) := \mathbb{E}[Y \mid X = x]$. Our focus in this work is to approximate F_* with kernel-based regularized least-squares algorithms, where we pay special attention to the case when $\mathcal{Y}$ is of high or infinite dimension—a setting including important applications in multitask learning, functional data analysis and inference with kernel mean embeddings. We consider the estimate $\hat{F}_\lambda : E_X \to \mathcal{Y}$, obtained by solving the convex optimization problem

$$\hat{F}_\lambda = \operatorname*{argmin}_{F \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{G}}^2 \right\}, \tag{1}$$

where a vector-valued reproducing kernel Hilbert space (vRKHS) $\mathcal{G}$ over E_X serves as the hypothesis space and $\lambda > 0$ is a regularization parameter. A vRKHS, as detailed in Section 2, is a generalization of the standard RKHS, allowing us to model functions that take values in a Hilbert space. This algorithm is commonly referred to as (vector-valued) *ridge regression* or simply the *regularized least squares* (RLS) algorithm—even though it can be understood as a special instance of *Tikhonov regularization* in the more general context of regularization theory. A central theoretical challenge in this context is to establish learning rates, either in expectation or in probability with respect to the distribution of D, for the error

$$\|\hat{F}_\lambda - F_*\| \tag{2}$$

in some relevant norm. In this paper, we investigate the behavior of Eq. (2) with respect to the norms of a continuum of suitable Hilbert spaces $[\mathcal{G}]^\gamma$ with $\mathcal{G} \subseteq [\mathcal{G}]^\gamma \subseteq L_2$; see Definition 2 for an exact definition. We focus on the class of vector-valued RKHSs induced by the vector-valued kernel

$$K(x, x') := k_X(x, x')T, \tag{3}$$

where $T : \mathcal{Y} \to \mathcal{Y}$ is a bounded positive-semidefinite self-adjoint operator and $k_X : E_X \times E_X \to \mathbb{R}$ is a scalar-valued kernel. This choice of kernel is the de-facto standard for infinite-dimensional learning problems, as it allows to practically compute $\hat{F}_\lambda$ defined by Eq. (1) conveniently in a variety of practical applications.

Relevant applications. Notable examples for such an infinite-dimensional learning setting are the estimation of dynamical systems (Song et al., 2009; Kostic et al., 2022, 2023), functional response regression (Kadri et al., 2016), structured prediction (Ciliberto et al., 2016, 2020), the estimation of linear operators (Mollenhauer and Koltai, 2020; Mollenhauer et al., 2022), the conditional mean embedding (Grünewälder et al., 2012a,b; Park and Muandet, 2020), causal effect estimation under observed covariates (Singh et al., 2023) and kernel regression with instrumental (Singh et al., 2019) and proximal (Mastouri et al., 2021) variables. We emphasize that in all of the above applications, vector-valued kernels of the form (3) are used, the identity operator $T = \mathrm{Id}_{\mathcal{Y}}$ being the most popular choice. An important reason for this choice of kernel is that, even in the infinite-dimensional case, it allows a convenient numerical evaluation of $\hat{F}_\lambda$ in terms of a vector-valued representer theorem (see e.g. Grünewälder et al., 2012a and Kadri et al., 2016).

Related work. The RLS algorithm has been extensively studied in literature (see e.g., Caponnetto and De Vito, 2007; Smale and Zhou, 2007; Steinwart et al., 2009; Blanchard and Mücke, 2018; Dicker et al., 2017; Lin et al., 2020; Fischer and Steinwart, 2020, and references therein). However, existing results concerning the optimal rates for RLS often cover the real-valued output space only. One exception is the work by Caponnetto and De Vito (2007), where the output space $\mathcal{Y}$ is potentially infinite-dimensional. Their analysis does not generally hold for the kernel K defined by Eq. (3), however, in the setting that $\mathcal{Y}$ is infinite-dimensional, as it relies on a trace condition which is violated when T is not trace class—a restriction which rules out the choice $T = \mathrm{Id}_{\mathcal{Y}}$, used for example in the analysis of conditional mean embedding (Grünewälder et al., 2012b; Li et al., 2022b). This issue has been noted by multiple authors in the context of specific applications (Grünewälder et al., 2012a,b; Kadri et al., 2016; Mollenhauer and Koltai, 2020; Park and Muandet, 2020). In addition, Caponnetto and De Vito (2007) assume that $\mathcal{Y}$ is finite dimensional in order to obtain the matching lower bound and only consider the well-specified case. We provide a comparison of our results with existing results in Table 1. For a detailed discussion, please see Section 7.

Contributions of this work. This manuscript extends the results from the earlier work of Li et al. (2022b) in multiple ways:

- ***Vector-valued interpolation spaces.*** While Li et al. (2022b) consider a specific instance of infinite-dimensional RLS in terms of the conditional mean embedding, by assuming that the response variable Y takes values in a RKHS, we offer an updated analysis of the RLS algorithm which applies to more general infinite-dimensional spaces $\mathcal{Y}$. Our study covers both the hard learning scenario (misspecified setting) when $F_* \notin \mathcal{G}$ and the easy learning scenario (well-specified setting) when $F_* \in \mathcal{G}$ (see Theorem 3). In order to cover the misspecified case, we construct novel interpolation spaces of vector-valued functions in terms of an isomorphism which allows to represent functions and linear operators in terms of a tensor product (Mollenhauer and Koltai, 2020; Mollenhauer et al., 2022). In both cases, when $\mathcal{Y}$ is real-valued, we recover the same rate as in Fischer and Steinwart (2020), the current best known rate for real-valued kernel ridge regression in the literature. We provide a thorough comparison to previous works after stating our results.

- ***Proof technique.*** Building upon our previous work, our definition of the vector-valued interpolation spaces allows to modify the integral operator technique while avoiding the aforementioned trace condition for T required by Caponnetto and De Vito (2007). We bypass the trace condition by making use of tensor product arguments. This allows to reduce the infinite-dimensional learning scenario to known real-valued arguments in multiple instances in our proofs.

- ***Finite dimensions.*** The results of this work naturally cover dimension-free rates for misspecified multitask learning in finite-dimensional spaces. To the best of our knowledge, such a setting has not been investigated before in the literature.

- ***Lower rates.*** With the so-called reduction technique, we obtain lower rates for the the general vector-valued learning setting even when $\mathcal{Y}$ is infinite-dimensional (see Theorem 5). These rates match our upper rates in many cases.

- ***Unbounded regression function.*** The available analysis of scalar-valued kernel ridge regression requires boundedness of the regression function ($\sup_{x\in E_X}|F_*(x)| < \infty$), see Fischer and Steinwart (2020). Recently, Zhang et al. (2023b) proved that same rates can be obtained by replacing the boundedness condition with the weaker assumption that $|F_*|^q$ is integrable for some $q \geq 2$. Later, Zhang et al. (2023a) demonstrated that this integrability assumption is automatically satisfied by scalar valued interpolation spaces (defined in Section 2). This so-called L_q*-embedding property* allows to completely remove the assumption that the target function is bounded or that its higher moments are integrable. Generalizing the ideas of Zhang et al. (2023b,a), we derive learning rates in the vector-valued setting without requiring boundedness or integration of the higher moments of F_* (see Theorem 3 and Remark 8). Key to this generalisation is Theorem 4, which states that the L_q–embedding property also holds for vector-valued interpolation spaces.

- ***Vector-valued Sobolev spaces.*** As a final contribution, we study RLS learning in the setting of vector-valued Sobolev spaces (see Definition 3) which was not covered in Li et al. (2022b). We obtain, for the first time, the minimax optimal learning rate for RLS learning in vector-valued Sobolev spaces (see Corollary 2 and 3). This contribution shows that our definition of the vector-valued interpolation spaces admits a natural interpretation in practical applications, rather than just being "the appropriate technical tool" to prove rates for misspecified vector-valued learning problems.

Organisation of this paper. In Section 2, we introduce the concept of a vRKHS and the required mathematical preliminaries. Section 3 discusses the formal construction of vector-valued interpolation spaces, which will be central to our analysis. We provide upper rates for the vector-valued learning problem in Section 4, while a corresponding lower bound on the rates is presented in Section 5. Section 6 sets our results in line with known rates for the scalar learning setting in the context of Sobolev spaces, and Section 7 compares our result with other contributions. In order to improve the readability, we defer proofs and technical auxiliary results to the appendices.

Readers familiar with the commonly used integral operator framework for kernel ridge regression (e.g. Caponnetto and De Vito, 2007) who are mostly interested in the structure of our upper and lower rates may directly refer to Section 4 and Section 5. The precise mathematical definition of the vector-valued interpolation space in Section 3 may be consulted afterwards—technically, it is used as a direct replacement for other source conditions found in the literature. The aforementioned three sections contain the fundamental additions to the known framework, while the technical setup and mathematical background from Section 2 are standard. For convenience, we provide a summary of our most important notation in Table 2 in Section 3.

	Kernel	Output space	Misspecified case	smoothness	Algorithm	Norm
Blanchard and Mücke (2018)	scalar $k_X(x,x')$	Real-valued	no	Hölder source condition	general	γ-norm
Fischer and Steinwart (2020)	scalar $k_X(x,x')$	Real-valued	yes	interpolation space	ridge regression	γ-norm
Caponnetto and De Vito (2007)	$K(x,x')$ trace class for all $x,x' \in E_X$	Vector-valued	no	Hölder source condition	ridge regression	L_2-norm
This work	$k_X(x,x')T$ with T psd.	Vector-valued	yes	vector-valued interpolation space	ridge regression	γ-norm

Table 1: An overview of articles providing lower rates and corresponding optimal upper rates for kernel-based least squares algorithms based on the integral operator technique in various scenarios under comparable assumptions on the underlying distributions. Note that the γ-norm is defined in Eq. (11).

2. Mathematical Preliminaries

Throughout the paper, we consider a random variable X (the covariate) defined on a second countable locally compact Hausdorff space[1] E_X endowed with its Borel σ-field $\mathcal{F}_{E_X}$, and the random variable Y (the output) defined on a potentially infinite dimensional separable real Hilbert space $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}})$ endowed with its Borel σ-field $\mathcal{F}_{\mathcal{Y}}$. We let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space with expectation operator $\mathbb{E}$. Let P be the pushforward of $\mathbb{P}$ under (X,Y) and π and ν be the marginal distributions on E_X and $\mathcal{Y}$, respectively; i.e., $X \sim \pi$ and $Y \sim \nu$. We use the Markov kernel $p : E_X \times \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$ to express the distribution of Y conditioned on X as

$$\mathbb{P}[Y \in A | X = x] = \int_A p(x, dy),$$

for all $x \in E_X$ and events $A \in \mathcal{F}_{\mathcal{Y}}$, see e.g. Dudley (2002).

We now introduce some notation related to linear operators and vector-valued integration. For more details, the reader may consult Weidmann (1980) and Diestel and Uhl (1977), respectively. We denote the space of real-valued Lebesgue square integrable functions on

[1] Under additional technical assumptions, the results in this paper can also be formulated when E_X is a more general topological space (for example when E_X is Polish). However, some properties of kernels defined on E_X such as the so-called *c_0-universality* (Carmeli et al., 2010) simplify the exposition when E_X is a second countable locally compact Hausdorff space; see Remark 3.

$(E_X, \mathcal{F}_{E_X})$ with respect to π as $L_2(E_X, \mathcal{F}_{E_X}, \pi)$ abbreviated $L_2(\pi)$ and similarly for ν we use $L_2(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}}, \nu)$ abbreviated $L_2(\nu)$. Let H be a separable real Hilbert space with inner product $\langle\cdot,\cdot\rangle_H$. We write $\mathcal{L}(H, H')$ as the Banach space of bounded linear operators from H to another Hilbert space H', equipped with the operator norm $\|\cdot\|_{H\to H'}$. When $H = H'$, we simply write $\mathcal{L}(H)$ instead. We also let $L_p(E_X, \mathcal{F}_{E_X}, \pi; H)$, abbreviated $L_p(\pi; H)$, the space of strongly $\mathcal{F}_{E_X} - \mathcal{F}_H$ measurable and Bochner p-integrable functions from E_X to H for $1 \le p \le \infty$ with the norms

$$\|f\|^p_{L_p(\pi;H)} = \int_{E_X} \|f\|^p_H \,\mathrm{d}\pi, \quad 1 \le p < \infty, \qquad \|f\|_{L_\infty(\pi;H)} = \inf\left\{C \ge 0 : \pi\{\|f\|_H > C\} = 0\right\}. \tag{4}$$

We denote the p-Schatten class $S_p(H, H')$ to be the space of all compact operators C from H to H' such that $\|C\|_{S_p(H,H')} := \|(\sigma_i(C))_{i\in J}\|_{\ell_p}$ is finite. Here $\|\,(\sigma_i(C))_{i\in J}\,\|_{\ell_p}$ is the ℓ_p sequence space norm of the sequence of the strictly positive singular values of C indexed by the countable set J. For $p = 2$, $S_2(H, H')$ is the Hilbert space of Hilbert-Schmidt operators from H to H'. Finally, for two Hilbert spaces H, H', we say that H is (continuously) embedded in H' and denote it as $H \hookrightarrow H'$ if H can be interpreted as a vector subspace of H' and the inclusion operator $i : H \to H'$ performing the change of norms with $ix = x$ for $x \in H$ is continuous; and we say that H is isometrically isomorphic to H' and denote it as $H \simeq H'$ if there is a linear isomorphism $\Psi : H \to H'$ which is an isometry. We write $H \cong H'$, if the sets coincide with equivalent norms.

Tensor Product of Hilbert Spaces (Aubin, 2000, Section 12)**:** Denote $H \otimes H'$ the tensor product of Hilbert spaces H, H'. The Hilbert space $H \otimes H'$ is the completion of the algebraic tensor product with respect to the norm induced by the inner product $\langle x_1 \otimes x'_1, x_2 \otimes x'_2\rangle_{H\otimes H'} = \langle x_1, x_2\rangle_H \langle x'_1, x'_2\rangle_{H'}$ for $x_1, x_2 \in H$ and $x'_1, x'_2 \in H'$ defined on the elementary tensors of $H\otimes H'$. This definition extends to $\operatorname{span}\{x\otimes x' | x \in H, x' \in H'\}$ and finally to its completion. The space $H \otimes H'$ is separable whenever both H and H' are separable. The element $x \otimes x' \in H \otimes H'$ is treated as the linear rank-one operator $x \otimes x' : H' \to H$ defined by $y' \to \langle y', x'\rangle_{H'} x$ for $y' \in H'$. Based on this identification, the tensor product space $H \otimes H'$ is isometrically isomorphic to the space of Hilbert-Schmidt operators from H' to H, i.e., $H \otimes H' \simeq S_2(H', H)$. We will hereafter not make the distinction between those two spaces and see them as identical. If $\{e_i\}_{i\in I}$ and $\{e'_j\}_{j\in J}$ are orthonormal basis in H and H', $\{e_i \otimes e'_j\}_{i\in I, j\in J}$ is an orthonormal basis in $H \otimes H'$.

Remark 1 (Aubin, 2000, Theorem 12.6.1) *Consider the Bochner space $L_2(\pi; H)$ where H is a separable Hilbert space. One can show that $L_2(\pi; H)$ is isometrically identified with the tensor product space $H \otimes L_2(\pi)$.*

Reproducing Kernel Hilbert Spaces, Covariance Operators: We let $k_X : E_X \times E_X \to \mathbb{R}$ be a symmetric and positive definite kernel function and $\mathcal{H}_X$ be a vector space of functions from E_X to $\mathbb{R}$, endowed with a Hilbert space structure via an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_X}$. We say k_X is a reproducing kernel of $\mathcal{H}_X$ if and only if for all $x \in E_X$ we have $k_X(\cdot, x) \in \mathcal{H}_X$ and for all $x \in E_X$ and $f \in \mathcal{H}_X$, we have $f(x) = \langle f, k_X(x,\cdot)\rangle_{\mathcal{H}_X}$. A space $\mathcal{H}_X$ which possesses a reproducing kernel is called a reproducing kernel Hilbert space (RKHS; Berlinet and Thomas-Agnan, 2011). We denote the canonical feature map of $\mathcal{H}_X$ as $\phi_X(x) = k_X(\cdot, x)$.

We require some technical assumptions on the previously defined RKHS and kernel:

1. $\mathcal{H}_X$ is separable, this is satisfied if k_X is continuous, given that E_X is separable[2];
2. $k_X(\cdot, x)$ is measurable for π-almost all $x \in E_X$;
3. $k_X(x,x) \leq \kappa_X^2$ for π-almost all $x \in E_X$.

Note that the above assumptions are not restrictive in practice, as well-known kernels such as the Gaussian, Laplacian and Matérn kernels satisfy all of the above assumptions on $E_X \subseteq \mathbb{R}^d$ (Sriperumbudur et al., 2011). We now introduce some facts about the interplay between $\mathcal{H}_X$ and $L_2(\pi)$, which has been extensively studied by Smale and Zhou (2004, 2005), De Vito et al. (2006) and Steinwart and Scovel (2012). We first define the (not necessarily injective) embedding $I_\pi : \mathcal{H}_X \to L_2(\pi)$, mapping a function $f \in \mathcal{H}_X$ to its π-equivalence class $[f]$. The embedding is a well-defined compact operator as long as its Hilbert-Schmidt norm is finite. In fact, this requirement is satisfied since its Hilbert-Schmidt norm can be computed as (Steinwart and Scovel, 2012, Lemma 2.2 & 2.3)

$$\|I_\pi\|_{S_2(\mathcal{H}_X, L_2(\pi))} = \|k_X\|_{L_2(\pi)} := \left(\int_{E_X} k_X(x,x) \mathrm{d}\pi(x)\right)^{1/2} < \infty.$$

The adjoint operator $S_\pi := I_\pi^* : L_2(\pi) \to \mathcal{H}_X$ is an integral operator with respect to the kernel k_X, i.e. for $f \in L_2(\pi)$ and $x \in E_X$ we have (Steinwart and Christmann, 2008, Theorem 4.27)

$$\left(S_\pi f\right)(x) = \int_{E_X} k_X\left(x, x'\right) f\left(x'\right) \mathrm{d}\pi\left(x'\right).$$

Next, we define the self-adjoint and positive semi-definite integral operators

$$L_X := I_\pi S_\pi : L_2(\pi) \to L_2(\pi) \quad \text{and} \quad C_{XX} := S_\pi I_\pi : \mathcal{H}_X \to \mathcal{H}_X.$$

These operators are trace class and their trace norms satisfy

$$\|L_X\|_{S_1(L_2(\pi))} = \|C_{XX}\|_{S_1(\mathcal{H}_X)} = \|I_\pi\|^2_{S_2(\mathcal{H}_X, L_2(\pi))} = \|S_\pi\|^2_{S_2(L_2(\pi), \mathcal{H}_X)}.$$

Vector-valued RKHS: We give a brief overview of the vector-valued reproducing kernel Hilbert space. We refer the reader to Carmeli et al. (2006) and Carmeli et al. (2010) for more details.

Definition 1 *Let $K : E_X \times E_X \to \mathcal{L}(\mathcal{Y})$ be an operator valued positive-semidefinite (psd) kernel such that $K(x,x') = K(x',x)^*$ for all $x, x' \in E_X$, and for all $x_1, \ldots, x_n \in E_X$ and $h_i, h_j \in \mathcal{Y}$,*

$$\sum_{i,j=1}^n \langle h_i, K(x_i, x_j) h_j \rangle_{\mathcal{Y}} \geq 0.$$

Fix K, $x \in E_X$, and $h \in \mathcal{Y}$, then $[K_x h](\cdot) := K(\cdot, x)h$ defines a function from E_X to $\mathcal{Y}$. We now consider

$$\mathcal{G}_{\text{pre}} := \operatorname{span}\{K_x h \mid x \in E_X, h \in \mathcal{Y}\}$$

[2] This follows from Steinwart and Christmann (2008, Lemma 4.33). Note that the lemma requires separability of E_X, which is satisfied since we assume that E_X is a second countable locally compact Hausdorff space.

with inner product on $\mathcal{G}_{\text{pre}}$ by linearly extending the expression

$$\left\langle K_x h, K_{x'} h' \right\rangle_{\mathcal{G}} := \left\langle h, K\left(x, x'\right) h' \right\rangle_{\mathcal{Y}}. \tag{5}$$

Let $\mathcal{G}$ be the completion of $\mathcal{G}_{\text{pre}}$ with respect to this inner product. We call $\mathcal{G}$ the vRKHS induced by the kernel K. The space $\mathcal{G}$ is a Hilbert space consisting of functions from E_X to $\mathcal{Y}$ with the reproducing property

$$\langle F(x), h \rangle_{\mathcal{Y}} = \langle F, K_x h \rangle_{\mathcal{G}}, \tag{6}$$

for all $F \in \mathcal{G}, h \in \mathcal{Y}$ and $x \in E_X$. For all $F \in \mathcal{G}$ we obtain

$$\|F(x)\|_{\mathcal{Y}} \leq \|K(x,x)\|_{\mathcal{Y}\to\mathcal{Y}}^{1/2} \|F\|_{\mathcal{G}}, \quad x \in E_X.$$

The inner product given by Eq. (5) implies that K_x is a bounded operator for all $x \in E_X$. For all $F \in \mathcal{G}$ and $x \in E_X$, Eq. (6) can be written as $F(x) = K_x^* F$. The linear operators $K_x : \mathcal{Y} \to \mathcal{G}$ and $K_x^* : \mathcal{G} \to \mathcal{Y}$ are bounded with

$$\|K_x\|_{\mathcal{Y}\to\mathcal{G}} = \|K_x^*\|_{\mathcal{G}\to\mathcal{Y}} = \|K(x,x)\|_{\mathcal{Y}\to\mathcal{Y}}^{1/2}$$

and we have $K_x^* K_{x'} = K\left(x, x'\right), x, x' \in E_X$. In the following, we will denote $\mathcal{G}$ as the vRKHS induced by the kernel $K : E_X \times E_X \to \mathcal{L}(\mathcal{Y})$ with

$$K(x,x') := k_X(x,x') \operatorname{Id}_{\mathcal{Y}}, \quad x, x' \in E_X.$$

Remark 2 (General multiplicative kernel) *Without loss of generality, we provide our results for the vRKHS $\mathcal{G}$ induced by the operator-valued kernel given by $K(x,x') = k_X(x,x') \operatorname{Id}_{\mathcal{Y}}$. However, with suitably adjusted constants in the assumptions, our results transfer directly to the more general vRKHS $\widetilde{\mathcal{G}}$ induced by the more general operator-valued kernel*

$$\widetilde{K}(x,x') := k_X(x,x') T$$

where $T : \mathcal{Y} \to \mathcal{Y}$ is any bounded positive-semidefinite self-adjoint operator. In fact, this setting is covered by straightforwardly replacing the response variable Y with the modified response $\widetilde{Y} := T^{1/2} Y$ in our learning problem. Equivalently, the space $\widetilde{\mathcal{G}}$ is obtained as the vRKHS induced by the kernel $K(x,x') = k_X(x,x') \operatorname{Id}_{\mathcal{Y}}$ by introducing $\langle y, y' \rangle_{\tilde{\mathcal{Y}}} := \langle y, Ty' \rangle_{\mathcal{Y}}$ for all $y, y' \in \mathcal{Y}$, which defines an inner product on the quotient space[3] *$\tilde{\mathcal{Y}} := \mathcal{Y} / \ker(T)$. This can readily be seen by the construction of $\widetilde{\mathcal{G}}$. By (5), we have*

$$\left\langle \widetilde{K}_x y, \widetilde{K}_{x'} y' \right\rangle_{\widetilde{\mathcal{G}}} = \left\langle y, k_X\left(x, x'\right) T y' \right\rangle_{\mathcal{Y}} = \left\langle y, k_X\left(x, x'\right) \operatorname{Id}_{\mathcal{Y}} y' \right\rangle_{\tilde{\mathcal{Y}}} \tag{7}$$

for all $x, x' \in E_X$ and $y, y' \in \mathcal{Y}$. In fact, we have $\tilde{\mathcal{G}} \simeq \mathcal{H} \otimes \tilde{\mathcal{Y}}$, see Carmeli et al. (2010, Example 3.2). In Section 4.1, we give the adjusted constants appearing in our learning rates when K is replaced with $\tilde{K}$.

An important property of $\mathcal{G}$ is that elements in $\mathcal{G}$ are isometrically isomorphic to the space of Hilbert-Schmidt operators between $\mathcal{H}_X$ and $\mathcal{Y}$.

[3] When T is not stricly positive definite, then elements in the nullspace $\ker(T)$ are simply interpreted as the element 0 in $\tilde{\mathcal{Y}}$, ensuring that $\langle \cdot, \cdot \rangle_{\tilde{\mathcal{Y}}}$ is a well-defined inner product.

Theorem 1 (Example 5 (i) in Carmeli et al., 2010) *For $g \in \mathcal{Y}$ and $f \in \mathcal{H}_X$, define the map $\bar{\Psi}$ on the elementary tensors as*

$$\left[\bar{\Psi}\left(g \otimes f\right)\right](x) \coloneqq f(x)g = \left(g \otimes f\right)\phi_X(x).$$

Then $\bar{\Psi}$ defines an isometric isomorphism between $S_2(\mathcal{H}_X, \mathcal{Y})$ and $\mathcal{G}$ through linearity and completion.

More details regarding Theorem 1 (for the special case where $\mathcal{Y}$ is a RKHS) can be found in Mollenhauer and Koltai (2020, Theorem 4.4). The isometric isomorphism $\bar{\Psi}$ induces the operator reproducing property stated below.

Corollary 1 *For every function $F \in \mathcal{G}$ there exists an operator $C \coloneqq \bar{\Psi}^{-1}(F) \in S_2(\mathcal{H}_X, \mathcal{Y})$ such that*

$$F(x) = C\phi_X(x) \in \mathcal{Y},$$

for all $x \in E_X$ with $\|C\|_{S_2(\mathcal{H}_X,\mathcal{Y})} = \|F\|_{\mathcal{G}}$ and vice versa. Conversely, for any pair $F \in \mathcal{G}$ and $C \in S_2(\mathcal{H}_X, \mathcal{Y})$, we have $C = \bar{\Psi}^{-1}(F)$ as long as $F(x) = C\phi_X(x)$ for all $x \in E_X$.

The proof of Corollary 1 is a simple extension of Lemma 15 in Ciliberto et al. (2016) and Corollary 4.5 in Mollenhauer and Koltai (2020). Corollary 1 shows that the vRKHS $\mathcal{G}$ is generated via the space of Hilbert-Schmidt operators $S_2(\mathcal{H}_X, \mathcal{Y})$

$$\mathcal{G} = \left\{F : E_X \to \mathcal{Y} \mid F = C\phi_X(\cdot), \quad C \in S_2(\mathcal{H}_X, \mathcal{Y})\right\}.$$

Vector-valued regression: We briefly recall the basic setup of regularized least squares regression with Hilbert space-valued random variables. The risk for vector-valued regression is

$$\mathcal{E}(F) \coloneqq \mathbb{E}\left[\|Y - F(X)\|_{\mathcal{Y}}^2\right] = \int_{E_X \times \mathcal{Y}} \|y - F(x)\|_{\mathcal{Y}}^2 p(x, dy)\pi(dx),$$

for measurable functions $F : E_X \to \mathcal{Y}$. The analytical minimiser of the risk over all those measurable functions is the *regression function* or the *conditional mean function* $F_\star \in L_2(\pi; \mathcal{Y})$ given by

$$F_*(x) \coloneqq \mathbb{E}[Y \mid X = x] = \int_{\mathcal{Y}} y\, p(x, dy), \quad x \in E_X.$$

This fact can for example be proven via a classical decomposition of the risk, see e.g. Mollenhauer and Koltai (2020, Theorem A.1). Throughout the paper, we assume that $\mathbb{E}[\|Y\|_{\mathcal{Y}}^2] < +\infty$, i.e., the random variable Y is square integrable. Note that this ensures that we have $F_\star \in L_2(\pi; \mathcal{Y})$.

We pick $\mathcal{G}$ as an hypothesis space of functions to estimate F_*. Given a data set $D = \{(x_i, y_i)\}_{i=1}^n$ independently and identically sampled from the joint distribution of X and Y, a regularized estimate of F_* is the solution of the following optimization problem:

$$\hat{F}_\lambda \coloneqq \arg\min_{F \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{G}}^2,$$

where $\lambda > 0$ is the regularization parameter. According to Corollary 1, $\hat{F}_\lambda(\cdot) := \bar{\Psi}\left(\hat{C}_\lambda\right)(\cdot) = \hat{C}_\lambda \phi_X(\cdot)$ where

$$\hat{C}_\lambda := \underset{C \in S_2(\mathcal{H}_X, \mathcal{Y})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|y_i - C\phi_X(x_i)\|_{\mathcal{Y}}^2 + \lambda \|C\|_{S_2(\mathcal{H}_X, \mathcal{Y})}^2, \tag{8}$$

An explicit solution is given by

$$\hat{F}_\lambda(x) = \hat{C}_\lambda \phi_X(x) = \sum_{i=1}^{n} y_i \beta_i(x), \qquad \beta(x) := \left[\mathbf{K}_{XX} + n\lambda \operatorname{Id}\right]^{-1} \mathbf{k}_{Xx} \in \mathbb{R}^n$$

$$\begin{aligned} (\mathbf{K}_{XX})_{ij} &= k(x_i, x_j) \qquad && i, j \in [n] \\ (\mathbf{k}_{Xx})_i &= k(x_i, x) \qquad && i \in [n] \end{aligned}$$

The above model is well-specified if the equivalence class $F_\star \in L_2(\pi; \mathcal{Y})$ admits a representative which is contained in $\mathcal{G}$. In what follows, we will simply write this scenario as $F_\star \in \mathcal{G}$ by abuse of notation. We note that universal consistency of this approach at least requires that $\mathcal{G}$ is dense in $L_2(\pi; \mathcal{Y})$ such that for every possible $F_\star$, we can achieve $\hat{F}_\lambda \to F_\star$ in the norm of $L_2(\pi; \mathcal{Y})$ either in expectation or with high probability with respect to the distribution of the samples D for some admissible regularization scheme $\lambda = \lambda_n \to 0$ whenever $n \to \infty$. This denseness is well-investigated and generally satisfied; see Remark 5.

Real-valued Interpolation Space: We now introduce the background required in order to characterize the Hilbert spaces used to deal with the misspecified setting $F_* \notin \mathcal{G}$. We review the results of Steinwart and Scovel (2012) and Fischer and Steinwart (2020) that set out the eigendecompositions of L_X and C_{XX}, and apply these in constructing the interpolation spaces used for the misspecified setting. By the spectral theorem for self-adjoint compact operators, there exists an at most countable index set I, a non-increasing sequence $(\mu_i)_{i\in I} > 0$, and a family $(e_i)_{i\in I} \in \mathcal{H}_X$, such that $([e_i])_{i\in I}$ is an orthonormal basis (ONB) of $\overline{\operatorname{ran} I_\pi} \subseteq L_2(\pi)$ and $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONB of $(\ker I_\pi)^\perp \subseteq \mathcal{H}_X$, and we have

$$L_X = \sum_{i\in I} \mu_i \langle \cdot, [e_i] \rangle_{L_2(\pi)} [e_i], \qquad C_{XX} = \sum_{i\in I} \mu_i \langle \cdot, \mu_i^{\frac{1}{2}} e_i \rangle_{\mathcal{H}_X} \mu_i^{\frac{1}{2}} e_i \tag{9}$$

For $\alpha \geq 0$, we define the α-interpolation space Steinwart and Scovel (2012) by

$$[\mathcal{H}]_X^\alpha := \left\{ \sum_{i\in I} a_i \mu_i^{\alpha/2} [e_i] : (a_i)_{i\in I} \in \ell_2(I) \right\} \subseteq L_2(\pi), \tag{10}$$

equipped with the α-power norm

$$\left\| \sum_{i\in I} a_i \mu_i^{\alpha/2} [e_i] \right\|_{[\mathcal{H}]_X^\alpha} := \|(a_i)_{i\in I}\|_{\ell_2(I)} = \left(\sum_{i\in I} a_i^2 \right)^{1/2}. \tag{11}$$

For $(a_i)_{i\in I} \in \ell_2(I)$, the α-interpolation space becomes a Hilbert space with inner product defined as

$$\left\langle \sum_{i\in I} a_i (\mu_i^{\alpha/2} [e_i]), \sum_{i\in I} b_i (\mu_i^{\alpha/2} [e_i]) \right\rangle_{[\mathcal{H}]_X^\alpha} = \sum_{i\in I} a_i b_i.$$

$$\begin{tikzcd}
S_2(L_2(\pi),\mathcal{Y}) \arrow[r, "\Psi"'] & L_2(\pi;\mathcal{Y}) \\
S_2(\mathcal{H}_X,\mathcal{Y}) \arrow[u, dotted, "\mathcal{I}_\pi"] \arrow[r, "\bar{\Psi}"'] & \mathcal{G}
\end{tikzcd}$$

Figure 1: Ψ and $\bar{\Psi}$ are the isometric isomorphisms between each pair of spaces. $\mathcal{I}_\pi$ denotes the canonical embedding between the two Hilbert-Schmidt spaces.

Moreover, $\left(\mu_i^{\alpha/2}[e_i]\right)_{i\in I}$ forms an ONB of $[\mathcal{H}]_X^\alpha$ and consequently $[\mathcal{H}]_X^\alpha$ is a separable Hilbert space. In the following, we use the abbreviation $\|\cdot\|_\alpha := \|\cdot\|_{[\mathcal{H}]_X^\alpha}$. For $\alpha = 0$ we have $[\mathcal{H}]_X^0 = \overline{\operatorname{ran} I_\pi} \subseteq L_2(\pi)$ with $\|\cdot\|_0 = \|\cdot\|_{L_2(\pi)}$. Moreover, for $\alpha = 1$ we have $[\mathcal{H}]_X^1 = \operatorname{ran} I_\pi$ and $[\mathcal{H}]_X^1$ is isometrically isomorphic to the closed subspace $(\ker I_\pi)^\perp$ of $\mathcal{H}_X$ via I_π, i.e. $\|[f]\|_1 = \|f\|_{\mathcal{H}_X}$ for $f \in (\ker I_\pi)^\perp$. For $0 < \beta < \alpha$, we have

$$[\mathcal{H}]_X^\alpha \hookrightarrow [\mathcal{H}]_X^\beta \hookrightarrow [\mathcal{H}]_X^0 \subseteq L_2(\pi). \tag{12}$$

For $\alpha > 0$, the α-interpolation space is given by the image of the fractional integral operator, namely

$$[\mathcal{H}]_X^\alpha = \operatorname{ran} L_X^{\alpha/2} \quad \text{and} \quad \left\| L_X^{\alpha/2} f \right\|_\alpha = \|f\|_{L_2(\pi)}$$

for $f \in \overline{\operatorname{ran} I_\pi}$.

Remark 3 (Universality) *Under assumptions 1 to 3 and E_X being a second-countable locally compact Hausdorff space, if $k_X(\cdot, x)$ is continuous and vanishing at infinity, then $[\mathcal{H}]_X^0 = L_2(\pi)$ if and only if $\mathcal{H}_X$ is dense in the space of continuous functions vanishing at infinity equipped with the uniform norm (Carmeli et al., 2010). Such RKHS are called c_0-universal. As a special case of Carmeli et al. (2010, Proposition 5.6), one can show that on $\mathbb{R}^d$, Gaussian, Laplacian, inverse multiquadrics and Matérn kernels are c_0-universal.*

Remark 4 (Interpolation space) *The name $\alpha-$interpolation space comes from the fact that for $0 < \alpha < 1$, the α-interpolation space can also be characterized in terms of classical interpolation theory of real vector spaces (see e.g., Triebel, 1995). In particular, Steinwart and Scovel (2012, Theorem 4.6) proved the fact*

$$[\mathcal{H}]_X^\alpha \cong \left[L_2(\pi), [\mathcal{H}]_X^1\right]_{\alpha,2},$$

where the notation $\left[L_2(\pi), [\mathcal{H}]_X^1\right]_{\alpha,2}$ denotes the classical Hilbert space interpolation of $L_2(\pi)$ and $[\mathcal{H}]_X^1$ of degree α ("interpolation of the real method"). As the precise construction of this space is fairly technical and merely used as an alternative interpretation of $[\mathcal{H}]_X^\alpha$ here, we refer to Appendix D for more details.

3. Approximation of F_* with Vector-valued Interpolation Space

In this section, we deal with the misspecified setting where $F_* \notin \mathcal{G}$. To do this, we first define the *vector-valued interpolation space* via the tensor product space. We recall from Remark 1 that $L_2(\pi;\mathcal{Y})$ is isometrically isomorphic to $S_2(L_2(\pi),\mathcal{Y})$ and we denote by Ψ the isometric isomorphism between the two spaces. Similarly, we have $\mathcal{G} \simeq S_2(\mathcal{H}_X,\mathcal{Y})$ and we denote by $\bar{\Psi}$ the isometric isomorphism between both spaces in accordance with Theorem 1. This is summarized in Figure 1. The second chain of spaces is not isometric to the first but can be naturally embedded into the first as follows. Recall that we denote by $I_\pi : \mathcal{H}_X \to L_2(\pi)$ the embedding that maps each function to its equivalence class, $I_\pi(f) = [f]$. We therefore naturally define the embedding $\mathcal{I}_\pi : S_2(\mathcal{H}_X,\mathcal{Y}) \to S_2(L_2(\pi),\mathcal{Y})$ through $\mathcal{I}_\pi(g \otimes f) = g \otimes I_\pi(f) = g \otimes [f]$ for all $f \in \mathcal{H}_X$, $g \in \mathcal{Y}$, and obtain the extension to the whole space by linearity and continuity.[4] Therefore, for $F \in \mathcal{G}$ we define $[F] := \Psi \circ \mathcal{I}_\pi \circ \bar{\Psi}^{-1}(F)$. In the rest of the paper, every embedding will be denoted using the notation $[\,\cdot\,]$. A stricter requirement would be to write $[\,\cdot\,]_\pi$ due to dependence on the measure π, but we omit the subscript for ease of notation.

Table 2: Notation for spaces and operators

Symbol	Description	Symbol	Description
E_X	Covariate space	L_X	L_2-integral operator
$\mathcal{Y}$	Output space	C_{XX}	$\mathcal{H}_X$-covariance operator
$\mathcal{H}_X$	Scalar-valued RKHS	$\mathcal{G}$	Vector-valued RKHS
$[\mathcal{H}]_X^\alpha$	Scalar-valued α-interpolation space	$[\mathcal{G}]^\alpha$	Vector-valued α-interpolation space
k_X	Scalar-valued kernel	K	Vector-valued kernel
I_π	Scalar L_2-embedding operator	$\mathcal{I}_\pi$	Vector-valued L_2-embedding operator

Definition 2 (Vector-valued interpolation space) *Let k_X be a real-valued kernel with associated RKHS $\mathcal{H}_X$ and let $[\mathcal{H}]_X^\alpha$ be the real-valued interpolation space associated to $\mathcal{H}_X$ with some $\alpha \geq 0$. Since $[\mathcal{H}]_X^\alpha \subseteq L_2(\pi)$, it is natural to define the vector-valued interpolation space $[\mathcal{G}]^\alpha$ as*

$$[\mathcal{G}]^\alpha := \Psi\left(S_2([\mathcal{H}]_X^\alpha,\mathcal{Y})\right) = \{F \mid F = \Psi(C),\ C \in S_2([\mathcal{H}]_X^\alpha,\mathcal{Y})\}.$$

$[\mathcal{G}]^\alpha$ is a Hilbert space equipped with the norm

$$\|F\|_\alpha := \|C\|_{S_2([\mathcal{H}]_X^\alpha,\mathcal{Y})} \qquad (F \in [\mathcal{G}]^\alpha),$$

where $C = \Psi^{-1}(F)$. For $\alpha = 0$, we retrieve,

$$\|F\|_0 = \|C\|_{S_2(L_2(\pi),\mathcal{Y})}.$$

[4] $\mathcal{I}_\pi$ is formally the tensor product of the operator I_π with the operator $\mathrm{Id}_\mathcal{Y}$, see Aubin (2000, Definition 12.4.1.)

Remark 5 (Interpolation space inclusions) *The vector-valued interpolation space $[\mathcal{G}]^\alpha$ allows us to study the approximation of F_* in the misspecified case. Note that we have $F_* \in L_2(\pi; \mathcal{Y})$ since $Y \in L_2(\mathbb{P}; \mathcal{Y})$ by assumption. In light of Eq. (12), for $0 < \beta < \alpha$ we have*

$$[\mathcal{G}]^\alpha \hookrightarrow [\mathcal{G}]^\beta \hookrightarrow [\mathcal{G}]^0 \subseteq L_2(\pi; \mathcal{Y}).$$

While the well-specified case corresponds to $F_* \in \mathcal{G}$, the misspecified case corresponds to $F_* \in [\mathcal{G}]^\beta$ for some $0 \leq \beta < 1$. One can see from Remark 3 that under assumptions 1 to 3 and E_X being a second-countable locally compact Hausdorff space, $[\mathcal{G}]^0 = L_2(\pi; \mathcal{Y})$ if and only if k_X is c_0-universal.

Akin to Remark 4 for scalar-valued interpolation spaces, our next theorem shows that for $0 < \alpha < 1$, $[\mathcal{G}]^\alpha$ can be characterized in terms of interpolation spaces.

Theorem 2 (Vector-valued interpolation norm equivalence) *For $0 < \alpha < 1$,*

$$[\mathcal{G}]^\alpha \cong \left[L_2(\pi; \mathcal{Y}), [\mathcal{G}]^1\right]_{\alpha,2}.$$

This result allows us to obtain learning rates for vector-valued Sobolev spaces, see Section 6 for details.

4. Upper Learning Rates

In this section, we derive the learning rate for the difference between $[\hat{F}_\lambda]$ and F_* in the interpolation norm. As our assumptions match those of Fischer and Steinwart (2020), we include their corresponding labels for ease of reference. Recall that $(\mu_i)_{i \in I}$ are the positive eigenvalues of the integral operator. We now list our assumptions:

5. For some constants $c_1 > 0$ and $p \in (0,1]$ and for all $i \in I$,
$$\mu_i \leq c_1 i^{-1/p}. \tag{EVD}$$

6. For $\alpha \in [p,1]$, the inclusion map $I_\pi^{\alpha,\infty} : [\mathcal{H}]_X^\alpha \hookrightarrow L_\infty(\pi)$ is continuous, there is a constant $A > 0$ such that
$$\|I_\pi^{\alpha,\infty}\|_{[\mathcal{H}]_X^\alpha \to L_\infty(\pi)} \leq A \tag{EMB}$$

7. There exists $0 < \beta$ and a constant $B \geq 0$ such that $F_* \in [\mathcal{G}]^\beta$
$$\|F_*\|_\beta \leq B. \tag{SRC}$$

   We let $C_* := \Psi^{-1}(F_*) \in S_2([\mathcal{H}]_X^\beta, \mathcal{Y})$.

8. We assume that there are constants $\sigma, R > 0$ such that
$$\int_{\mathcal{Y}} \|y - F_*(x)\|_{\mathcal{Y}}^q p(x, dy) \leq \frac{1}{2} q! \sigma^2 R^{q-2}, \tag{MOM}$$

   is satisfied for π-almost all $x \in E_X$ and all $q \geq 2$.

(EVD) is a standard assumption on the eigenvalue decay of the integral operator (see more details in Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020). Property (EMB) is referred to as the *embedding property* in Fischer and Steinwart (2020). It can be shown that it holds if and only if there exists a constant $A \geq 0$ with $\sum_{i\in I} \mu_i^{\alpha} e_i^2(x) \leq A^2$ for π-almost all $x \in E_X$ (Fischer and Steinwart, 2020, Theorem 9). Since we assume k_X to be bounded, the embedding property always hold true when $\alpha = 1$. Furthermore, (EMB) implies a polynomial eigenvalue decay of order $1/\alpha$, which is why we take $\alpha \geq p$. (SRC) is justified by Remark 5 and is often referred to as the source condition in literature (Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020; Lin and Cevher, 2018; Lin et al., 2020). It measures the smoothness of the regression function F_*. In particular, when $\beta \geq 1$, the source condition implies that F_* has a representative from $\mathcal{G}$, indicating the well-specified scenario. However, once we let $\beta < 1$, we are in the misspecified learning setting, which is the main interest in this manuscript. Finally, the (MOM) condition on the Markov kernel $p(x, dy)$ is a Bernstein moment condition used to control the noise of the observations (see Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020 for more details). If Y is almost surely bounded, for example $\|Y\|_{\mathcal{Y}} \leq Y_\infty$ almost surely, then (MOM) is satisfied with $\sigma = R = 2Y_\infty$. It is possible to prove that the Bernstein condition is equivalent to sub-exponentiality, see Mollenhauer et al. (2022, Remark 4.9).

Theorem 3 (Upper learning rates) *Let $\mathcal{H}_X$ be a RKHS on E_X with respect to a kernel k_X such that assumptions 1 to 3 hold. Furthermore, let the conditions* (EVD), (EMB), (MOM) *be satisfied for some $0 < p \leq \alpha \leq 1$. For $0 \leq \gamma \leq 1$, if* (SRC) *is satisfied with $\gamma < \beta \leq 2$, and then*

1. *in the case $\beta + p \leq \alpha$, let $\lambda_n = \Theta\left(\left(n/\log^{\theta}(n)\right)^{-\frac{1}{\alpha}}\right)$ for some $\theta > 1$, for all $\tau > \log(5)$ and sufficiently large $n \geq 1$, there is a constant $J > 0$ independent of n and τ such that*

$$\left\|[\hat{F}_{\lambda}] - F_*\right\|_{\gamma}^{2} \leq \tau^2 J \left(\frac{n}{\log^{\theta} n}\right)^{-\frac{\beta-\gamma}{\alpha}}$$

   is satisfied with P^n-probability not less than $1 - 5e^{-\tau}$.

2. *in the case $\beta + p > \alpha$, let $\lambda_n = \Theta\left(n^{-\frac{1}{\beta+p}}\right)$, for all $\tau > \log(5)$ and sufficiently large $n \geq 1$, there is a constant $J > 0$ independent of n and τ such that*

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_{\gamma}^{2} \leq \tau^2 J n^{-\frac{\beta-\gamma}{\beta+p}}$$

   is satisfied with P^n-probability not less than $1 - 5e^{-\tau}$.

Remark 6 (Constants) *The index bound hidden in the phrase "sufficiently large n" depends on the parameters and constants from* (EVD) *and* (EMB), *on τ, on a lower bound $0 < c \leq 1$ for the operator norm $c \leq \|C_{XX}\|$, on $\|F_*\|_{L_{q_{\alpha,\beta}}(\pi;\mathcal{Y})}$ (see Remark 8 and Theorem 4 below) and on the regularization parameter sequence $(\lambda_n)_{n\geq 1}$. We can see that the constant J in Theorem 3 does not depend on $\|F_*\|_\infty$, and only depends on the parameters and constants from* (EVD), (EMB), (MOM), (SRC), *on $\|F_*\|_{L_{q_{\alpha,\beta}}(\pi;\mathcal{Y})}$, and on the regularization parameter sequence $(\lambda_n)_{n\geq 1}$.*

Theorem 3 states that the learning rate for $[\hat{F}_\lambda]$ is governed by the interplay between p, α, and β. To simplify the discussion, we focus on the $L_2(\pi;\mathcal{Y})$ learning rate, corresponding to $\gamma = 0$. The exponent $\beta/\max\{\alpha,\beta+p\}$ explicitly provides the learning rate. For example, if we have $\alpha \leq \beta$, we obtain a learning rate of $\beta/(\beta+p)$. In particular, for a Gaussian kernel on a bounded convex set E_X with π uniform on E_X, p and α are arbitrarily close to 0 (see e.g., Meunier et al., 2023, Example 2), and our learning rate can achieve $O(\log(n)/n)$ rate simply by taking $\lambda_n = \Theta\left((\log(n)/n)^{1/\beta}\right)$. We address the case of kernels with slower eigenvalue decay such as the Matérn kernel in Section 6.

Remark 7 (Saturation effect) *We note that for $\beta > 2$, the upper learning rate is still valid but saturates, i.e. the best upper bound of the generalization error (in L_2-norm) is $n^{-\frac{2}{2+p}}$. This phenomenon is commonly called the Tikhonov saturation effect and well known in classical regularisation theory (Engl et al., 1996) and kernel learning literature (see e.g. Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020; Rudi et al., 2015). In the real-valued setting, under the additional assumption that the kernel k_X is Hölder continuous and that the conditional variance of the noise is bounded away from* 0*, Li et al. (2022a) recently demonstrated that for any regularization parameter λ_n, the generalization error for kernel ridge regression is lower bounded with high probability by $n^{-\frac{2}{2+p}}$. This proves that the saturation effect is unavoidable when the algorithm employs Tikhonov regularization. To benefit from smoothness of the regression function beyond that saturation point at $\beta = 2$, one can employ different spectral regularization algorithms as explored by Blanchard and Mücke (2018). Proving the saturation effect for vector-valued regression with Tikhonov regularization, as well as exploring alternative spectral regularization algorithms, are very important research directions that we leave open for future works.*

Remark 8 (Boundedness condition) *When analyzing the RLS algorithm for both scalar and vector-valued outputs, it is standard to assume that the regression function $F_\star$ is bounded, as discussed in prior studies (see for example Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020). This boundedness is inherently met when $\beta \geq \alpha$, which falls in line with the assumption* (EMB)*. However, when $\beta < \alpha$, this condition must be explicitly assumed. Our previous work, Li et al. (2022b), did not provide a way to relax this assumption (see Appendix A). Nonetheless, when $\mathcal{Y} = \mathbb{R}$, recent insights by Zhang et al. (2023b) suggest that this boundedness criterion can be substituted with the requirement that F_* belongs to $L_q(\pi;\mathbb{R})$ for some $q \geq 2$. Furthermore, Zhang et al. (2023a) demonstrated that $F_* \in L_q(\pi;\mathbb{R})$ is automatically satisfied when $F_* \in [\mathcal{H}]_X^\beta$ leading to the L_q-embedding property of scalar valued interpolation spaces. Adopting a similar methodology, our Theorem 3 extends the scope of Zhang et al. (2023b,a) to the case where $\mathcal{Y}$ can be any Hilbert space, not just a subset of $\mathbb{R}$. To achieve this goal we first show that the boundedness requirement on F_* can be weakened to the assumption that $F_* \in L_q(\pi;\mathcal{Y})$ for some $q \geq 2$, and then we remove this assumption by showing that we have a continuous embedding $[\mathcal{G}]^\beta \hookrightarrow L_q(\pi;\mathcal{Y})$. This L_q-embedding property of vector-valued interpolation spaces is given in the next theorem. As will be demonstrated in subsequent sections, these enhancements, building upon the findings in Li et al. (2022b), are pivotal in achieving minimax rates for the RLS algorithm within many vector-valued RKHS, including vector-valued Sobolev spaces.*

Theorem 4 (**L_q-embedding property**) *Let Assumption (EMB) be satisfied with parameter $\alpha \in (0,1]$. For any $\beta \in (0,\alpha]$, the inclusion map*

$$I_\pi^{q_{\alpha,\beta}} : [\mathcal{G}]^\beta \hookrightarrow L_{q_{\alpha,\beta}}(\pi;\mathcal{Y})$$

is continuous, where $q_{\alpha,\beta} := \frac{2\alpha}{\alpha-\beta}$.

Notice that when we let $\beta \to \alpha$, we have $q_{\alpha,\beta} \to +\infty$ and we retrieve the property that $[\mathcal{G}]^\alpha \hookrightarrow L_\infty(\pi;\mathcal{Y})$. On the other hand, when $\beta \to 0$, we find $q_{\alpha,\beta} \to 2$ and we retrieve the property that $[\mathcal{G}]^\beta \hookrightarrow L_2(\pi;\mathcal{Y})$ for all $\beta \geq 0$. The L_q–embedding property allows to characterise the integrability of elements of $[\mathcal{G}]^\beta$ in the intermediate situations where $0 < \beta < \alpha$.

4.1 Rates for the general multiplicative kernel

As previously discussed in Remark 2, we show that the rates from Theorem 3 also hold for the kernel

$$\widetilde{K}(x,x') := k_X(x,x')T$$

where $T : \mathcal{Y} \to \mathcal{Y}$ is a bounded positive-semidefinite self-adjoint operator. Let $\tilde{\mathcal{G}}$ be the vRKHS induced by the kernel $\tilde{K}$. By Remark 2 (see also Carmeli et al., 2010, Example 3.2), we obtain upper rates for learning the conditional mean function $F_\star \in L_2(\pi;\mathcal{Y})$ with the general kernel $\tilde{K}$ by applying the transformation $y \mapsto T^{1/2}y$ for all $y \in \mathcal{Y}$ and simply invoking Theorem 3. That is, we are learning $\tilde{F}_\star \in L_2(\pi;\mathcal{Y})$ given by

$$\tilde{F}_\star(\cdot) := \mathbb{E}[T^{1/2}Y \mid X = \cdot]$$

with the kernel $K(x,x') = k_X(x,x')\,\mathrm{Id}_\mathcal{Y}$. We first notice that the conditions (EVD) and (EMB) do not depend on the choice of T (or equivalently, the choice of norm on $\mathcal{Y}$). Therefore, it remains to investigate the constants for which (MOM) and (SRC) hold for $\tilde{F}_\star$ with respect to the kernel K, under the assumption that $F_\star$ satisfies (MOM) and (SRC) with respect to the kernel $\tilde{K}$— this allows to apply Theorem 3 and directly extends the upper rates to the general case with adjusted constants.

We first verify (MOM) for $\tilde{F}_\star$ under the assumption that $F_\star$ satisfies (MOM) for some $\sigma, R > 0$. We have

$$\begin{aligned}\int_\mathcal{Y} \|T^{1/2}y - \tilde{F}_*(x)\|_\mathcal{Y}^q \, p(x,dy) &= \int_\mathcal{Y} \|T^{1/2}(y - F_*(x))\|_\mathcal{Y}^q \, p(x,dy) \\ &\leq \|T\|_{\mathcal{Y}\to\mathcal{Y}}^{q/2} \int_\mathcal{Y} \|y - F_*(x)\|_\mathcal{Y}^q \, p(x,dy) \leq \frac{1}{2} q! \tilde{\sigma}^2 \tilde{R}^{q-2}\end{aligned}$$

with $\tilde{\sigma} := \|T\|_{\mathcal{Y}\to\mathcal{Y}}^{1/2}\sigma$ and $\tilde{R} := \|T\|_{\mathcal{Y}\to\mathcal{Y}}^{1/2}R$.

We now assume $F_\star$ satisfies (SRC) with respect to interpolation space $[\tilde{\mathcal{G}}]^\beta$ induced by the kernel $\tilde{K}$. That is, we have $\|F_\star\|_{[\tilde{\mathcal{G}}]^\beta} < B$ for some $B \geq 0$ and $\beta > 0$. We recall $\tilde{\mathcal{G}} \simeq \mathcal{H} \otimes \tilde{\mathcal{Y}}$, where $\tilde{\mathcal{Y}} = \mathcal{Y}/\ker(T)$ equipped with the inner product $\langle y, y' \rangle_{\tilde{\mathcal{Y}}} = \langle y, Ty' \rangle_\mathcal{Y}$. Analogously to the interpolation space $[\mathcal{G}]^\beta$, we obtain $[\tilde{\mathcal{G}}]^\beta \simeq S_2([\mathcal{H}]_X^\beta, \tilde{\mathcal{Y}})$. Hence, there exists an orthogonal sequence $\{h_i\}_{i\in I}$ in $[\mathcal{H}]^\beta$ and some sequence $\{y_i\}_{i\in I}$ in $\mathcal{Y}$, such that isometrically, we have $F_\star \simeq \sum_{i\in I} y_i \otimes h_i$. By orthogonality of the $\{h_i\}_{i\in I}$, we have

$$\|F_\star\|^2_{[\tilde{\mathcal{G}}]^\beta} = \sum_{i\in I} \|h_i\|^2_{[\mathcal{H}]^\beta} \|T^{1/2} y_i\|^2_{\mathcal{Y}} = \|\tilde{F}_\star\|^2_{[\mathcal{G}]^\beta},$$

confirming (SRC) for $\tilde{F}_\star$ with respect to the interpolation space $[\mathcal{G}]^\beta$ without adjusting the constant B.

5. Lower Bound

Our final theorem provides a lower bound for the convergence rates, which eventually allows us to confirm the optimality of the learning rates given in the preceding section. In deriving the lower bound, we need the following extra assumption.

8. For some constants $c_1, c_2 > 0$ and $p \in (0,1]$ and for all $i \in I$,

$$c_2 i^{-1/p} \leq \mu_i \leq c_1 i^{-1/p} \tag{EVD+}$$

Theorem 5 (Lower learning rates) *Let k_X be a kernel on E_X such that assumptions 1 to 3 hold and π be a probability distribution on E_X such that* (EVD+) *holds with $0 < p \leq 1$. Then for all parameters $0 < \beta \leq 2$, $0 \leq \gamma \leq 1$ with $\gamma < \beta$ and all constants σ, R, B, there exist constants $J_0, J, \theta > 0$ such that for all learning methods $D \to \hat{F}_D$ ($D := \{(x_i, y_i)\}_{i=1}^n$), all $\tau > 0$, and all sufficiently large $n \geq 1$ there is a distribution P defined on $E_X \times \mathcal{Y}$ used to sample D, with marginal distribution π on E_X, such that* (SRC) *with respect to B, β and* (MOM) *with respect to σ, R are satisfied, and with P^n-probability not less than $1 - J_0 \tau^{1/\theta}$,*

$$\|[\hat{F}_D] - F_*\|^2_\gamma \geq \tau^2 J n^{-\frac{\beta-\gamma}{\beta+p}}.$$

Theorem 5 states that under the assumptions of Theorem 3 and Assumption (EVD+) no learning method can achieve a learning rate faster than

$$n^{-\frac{\beta-\gamma}{\beta+p}}. \tag{13}$$

To our knowledge, this is the first analysis that demonstrates the lower rate for vector-valued regression in infinite dimension. In the context of regularized regression, Caponnetto and De Vito (2007), Steinwart et al. (2009) and Blanchard and Mücke (2018) provide lower bounds on the learning rate under comparable assumptions. However, one key difference in our analysis is that the output of the regression learning now takes values in a potentially infinite dimensional Hilbert space $\mathcal{Y}$, rather than in $\mathbb{R}$ or $\mathbb{R}^d$.

Our analysis reveals that for $\beta \geq \alpha - p$, the RLS estimator leads to the minimax optimal rate (by combining Theorem 5 and case 2 in Theorem 3), namely $O(n^{-(\beta-\gamma)/(\beta+p)})$. This scenario is particularly relevant for vector-valued Sobolev RKHSs where $p = \alpha$, a topic we will explore in the following section. We point out that finding the optimal rate for $\beta < \alpha - p$ remains a longstanding challenge, even when the output is in $\mathbb{R}$.

6. Example: Vector-valued Sobolev Space

In this section we illustrate our main results in the case of vector-valued Sobolev RKHSs. To this end, we assume that $E_X \subseteq \mathbb{R}^d$ is a bounded domain with smooth boundary equipped with the Lebesgue measure μ. $L_2(E_X;\mathcal{Y}) \coloneqq L_2(E_X,\mu;\mathcal{Y})$ denotes the corresponding Bochner space. We start by introducing vector-valued Sobolev spaces.

Definition 3 (vSobolev space) *For $m \in \mathbb{N}$, the vector-valued Sobolev space $W^{m,2}(E_X;\mathcal{Y})$ is the Hilbert space of all $f \in L_2(E_X;\mathcal{Y})$ whose weak derivatives of all orders*[5] *$|r| \leqslant m$ exist and belong to $L_2(E_X;\mathcal{Y})$, endowed with the norm*

$$\|f\|^2_{W^{m,2}(E_X;\mathcal{Y})} \coloneqq \sum_{|r|\leqslant m} \|\partial^r f\|^2_{L_2(E_X;\mathcal{Y})}\,.$$

For $m = 0$, $W^{0,2}(E_X;\mathcal{Y}) \coloneqq L_2(E_X;\mathcal{Y})$.

For the definition of weak derivatives of functions in $L_2(E_X;\mathcal{Y})$ see Aubin (2000, Section 12.7). The following theorem allows us to connect vector-valued Sobolev spaces to our framework.

Theorem 6 (Aubin 2000, Theorem 12.7.1) *For $m \in \mathbb{N}$, the vSobolev space $W^{m,2}(E_X;\mathcal{Y})$ is isometric to the Hilbert tensor product $\mathcal{Y} \otimes W^{m,2}(E_X)$, where $W^{m,2}(E_X) \coloneqq W^{m,2}(E_X;\mathbb{R})$ is the standard scalar-valued Sobolev space.*

When k_X is a translation invariant kernel on $\mathbb{R}^d$ whose Fourier transform behaves as $\left(1 + \|\cdot\|_2^2\right)^{-m}$ with $m > d/2$, such as the Matérn kernel (see Definition 5), the induced RKHS $\mathcal{H}_X$ restricted to E_X coincides with $W^{m,2}(E_X)$, and their norms are equivalent, see Wendland (2004, Corollary 10.13 and Theorem 10.46). Therefore, if we choose such a kernel k_X to construct $K = k_X \operatorname{Id}_{\mathcal{Y}}$ and $\mathcal{G}$, we obtain by Theorem 1 and Theorem 6,

$$\mathcal{G} \simeq \mathcal{Y} \otimes \mathcal{H}_X \simeq \mathcal{Y} \otimes W^{m,2}(E_X) \simeq W^{m,2}(E_X;\mathcal{Y}).$$

The induced vector-valued RKHS therefore corresponds to a vector-valued Sobolev space. Furthermore, the interpolation spaces $[\mathcal{G}]^\alpha$, $\alpha \geq 0$, can be characterized as a vector-valued fractional Sobolev space (see e.g., Section 5.6 Hytönen et al., 2016, for more details).

Definition 4 (Vector-Valued Fractional Sobolev Space) *Fix $r > 0$, and let $m \coloneqq \min\{s \in \mathbb{N} : s > r\}$. The vector-valued fractional Sobolev space $W^{r,2}(E_X;\mathcal{Y})$ is defined by means of the real interpolation method, namely*

$$W^{r,2}(E_X;\mathcal{Y}) \coloneqq \left[L_2(E_X;\mathcal{Y}), W^{m,2}(E_X;\mathcal{Y})\right]_{r/m,2}\,.$$

Proposition 1 *For all $r \geq 0$, $W^{r,2}(E_X;\mathcal{Y}) \cong \mathcal{Y} \otimes W^{r,2}(E_X)$. Furthermore, if k_X is a kernel on E_X such that $\mathcal{G} \simeq W^{m,2}(E_X;\mathcal{Y})$ with $m > d/2$, then for all $r \geq 0$,*

$$[\mathcal{G}]^{r/m} \cong W^{r,2}(E_X;\mathcal{Y}).$$

i.e., $[\mathcal{G}]^{r/m} = W^{r,2}(E_X;\mathcal{Y})$ with equivalent norms.

[5] $r \coloneqq (r_1,\ldots,r_d) \in \mathbb{N}^d$ is a multi-index and $|r|$ denotes the sum of its values.

Definition 5 (Matérn kernel) *For $m \in \mathbb{N}$ with $m > d/2$, the Matérn kernel of order m is defined as for all $x, x' \in \mathbb{R}^d$*

$$k_X\left(x', x\right) = \frac{1}{2^{m-d/2-1}\Gamma(m-d/2)}\left(\sqrt{2(m-d/2)}\left\|x'-x\right\|\right)^{m-d/2}\mathcal{K}_{m-d/2}\left(\sqrt{2(m-d/2)}\left\|x'-x\right\|\right)$$

where $\mathcal{K}_{m-d/2}$ is the modified Bessel function of the second kind of order $m-d/2$ and Γ is the Gamma function (see e.g., Kanagawa et al., 2018, Examples 2.2 and 2.6).

We now specify Theorem 3 and Theorem 5 in the setting of vector-valued Sobolev spaces. We make the assumption that the marginal π is equivalent to the Lebesgue measure so that $L_2(E_X;\mathcal{Y}) \simeq L_2(E_X,\pi;\mathcal{Y})$.

Corollary 2 (vSobolev Upper Rates) *Let k_X be a kernel on E_X such that assumptions 1 to 3 hold and such that $\mathcal{G} \simeq W^{m,2}(E_X;\mathcal{Y})$ with $m > d/2$, and let P be a probability distribution on $E_X \times \mathcal{Y}$ such that $\pi := P_{E_X}$ (the marginal distribution on E_X) is equivalent to the Lebesgue measure μ on E_X. Furthermore, let $B > 0$ be a constant such that $\|F_*\|_{W^{s,2}(E_X;\mathcal{Y})} \leq B$ for some $0 < s \leq 2m$, and* (MOM) *be satisfied. Then, for $0 \leq t < s$ and a choice $\lambda_n = \Theta\left(n^{-\frac{m}{s+d/2}}\right)$, for all $\tau > \log(5)$ and sufficiently large $n \geq 1$, there is a constant $J > 0$ independent of n and τ such that*

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|^2_{W^{t,2}(E_X;\mathcal{Y})} \leq \tau^2 J n^{-\frac{s-t}{s+d/2}}$$

is satisfied with P^n-probability not less than $1 - 5e^{-\tau}$.

Corollary 3 (vSobolev Lower Rates) *Let k_X be a kernel on E_X such that assumptions 1 to 3 hold and such that $\mathcal{G} \simeq W^{m,2}(E_X;\mathcal{Y})$ with $m > d/2$, P be a probability distribution on $E_X \times \mathcal{Y}$ such that $\pi := P_{E_X}$ (the marginal distribution on E_X) is equivalent to the Lebesgue measure μ on E_X. Then for all parameters $0 \leq t < s \leq 2m$, and all constants $\sigma, R, B > 0$ there exist constants $J_0, J, \theta > 0$ such that for all learning methods $D \to \hat{F}_D$ ($D := \{(x_i, y_i)\}_{i=1}^n$), all $\tau > 0$, and all sufficiently large $n \geq 1$ there is a distribution P defined on $E_X \times \mathcal{Y}$ used to sample D, with marginal distribution π on E_X, such that $\|F_*\|_{W^{s,2}(E_X;\mathcal{Y})} \leq B$ and* (MOM) *with respect to σ, R, are satisfied, and with P^n-probability not less than $1 - J_0\tau^{1/\theta}$,*

$$\|[\hat{F}_D] - F_*\|^2_{W^{t,2}(E_X;\mathcal{Y})} \geq \tau^2 J n^{-\frac{s-t}{s+d/2}}.$$

Corollary 2 and 3 are proved by inserting the values for p, α and β from (EVD), (EMB) and (SRC) into Theorem 3 and 5. For $\mathcal{G} \simeq W^{m,2}(E_X;\mathcal{Y})$ and π equivalent to the Lebesgue measure, we show in the appendix that $p = \frac{d}{2m}$, $\alpha = p + \epsilon$ for all $\epsilon > 0$ (Proposition 4 in the appendix), and by Proposition 1, $W^{s,2}(E_X;\mathcal{Y}) \cong [\mathcal{G}]^{s/m}$, which implies $\beta = s/m$. Since $\alpha - p$ is arbitrarily close to zero, we are in the regime $\beta + p > \alpha$ and achieve the rate $n^{-\frac{\beta-\gamma}{\beta+p}} = n^{-\frac{s-t}{s+d/2}}$ by Theorem 3.

Our results show that the RLS estimator in Eq. (1) leads to minimax optimal rates for any $\beta \in (0, 2]$ when $\mathcal{G}$ is a vector-valued Sobolev RKHS, since the rates obtained in Corollary 2 match the lower bound in Corollary 3. This aligns with the recent findings obtained in Zhang et al. (2023b) for scalar-valued Sobolev RKHS.

7. Related Work

In this section, we compare our results with learning rates obtained in the literature. Due to the large amount of available types of rates for the scalar learning setting, we primarily focus on optimal rates derived under comparable assumptions on the underlying distributions. For the much less investigated vector-valued learning case, we provide a more general overview of recent results.

As discussed previously, the closest work to our results is Caponnetto and De Vito (2007). By assuming that $F_* \in \mathcal{G}$, Caponnetto and De Vito (2007) provide the first analysis of RLS with Tikhonov regularization for when $\mathcal{Y}$ is infinite dimensional. In this work, the smoothness of the Bayes function is naturally expressed as an element of the range of a power iterate of the corresponding covariance operator (this is known as a *Hölder source condition* in regularization theory, see e.g. Blanchard and Mücke, 2018). They show the L_2 learning rate $n^{-\beta/(\beta+p)}$, when $K(x,x')$ is trace class for all $x, x' \in E_X$—this condition is violated for the standard choice of kernel $K(x,x') = k_X(x,x')\,\mathrm{Id}_{\mathcal{Y}}$ whenever $\mathcal{Y}$ is infinite dimensional. For finite dimensional $\mathcal{Y}$, Caponnetto and De Vito (2007) obtain the matching lower bound. In contrast, we study the RLS algorithm beyond the well-specified setting. Our analysis covers both the well-specified case and the hard learning scenario with $F_* \notin \mathcal{G}$, without assuming the trace class condition for the kernel K. When $F_* \in \mathcal{G}$, the construction of our vector-valued interpolation space can be interpreted as a generalisation of the Hölder source condition (this is seen by interpreting the covariance operator in terms of the tensor product structure of $\mathcal{G}$, see Mollenhauer et al., 2022). Hence, in the well-specified case, our rates recover the same rate as Caponnetto and De Vito (2007). Moreover, instead of L_2-rate, we derive a general γ-learning rate, such that the L_2 learning rate is recovered when $\gamma = 0$. Finally, we obtain the dimension-free matching lower rate without requiring finite dimensional $\mathcal{Y}$.

In the real-valued RLS setting, Blanchard and Mücke (2018) and Fischer and Steinwart (2020) provide the γ-learning rate with general regularization schemes for the well-specified case under the Hölder source condition (Blanchard and Mücke, 2018) and Tikhonov regularization for the misspecified case based on real-valued interpolation spaces (Fischer and Steinwart, 2020) respectively. For Tikhonov RLS in the well-specified regime, they obtain the same L_2 learning rate as in Caponnetto and De Vito (2007) given by $n^{-\beta/(\beta+p)}$. They both provide the matching lower bound when $F_* \in \mathcal{G}$. A key difference between the two is that Fischer and Steinwart (2020) extend the learning rate analysis to the hard learning scenario by employing the embedding property. We use similar techniques to Fischer and Steinwart (2020), and generalize the study to the vector-valued RLS setting through our construction of vector-valued interpolation spaces. Thus, when $\mathcal{Y}$ is real-valued, our results recover the known kernel ridge regression rate of Fischer and Steinwart (2020).

In addition to the previously mentioned work, there are some comparable results for infinite-dimensional RLS which do not explicitly contain optimal upper rates and/or do not provide corresponding lower bounds. To our knowledge, Mollenhauer et al. (2022) derive the first upper learning rates for the infinite-dimensional RLS algorithm for the case of general regularisation schemes which are not exclusively based on vector-valued RKHSs. These rates hold for the γ-norm and cover our setting with the kernel $K(x,x') = k_X(x,x')\,\mathrm{Id}_{\mathcal{Y}}$ as a special case. Technically, their approach is similar to the real-valued analysis by Blanchard and Mücke (2018)—thus, they only cover the well-specified setting under Hölder source conditions. As a major difference compared to our results, Mollenhauer et al. (2022) only consider rates up to the order $O(n^{-1/2})$ without additional assumptions about the marginal of X (which are needed for faster rates). Singh et al. (2019) study the vector-valued RLS problem in a setting which is similar to ours. They obtain a suboptimal $O(n^{-1/4})$ upper rate in the well-specified setting, however, due to the use of a less sharp concentration bound. Finally, there are extensive studies concerning real-valued RLS (see e.g., Bauer et al., 2007; Smale and Zhou, 2007; Dicker et al., 2017; Lin et al., 2020; Lin and Cevher, 2018; Steinwart and Christmann, 2008; Steinwart et al., 2009, and references therein). In particular, Lin and Cevher (2018) derive the γ-learning rate of $n^{-(\beta-\gamma)/\max\{\beta+p,1\}}$ using the integral operator technique, while Steinwart et al. (2009) obtain an L_2 rate of $n^{-\beta/\max\{\beta+p,1\}}$ using an empirical process technique.

Acknowledgements

MM is partly supported by the Deutsche Forschungsgemeinschaft (DFG) through grant EXC 2046 *MATH+*, Project EF1-19: *Machine Learning Enhanced Filtering Methods for Inverse Problems*. ZL, DM and AG are supported by the Gatsby Charitable Foundation. The authors wish to thank Haobo Zhang, Zikai Shen and Ingo Steinwart for helpful discussions.

References

Robert Adams and John Fournier. ***Sobolev spaces***. Elsevier, 2003.

Jean-Pierre Aubin. ***Applied Functional Analysis***. John Wiley & Sons, Inc., 2nd edition, 2000.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. ***Journal of Complexity***, 23(1):52–72, 2007.

Alain Berlinet and Christine Thomas-Agnan. ***Reproducing Kernel Hilbert Spaces in Probability and Statistics***. Springer Science & Business Media, 2011.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. ***Foundations of Computational Mathematics***, 18(4):971–1013, 2018.

V.I. Bogachev. ***Gaussian Measures***. American Mathematical Society, 1998.

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

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. ***Analysis and Applications***, 4(04):377–408, 2006.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. ***Analysis and Applications***, 8(01): 19–61, 2010.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. ***Advances in Neural Information Processing Systems***, 29, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. ***The Journal of Machine Learning Research***, 21(1):3852–3918, 2020.

Ernesto De Vito, Lorenzo Rosasco, and Andrea Caponnetto. Discretization error analysis for tikhonov regularization. ***Analysis and Applications***, 4(01):81–99, 2006.

L.H. Dicker, D.P. Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. ***Electronic Journal of Statistics***, 11:1022–1047, 2017.

Joe Diestel and J.J. Uhl. ***Vector Measures***. American Mathematical Society, 1977.

R.M. Dudley. ***Real Analysis and Probability***. Cambridge University Press, 2nd edition edition, 2002.

D.E. Edmunds and Hans Triebel. ***Function Spaces, Entropy Numbers, Differential Operators***. Cambridge University Press, 1996.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. ***Regularization of Inverse Problems***, volume 375 of ***Mathematics and its Applications***. Kluwer Academic Publishers Group, 1996.

Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. ***Journal of Machine Learning Research***, 2:243–264, 2002.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. ***Journal Of Machine Learning Research***, 21:205–1, 2020.

Arthur Gretton. Introduction to RKHS, and some simple kernel algorithms. ***Lecture Notes, University College London***, 2013.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In ***Proceedings of the 29th International Coference on International Conference on Machine Learning***, pages 1803—-1810, 2012a.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimilano Pontil, and Arthur Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In ***Proceedings of the 29th International Conference on Machine Learning***, pages 535–542, 2012b.

Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. ***Analysis in Banach spaces. Volume I: Martingales and Littlewood-Paley Theory***. Springer, 2016.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. ***Journal of Machine Learning Research***, 17(20):1–54, 2016.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. ***arXiv preprint arXiv:1807.02582***, 2018.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces. ***Advances in Neural Information Processing Systems***, 35:4017–4031, 2022.

Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Koopman operator learning: Sharp spectral rates and spurious eigenvalues. ***arXiv preprint arXiv:2302.02004***, 2023.

Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In ***The Eleventh International Conference on Learning Representations***, 2022a.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In ***Advances in Neural Information Processing Systems***, volume 35, pages 4433–4445, 2022b.

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In ***International Conference on Machine Learning***, pages 3092–3101. PMLR, 2018.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. ***Applied and Computational Harmonic Analysis***, 48(3):868–890, 2020.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In ***International Conference on Machine Learning***, pages 7512–7523. PMLR, 2021.

Dimitri Meunier, Zhu Li, Arthur Gretton, and Samory Kpotufe. Nonlinear meta-learning can guarantee faster rates. ***arXiv preprint arXiv:2307.10870***, 2023.

Mattes Mollenhauer and Péter Koltai. Nonparametric approximation of conditional expectation operators. ***arXiv preprint arXiv:2012.12917***, 2020.

Mattes Mollenhauer, Nicole Mücke, and T.J. Sullivan. Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem. ***arXiv preprint arXiv:2211.08875***, 2022.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. ***Advances in Neural Information Processing Systems***, 33: 21247–21259, 2020.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In ***Advances in Neural Information Processing Systems***, pages 1657–1665, 2015.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. ***Advances in Neural Information Processing Systems***, 32, 2019.

Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. ***Biometrika***, page asad042, 07 2023. ISSN 1464-3510.

Steve Smale and Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. ***Bulletin of the American Mathematical Society***, 41(3):279–305, 2004.

Steve Smale and Ding-Xuan Zhou. Shannon sampling II: Connections to learning theory. ***Applied and Computational Harmonic Analysis***, 19(3):285–302, 2005.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. ***Constructive approximation***, 26(2):153–172, 2007.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In ***Proceedings of the 26th Annual International Conference on Machine Learning***, pages 961–968. ACM, 2009.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. ***Journal of Machine Learning Research***, 12(Jul):2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. ***Support Vector Machines***. Springer, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. ***Constructive Approximation***, 35(3):363–417, 2012.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In ***Proceedings of the 22nd Annual Conference on Learning Theory***, pages 79–93, 2009.

Hans Triebel. ***Interpolation Theory, Function Spaces, Differential Operators***. J.A. Barth, 2nd edition, 1995.

Joachim Weidmann. ***Linear Operators in Hilbert Spaces***. Springer, 1980.

Holger Wendland. ***Scattered Data Approximation***. Cambridge University Press, 2004.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. ***arXiv preprint arXiv:2303.14942***, 2023a.

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. On the optimality of misspecified kernel ridge regression. ***arXiv preprint arXiv:2305.07241***, 2023b.

Appendices

We now report proofs which were omitted in the main text. As discussed in Remark 8, the proof for the upper rates given in Theorem 3 builds on the results from conditional mean embedding learning in Li et al. (2022b). In particular, we proceed in two steps to prove Theorem 3. Firstly, in Section A, we extend the analysis in Li et al. (2022b) to general vector-valued regression setting with a bounded regression function. Secondly, in Section B, we replace this boundedness assumption by a weaker integrability condition. Finally, using Theorem 4, we show that this integrability condition can be removed, leading to Theorem 3. The proof of Theorem 4 is provided at the end of Section B. In Section C, we prove the lower bound on the rates given in Theorem 5. Section D contains the proof Theorem 2. Section E contains the proofs for the results related to Sobolev spaces presented in Section 6. Finally, in Section F, we collect some technical supporting results.

Appendix A. Learning rates for bounded regression function

Theorem 7 *Let $\mathcal{H}_X$ be a RKHS on E_X with respect to a kernel k_X such that assumptions 1 to 3 hold. Let P be a probability distribution on $E_X \times \mathcal{Y}$ with $\pi := P_{E_X}$ (the marginal distribution on E_X). Furthermore, let the conditions* (EVD), (EMB), (MOM) *be satisfied for some $0 < p \leq \alpha \leq 1$ and let $B_\infty > 0$ be a constant with $\|F_*\|_{L_\infty(\pi;\mathcal{Y})} \leq B_\infty$. Then for $0 \leq \gamma \leq 1$, if* (SRC) *is satisfied with $\gamma < \beta \leq 2$,*

1. *in the case $\beta + p \leq \alpha$ and $\lambda_n = \Theta\left(\left(n/\log^\theta(n)\right)^{-\frac{1}{\alpha}}\right)$ for some $\theta > 1$, for all $\tau > \log(4)$ and sufficiently large $n \geq 1$, there is a constant $J > 0$ independent of n and τ such that*

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 \leq \tau^2 J \left(\frac{n}{\log^\theta n}\right)^{-\frac{\beta-\gamma}{\alpha}}$$

   is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$.

2. *in the case $\beta + p > \alpha$ and $\lambda_n = \Theta\left(n^{-\frac{1}{\beta+p}}\right)$, for all $\tau > \log(4)$ and sufficiently large $n \geq 1$, there is a constant $J > 0$ independent of n and τ such that*

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 \leq \tau^2 J n^{-\frac{\beta-\gamma}{\beta+p}}$$

   is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$.

Remark 9 *The proof of Theorem 7 reveals that the index bound hidden in the phrase "sufficiently large n" just depends on the parameters and constants from* (EVD) *and* (EMB), *on τ, on a lower bound $0 < c \leq 1$ for the operator norm $c \leq \|C_{XX}\|$, and on the regularization parameter sequence $(\lambda_n)_{n\geq1}$. Moreover, the constant J only depends on the parameters and constants from* (EVD), (EMB), (MOM), (SRC), *on B_∞, and on the regularization parameter sequence $(\lambda_n)_{n\geq1}$.*

Structure of the proof. Recall that $\hat{F}_\lambda \in \mathcal{G}$ is defined as $\hat{F}_\lambda := \bar{\Psi}\left(\hat{C}_\lambda\right)$ where $\hat{C}_\lambda$ is solution of Eq. (8). We introduce its population counterpart, the solution of the following problem:

$$C_\lambda := \underset{C \in S_2(\mathcal{H}_X, \mathcal{Y})}{\arg\min}\ \mathbb{E}_P \|Y - C\phi_X(X)\|_{\mathcal{Y}}^2 + \lambda \|C\|_{S_2(\mathcal{H}_X,\mathcal{Y})}^2, \qquad F_\lambda := \bar{\Psi}\left(C_\lambda\right) \in \mathcal{G}.$$

It can be readily shown (see for example Appendix D.1 Grünewälder et al., 2012b and Corollary 7.4 Mollenhauer and Koltai, 2020) that

$$\begin{aligned} C_\lambda &= C_{YX}\left(C_{XX} + \lambda Id_{\mathcal{H}_X}\right)^{-1}, \\ \hat{C}_\lambda &= \hat{C}_{YX}\left(\hat{C}_{XX} + \lambda Id_{\mathcal{H}_X}\right)^{-1}, \end{aligned}$$

where $Id_{\mathcal{H}_X}$ is the identity operator in $\mathcal{H}_X$ and

$$\begin{aligned} C_{XX} &= \mathbb{E}[\phi_X(X) \otimes \phi_X(X)] & C_{YX} &= \mathbb{E}[Y \otimes \phi_X(X)] \\ \hat{C}_{XX} &= \frac{1}{n}\sum_{i=1}^n \phi_X(x_i) \otimes \phi_X(x_i) & \hat{C}_{YX} &= \frac{1}{n}\sum_{i=1}^n y_i \otimes \phi_X(x_i). \end{aligned}$$

Finally, recall that $F_* \in L_2(\pi;\mathcal{Y})$ and $C_* := \Psi^{-1}\left(F_*\right)$ is in $S_2(L_2(\pi),\mathcal{Y})$. From the definition of the vector-valued interpolation norm we introduce the following decomposition,

$$\begin{aligned} \left\|[\hat{F}_\lambda] - F_*\right\|_\gamma &\le \left\|\left[\hat{F}_\lambda - F_\lambda\right]\right\|_\gamma + \left\|[F_\lambda] - F_*\right\|_\gamma \\ &= \left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|_{S_2\left([\mathcal{H}]_X^\gamma,\mathcal{Y}\right)} + \left\|[C_\lambda] - C_*\right\|_{S_2\left([\mathcal{H}]_X^\gamma,\mathcal{Y}\right)} \end{aligned} \tag{14}$$

We can see that the error for the first term is mainly due to the sample approximation. We therefore refer to the first term as the *Variance*. We refer to the second term as the *Bias*. Our proof of convergence of the bias adapts the proof by Fischer and Steinwart (2020), and utilizes the fact that C_* is Hilbert-Schmidt to obtain a sharp rate.

A.1 Bounding the Bias

In this section, we establish the bound on the bias. The key insight is that due to Aubin (2000, Theorem 12.6.1) the conditional mean function can be expressed as a Hilbert-Schmidt operator. The proof generalizes Fischer and Steinwart (2020, Lemma 14), which addresses the scalar case; and Singh et al. (2019, Theorem 6).

Lemma 1 *If $F_* \in [\mathcal{G}]^\beta$ is satisfied for some $0 \le \beta \le 2$, then the following bound is satisfied, for all $\lambda > 0$ and $0 \le \gamma \le \beta$:*

$$\left\|[F_\lambda] - F_*\right\|_\gamma^2 \le \|F_*\|_\beta^2\, \lambda^{\beta-\gamma} \tag{15}$$

Proof We first recall that since $F_* \in [\mathcal{G}]^\beta$, $F_* = \Psi\left(C_*\right)$ with $C_* \in S_2([\mathcal{H}]_X^\beta,\mathcal{Y})$, furthermore $F_\lambda = \bar{\Psi}\left(C_\lambda\right)$ with $C_\lambda \in S_2(\mathcal{H}_X,\mathcal{Y})$. Hence, $\|[F_\lambda] - F_*\|_\gamma = \|[C_\lambda] - C_*\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}$ and $\|F_*\|_\beta = \|C_*\|_{S_2\left([\mathcal{H}]_X^\beta,\mathcal{Y}\right)}$. We first decompose $[C_\lambda] - C_*$, and follow this by establishing an upper bound on the bias. Since $C_* \in S_2([\mathcal{H}]_X^\beta,\mathcal{Y}) \subseteq S_2(\overline{\operatorname{ran} I_\pi},\mathcal{Y})$, it admits the decomposition

$$C_* = \sum_{i \in I}\sum_{j \in J} \breve{a}_{ij} d_j \otimes [e_i].$$

where $(d_j)_{j\in J}$ is any countable basis of $\mathcal{Y}$ and $\sum_{i\in I}\sum_{j\in J}\check{a}_{ij}^2 < +\infty$ with $\check{a}_{ij} = \langle C_*, d_j \otimes [e_i]\rangle_{S_2(L_2(\pi),\mathcal{Y})} = \langle C_*[e_i], d_j\rangle_{\mathcal{Y}}$ for all $i \in I, j \in J$ (see e.g. Gretton (2013), Lecture on "testing statistical dependence"). On the other hand, $C_\lambda = C_{YX}\left(C_{XX} + \lambda Id_{\mathcal{H}_X}\right)^{-1}$. Since $\left(\mu_i^{1/2} e_i\right)_{i\in I}$ is an ONB of $(\ker I_\pi)^\perp$, we can complete it with an at most countable basis $(\bar{e}_i)_{i\in I'}$ (with $I \cap I' = \varnothing$) of $\ker I_\pi$ such that the union of the family forms a basis of $\mathcal{H}_X$. We get a basis of $S_2(\mathcal{H}_X,\mathcal{Y})$ through $(d_j \otimes f_i)_{i\in I\cup I', j\in J}$ where $f_i = \mu_i^{1/2} e_i$ if $i \in I$ and $f_i = \bar{e}_i$ if $i \in I'$. By the spectral decomposition of C_{XX} Eq. (9), for $a > 0$ we then have

$$\left(C_{XX} + \lambda Id_{\mathcal{H}_X}\right)^{-a} = \sum_{i\in I}\left(\mu_i + \lambda\right)^{-a}\left\langle \mu_i^{1/2} e_i, \cdot\right\rangle_{\mathcal{H}_X} \mu_i^{1/2} e_i + \lambda^{-a}\sum_{i\in I'}\langle \bar{e}_i, \cdot\rangle_{\mathcal{H}_X}\bar{e}_i.$$

Furthermore,

$$\begin{aligned}
C_{YX} &= \mathbb{E}_P\left[Y \otimes \phi_X(X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[Y\right] \otimes \phi_X(X)\right] \\
&= \mathbb{E}_X\left[F_*(X) \otimes \phi_X(X)\right] \\
&= \mathbb{E}_X\left[\Psi\left(C_*\right)(X) \otimes \phi_X(X)\right] \\
&= \sum_{i\in I}\sum_{j\in J}\check{a}_{ij}\mathbb{E}_X\left[\Psi\left(d_j \otimes [e_i]\right)(X) \otimes \phi_X(X)\right] \\
&= \sum_{i\in I}\sum_{j\in J}\check{a}_{ij}\mathbb{E}_X\left[[e_i](X) d_j \otimes \phi_X(X)\right].
\end{aligned}$$

In the last step we used the explicit form of the isomorphism between $L_2(\pi;\mathcal{Y})$ and $S_2(L_2(\pi),\mathcal{Y})$ mentioned in Remark 1: Ψ is characterized by $\Psi(g \otimes f) = (x \mapsto g f(x))$, for all $g \in \mathcal{Y}, f \in L_2(\pi)$. Then, using that $([e_i])_{i\in I}$ is an ONS in $L_2(\pi)$,

$$[C_\lambda] = \sum_{i\in I}\sum_{j\in J}\check{a}_{ij}\frac{\mu_i}{\lambda + \mu_i} d_j \otimes [e_i],$$

and hence

$$[C_\lambda] - C_* = -\sum_{i\in I}\sum_{j\in J}\check{a}_{ij}\frac{\lambda}{\lambda + \mu_i} d_j \otimes [e_i].$$

We are now ready to compute the upper bound. Parseval's identity w.r.t. the ONB $\left(d_j \otimes \mu_i^{\gamma/2}[e_i]\right)_{i\in I, j\in J}$ of $S_2\left([\mathcal{H}]_X^\gamma, \mathcal{Y}\right)$ yields

$$\begin{aligned}
\left\|[C_\lambda] - C_*\right\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}^2 &= \left\|\sum_{i\in I}\sum_{j\in J}\check{a}_{ij}\frac{\lambda}{\lambda + \mu_i} d_j \otimes [e_i]\right\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}^2 \\
&= \sum_{i\in I}\sum_{j\in J}\check{a}_{ij}^2\left(\frac{\lambda}{\lambda + \mu_i}\right)^2 \mu_i^{-\gamma}.
\end{aligned}$$

Next we notice that,

$$\begin{aligned}\left(\frac{\lambda}{\mu_i+\lambda}\right)^2\mu_i^{-\gamma} &= \left(\frac{\lambda}{\mu_i+\lambda}\right)^2\mu_i^{-\gamma}\left(\frac{\lambda}{\lambda}\frac{\mu_i+\lambda}{\mu_i+\lambda}\right)^{\beta-\gamma}\\ &= \lambda^{\beta-\gamma}\mu_i^{-\beta}\left(\frac{\lambda}{\mu_i+\lambda}\right)^2\left(\frac{\mu_i}{\mu_i+\lambda}\right)^{\beta-\gamma}\left(\frac{\mu_i+\lambda}{\lambda}\right)^{\beta-\gamma}\\ &= \lambda^{\beta-\gamma}\mu_i^{-\beta}\left(\frac{\mu_i}{\mu_i+\lambda}\right)^{\beta-\gamma}\left(\frac{\lambda}{\lambda+\mu_i}\right)^{2-\beta+\gamma}\\ &\le \lambda^{\beta-\gamma}\mu_i^{-\beta},\end{aligned}$$

where we used $\beta-\gamma\ge 0$ and $2-\beta+\gamma\ge 0$. Hence,

$$\begin{aligned}\|[C_\lambda]-C_*\|^2_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} &\le \lambda^{\beta-\gamma}\sum_{i\in I}\sum_{j\in J}\check{a}^2_{ij}\mu_i^{-\beta}\\ &= \lambda^{\beta-\gamma}\|C_*\|^2_{S_2\left([\mathcal{H}]^\beta_X,\mathcal{Y}\right)}\end{aligned}$$

■

A.2 Bounding the Variance

The proof will require several lemmas in its construction, which we now present. We start with a lemma that allows to go from the γ-norm of embedded vector-valued maps to their norm in the original Hilbert-Schmidt space.

Lemma 2 *For $0\le\gamma\le 1$ and $F\in\mathcal{G}$ the inequality*

$$\|[F]\|_\gamma \le \left\|CC_{XX}^{\frac{1-\gamma}{2}}\right\|_{S_2(\mathcal{H}_X,\mathcal{Y})} \tag{16}$$

holds, where $C=\bar{\Psi}^{-1}(F)\in S_2(\mathcal{H}_X,\mathcal{Y})$. If, in addition, $\gamma<1$ or $C\perp\mathcal{Y}\otimes\ker I_\pi$ is satisfied, then the result is an equality.

Proof Let us fix $F\in\mathcal{G}$, and define $C:=\bar{\Psi}^{-1}(F)\in S_2(\mathcal{H}_X,\mathcal{Y})$. Since $\left(\mu_i^{1/2}e_i\right)_{i\in I}$ is an ONB of $(\ker I_\pi)^\perp$, we can complete it with a basis $(\bar{e}_i)_{i\in I'}$ (with $I\cap I'=\varnothing$) of $\ker I_\pi$ such that the union of the family forms a basis of $\mathcal{H}_X$. Let $(d_j)_{j\in J}$ be a basis of $\mathcal{Y}$, we get a basis of $S_2(\mathcal{H}_X,\mathcal{Y})$ through $(d_j\otimes f_i)_{i\in I\cup I',j\in J}$ where $f_i=\mu_i^{1/2}e_i$ if $i\in I$ and $f_i=\bar{e}_i$ if $i\in I'$. Then C admits the decomposition

$$C=\sum_{i\in I}\sum_{j\in J}a_{ij}d_j\otimes\mu_i^{1/2}e_i+\sum_{i\in I'}\sum_{j\in J}a_{ij}d_j\otimes\bar{e}_i,$$

where $a_{ij}=\langle C,d_j\otimes f_i\rangle_{S_2(\mathcal{H}_X,\mathcal{Y})}=\langle Cf_i,d_j\rangle_{\mathcal{Y}}$ for all $i\in I\cup I',j\in J$. Since

$$[C]=\sum_{i\in I}\sum_{j\in J}a_{ij}d_j\otimes\mu_i^{1/2}[e_i],$$

with Parseval's identity w.r.t. the ONB $\left(d_j \otimes \mu_i^{\gamma/2}[e_i]\right)_{i\in I, j\in J}$ of $S_2([\mathcal{H}]_X^\gamma, \mathcal{Y})$ this yields

$$\|[C]\|^2_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})} = \left\|\sum_{i\in I}\sum_{j\in J} a_{ij}\mu_i^{\frac{1-\gamma}{2}} d_j \otimes \mu_i^{\gamma/2}[e_i]\right\|^2_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})} = \sum_{i\in I}\sum_{j\in J} a_{ij}^2\mu_i^{1-\gamma}.$$

For $\gamma < 1$, the spectral decomposition of C_{XX} Eq. (9) together with the fact that $\left(d_j \otimes \mu_i^{1/2} e_i\right)_{i\in I, j\in J}$ is an ONS in $S_2(\mathcal{H}_X, \mathcal{Y})$ yields

$$\begin{aligned}
\left\|C C_{XX}^{\frac{1-\gamma}{2}}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} &= \left\|C \sum_{l\in I} \mu_l^{\frac{1-\gamma}{2}} \langle \cdot, \mu_l^{\frac{1}{2}} e_l\rangle_{\mathcal{H}_X} \mu_l^{\frac{1}{2}} e_l\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&= \sum_{i\in I}\left\|\sum_{l\in I} \mu_l^{\frac{1-\gamma}{2}} \langle \mu_i^{\frac{1}{2}} e_i, \mu_l^{\frac{1}{2}} e_l\rangle_{\mathcal{H}_X} \mu_l^{\frac{1}{2}} C e_l\right\|^2_{\mathcal{Y}} + \sum_{i\in I'}\left\|\sum_{l\in I} \mu_l^{\frac{1-\gamma}{2}} \langle \bar{e}_i, \mu_l^{\frac{1}{2}} e_l\rangle_{\mathcal{H}_X} \mu_l^{\frac{1}{2}} C e_l\right\|^2_{\mathcal{Y}} \\
&= \sum_{i\in I}\left\|\mu_i^{\frac{1-\gamma}{2}} \mu_i^{\frac{1}{2}} C e_i\right\|^2_{\mathcal{Y}} \\
&= \sum_{i\in I}\sum_{j\in J} \mu_i^{1-\gamma}\left\langle C\left(\mu_i^{\frac{1}{2}} e_i\right), d_j\right\rangle^2_{\mathcal{Y}} \\
&= \sum_{i\in I}\sum_{j\in J} a_{ij}^2 \mu_i^{1-\gamma}.
\end{aligned}$$

This proves the claimed equality in the case of $\gamma < 1$. For $\gamma = 1$, we have $C_{XX}^{\frac{1-\gamma}{2}} = \mathrm{Id}_{\mathcal{H}_X}$ and the Pythagorean theorem together with Parseval's identity yields

$$\begin{aligned}
\left\|C C_{XX}^{\frac{1-\gamma}{2}}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} &= \left\|\sum_{i\in I}\sum_{j\in J} a_{ij} d_j \otimes \mu_i^{1/2} e_i + \sum_{i\in I'}\sum_{j\in J} a_{ij} d_j \otimes \bar{e}_i\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&= \left\|\sum_{i\in I}\sum_{j\in J} a_{ij} d_j \otimes \mu_i^{1/2} e_i\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} + \left\|\sum_{i\in I'}\sum_{j\in J} a_{ij} d_j \otimes \bar{e}_i\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&= \sum_{i\in I}\sum_{j\in J} a_{ij}^2 + \left\|\sum_{i\in I'}\sum_{j\in J} a_{ij} d_j \otimes \bar{e}_i\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})}
\end{aligned}$$

This gives the claimed equality if $C \perp \mathcal{Y} \otimes \ker I_\pi$, as well as the claimed inequality for general $C \in S_2(\mathcal{H}_X, \mathcal{Y})$. We conclude with $\|[F]\|_\gamma = \|[C]\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}$ by definition. ■

Lemma 3 *If $F_* \in [\mathcal{G}]^\beta$ is satisfied for some $0 \le \beta \le 2$, then the following bounds is satisfied, for all $\lambda > 0$ and $\gamma \ge 0$:*

$$\|[F_\lambda]\|^2_\gamma \le \|F_*\|^2_{\min\{\gamma,\beta\}} \lambda^{-(\gamma-\beta)_+}. \tag{17}$$

Proof By Parseval's identity

$$\|[F_\lambda]\|^2_\gamma = \sum_{i\in I}\sum_{j\in J}\left(\frac{\mu_i}{\mu_i + \lambda}\right)^2 \mu_i^{-\gamma} \check{a}_{ij}^2.$$

where $\check{a}_{ij} = \langle C_*[e_i], d_j \rangle_{\mathcal{Y}}$ for all $i \in I, j \in J$ as in the proof of Lemma 1. In the case of $\gamma \leq \beta$ we bound the fraction by 1 and then Parseval's identity gives us

$$\|[F_\lambda]\|_\gamma^2 \leq \sum_{i \in I} \sum_{j \in J} \mu_i^{-\gamma} \check{a}_{ij}^2 = \|F_*\|_\gamma^2 .$$

In the case of $\gamma > \beta$,

$$\|[F_\lambda]\|_\gamma^2 = \sum_{i \in I} \sum_{j \in J} \left(\frac{\mu_i^{1-\frac{\gamma-\beta}{2}}}{\mu_i + \lambda} \right)^2 \mu_i^{-\beta} \check{a}_{ij}^2 \leq \lambda^{-(\gamma-\beta)} \sum_{i \in I} \sum_{j \in J} \mu_i^{-\beta} \check{a}_{ij}^2 = \lambda^{-(\gamma-\beta)} \|F_*\|_\beta^2 ,$$

where we used Parseval's identity in the equality and Lemma 25 from Fischer and Steinwart (2020). ■

By (EMB), the inclusion map $I_\pi^{\alpha,\infty} : [\mathcal{H}]_X^\alpha \hookrightarrow L_\infty(\pi)$ has bounded norm $A > 0$ i.e. for $f \in [\mathcal{H}]_X^\alpha$, f is $\pi-$a.e. bounded and $\|f\|_\infty \leq A \|f\|_\alpha$. We now show that (EMB) automatically implies that the inclusion operator for $[\mathcal{G}]^\alpha$ is bounded.

Lemma 4 *Under* (EMB) *the inclusion operator* $\mathcal{I}_\pi^{\alpha,\infty} : [\mathcal{G}]^\alpha \hookrightarrow L_\infty(\pi; \mathcal{Y})$ *is bounded with operator norm less than or equal to* A.

$L_\infty(\pi; \mathcal{Y})$ denotes the space of $\mathcal{F}_{E_X} - \mathcal{F}_{\mathcal{Y}}$ measurable $\mathcal{Y}$-valued functions (gathered by π-equivalent classes) that are essentially bounded with respect to π. $L_\infty(\pi; \mathcal{Y})$ is endowed with the norm $\|F\|_\infty := \inf\{c \geq 0 : \|F(x)\|_{\mathcal{Y}} \leq c$ for π-almost all $x \in E_X\}$.

Proof For every $F \in [\mathcal{G}]^\alpha$, there is a sequence $b_{ij} \in \ell_2(I \times J)$ such that for $\pi-$almost all $x \in E_X$,

$$F(x) = \sum_{i \in I, j \in J} b_{ij} d_j \mu_i^{\alpha/2} [e_i](x)$$

where $(d_j)_{j \in J}$ is any orthonormal basis of $\mathcal{Y}$ and $\|F\|_\alpha^2 = \sum_{i \in I, j \in J} b_{ij}^2$. We consider $F \in [\mathcal{G}]^\alpha$ such that $\sum_{i \in I, j \in J} b_{ij}^2 \leq 1$. For $\pi-$almost all $x \in E_X$,

$$\begin{aligned} \|F(x)\|_{\mathcal{Y}}^2 &= \left\| \sum_{j \in J} \left(\sum_{i \in I} b_{ij} \mu_i^{\alpha/2} [e_i](x) \right) d_j \right\|_{\mathcal{Y}}^2 \\ &= \sum_{j \in J} \left(\sum_{i \in I} b_{ij} \mu_i^{\alpha/2} [e_i](x) \right)^2 \\ &\leq \sum_{j \in J} \left(\sum_{i \in I} b_{ij}^2 \right) \left(\sum_{i \in I} \mu_i^\alpha [e_i]^2(x) \right) \\ &\leq A^2 \sum_{j \in J} \sum_{i \in I} b_{ij}^2 \\ &\leq A^2 \end{aligned}$$

where we used the Cauchy-Schwarz inequality for each $j \in J$ for the first inequality and a consequence of (EMB) in the second inequality (see Theorem 9 in Fischer and Steinwart, 2020). We therefore conclude $\|\mathcal{I}_\pi^{\alpha,\infty}\| \leq A$. ■

Combining Lemmas 1, 3 and 4 we have the following result.

Lemma 5 *If $F_* \in [\mathcal{G}]^\beta$ and* (EMB) *are satisfied for some $0 \leq \beta \leq 2$ and $0 < \alpha \leq 1$, then the following bounds are satisfied, for all $0 < \lambda \leq 1$:*

$$\|[F_\lambda] - F_*\|_{L_\infty}^2 \leq \left(\|F_*\|_{L_\infty} + A\|F_*\|_\beta\right)^2 \lambda^{\beta-\alpha}, \tag{18}$$

$$\|[F_\lambda]\|_{L_\infty}^2 \leq A^2 \|F_*\|_{\min\{\alpha,\beta\}}^2 \lambda^{-(\alpha-\beta)_+}. \tag{19}$$

Proof For Eq. (19), we use Lemma 4 and Eq. (17) in Lemma 3.

$$\|[F_\lambda]\|_\infty^2 \leq A^2 \|[F_\lambda]\|_\alpha^2 \leq A^2 \|F_*\|_{\min\{\alpha,\beta\}}^2 \lambda^{-(\alpha-\beta)_+}$$

To show Eq. (18), in the case $\beta \leq \alpha$ we use the triangle inequality, Eq. (19) and $\lambda \leq 1$ to obtain

$$\begin{aligned}\|[F_\lambda] - F_*\|_\infty &\leq \|F_*\|_\infty + \|[F_\lambda]\|_\infty \\ &\leq \left(\|F_*\|_\infty + A\|F_*\|_\beta\right)\lambda^{-\frac{\alpha-\beta}{2}}\end{aligned}$$

In the case $\beta > \alpha$, Eq. (18) is a consequence of Lemma 4 and Eq. (15) in Lemma 1 with $\gamma = \alpha$,

$$\|[F_\lambda] - F_*\|_\infty^2 \leq A^2 \|[F_\lambda] - F_*\|_\alpha^2 \leq A^2 \|F_*\|_\beta^2 \lambda^{\beta-\alpha} \leq \left(\|F_*\|_\infty + A\|F_*\|_\beta\right)^2 \lambda^{\beta-\alpha}.$$

■

Theorem 8 *Let $\mathcal{H}_X$ be a RKHS on E_X with respect to a kernel k_X such that assumptions 1 to 3 hold. Let P be a probability distribution on $E_X \times \mathcal{Y}$ with $\pi := P_{E_X}$ (the marginal distribution on E_X). Furthermore, let $\|F_*\|_\infty < \infty$,* (EMB) *and* (MOM) *be satisfied. We define*

$$\begin{aligned}M(\lambda) &= \|[F_\lambda] - F_*\|_\infty, \\ \mathcal{N}(\lambda) &= \operatorname{tr}\left(C_{XX}(C_{XX} + \lambda \mathrm{Id}_{\mathcal{H}_X})^{-1}\right), \\ Q_\lambda &= \max\{M(\lambda), R\}, \\ g_\lambda &= \log\left(2e\mathcal{N}(\lambda)\frac{\|C_{XX}\| + \lambda}{\|C_{XX}\|}\right).\end{aligned}$$

Then, for $0 \leq \gamma \leq 1$, $\tau \geq 1$, $\lambda > 0$ and $n \geq 8A^2\tau g_\lambda \lambda^{-\alpha}$, with probability $1 - 4e^{-\tau}$:

$$\left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}^2 \leq \frac{576\tau^2}{n\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|_{L_2(\pi;\mathcal{Y})}^2 A^2}{\lambda^\alpha} + \frac{2Q_\lambda^2 A^2}{n\lambda^\alpha}\right)$$

Proof We first decompose the variance term as

$$\begin{aligned}
&\left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} \\
&= \left\|\left[\hat{C}_{YX}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1} - C_{YX}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\right]\right\|_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} \\
&\le \left\|\left(\hat{C}_{YX}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1} - C_{YX}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\right) C_{XX}^{\frac{1-\gamma}{2}}\right\|_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&\le \left\|\left(\hat{C}_{YX} - C_{YX}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}}\right\|_{S_2(\mathcal{H}_X,\mathcal{Y})} \quad (20) \\
&\cdot \left\|\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{\frac{1}{2}}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{\frac{1}{2}}\right\|_{\mathcal{H}_X\to\mathcal{H}_X} \quad (21) \\
&\cdot \left\|\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}} C_{XX}^{\frac{1-\gamma}{2}}\right\|_{\mathcal{H}_X\to\mathcal{H}_X} \quad (22)
\end{aligned}$$

where we used Lemma 2 in the first inequality. Eq. (21) is bounded as in Lemma 17 and Theorem 16 in Fischer and Steinwart (2020),

$$\left\|\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{\frac{1}{2}}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{\frac{1}{2}}\right\|_{\mathcal{H}_X\to\mathcal{H}_X} \le 3$$

for $n \ge 8A^2\tau g_\lambda \lambda^{-\alpha}$ with probability $1-2e^{-\tau}$ for all $\tau \ge 1$. For Eq. (22) we have, using Lemma 25 from Fischer and Steinwart (2020)

$$\left\|\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}} C_{XX}^{\frac{1-\gamma}{2}}\right\|_{\mathcal{H}_X\to\mathcal{H}_X} \le \sqrt{\sup_i \frac{\mu_i^{1-\gamma}}{\mu_i+\lambda}} \le \lambda^{-\frac{\gamma}{2}}.$$

Finally for the bound of Eq. (20) Lemma 6 show that for $\tau \ge 1$, $\lambda > 0$ and $n \ge 1$ with probability $1-2e^{-\tau}$:

$$\begin{aligned}
&\left\|\left(\hat{C}_{YX} - C_{YX}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&\le \frac{64\tau^2}{n}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q_\lambda^2 A^2}{n\lambda^\alpha}\right).
\end{aligned}$$

■

Lemma 6 *Assume the conditions in Theorem 8 hold. Then for $\tau \ge 1$, $\lambda > 0$ and $n \ge 1$ with probability $1-2e^{-\tau}$:*

$$\begin{aligned}
&\left\|\left(\hat{C}_{YX} - C_{YX}\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX}+\lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \\
&\le \frac{64\tau^2}{n}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q_\lambda^2 A^2}{n\lambda^\alpha}\right).
\end{aligned}$$

Proof We begin with the decomposition

$$\begin{aligned}
&\hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right) \\
&= \hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X} + \hat{C}_{XX} - C_{XX}\right) \\
&= \hat{C}_{YX} - C_{YX} + C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\left(C_{XX} - \hat{C}_{XX}\right) \\
&= \hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\hat{C}_{XX} - \left(C_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}C_{XX}\right) \\
&= \hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\hat{\mathbb{E}}[\phi_X(X)\otimes\phi_X(X)] - \left(C_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\mathbb{E}[\phi_X(X)\otimes\phi_X(X)]\right) \\
&= \hat{\mathbb{E}}\left[(Y - F_\lambda(X))\otimes\phi_X(X)\right] - \mathbb{E}\left[(Y - F_\lambda(X))\otimes\phi_X(X)\right]
\end{aligned}$$

where we denote $\hat{\mathbb{E}}[\phi_X(X)\otimes\phi_X(X)] = \frac{1}{n}\sum_{i=1}^n \phi_X(x_i)\otimes\phi_X(x_i)$. We wish to apply Theorem 13 with $H = S_2(\mathcal{H}_X, \mathcal{Y})$. Consider the random variables $\xi_0, \xi_2 : E_X \times \mathcal{Y} \to \mathcal{Y}\otimes\mathcal{H}_X$ defined by

$$\xi_0(x,y) := (y - F_\lambda(x))\otimes\phi_X(x), \tag{23}$$

$$\xi_2(x,y) := \xi_0(x,y)\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1/2}. \tag{24}$$

Since our kernel k_X is bounded,

$$\begin{aligned}
\|\xi_0(x,y)\|_{S_2(\mathcal{H}_X,\mathcal{Y})} &= \|y - F_\lambda(x)\|_{\mathcal{Y}}\,\|\phi_X(x)\|_{\mathcal{H}_X} \\
&\le \|y - F_\lambda(x)\|_{\mathcal{Y}}\,\kappa_X \\
&\le \left(\|y\|_{\mathcal{Y}} + \|F_\lambda\|_{L_\infty(\pi;\mathcal{Y})}\right)\kappa_X,
\end{aligned}$$

is satisfied for $\pi-$almost all $x \in E_X$ and F_λ is π-almost surely bounded by Lemma 5. Since $y \in L_2(\pi;\mathcal{Y})$, we have that $y \in L_1(\pi;\mathcal{Y})$. This yields

$$\frac{1}{n}\sum_{i=1}^n\left(\xi_2(x_i,y_i) - \mathbb{E}\xi_2\right) = \hat{\mathbb{E}}\xi_2 - \mathbb{E}\xi_2 = \left(\hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}}, \tag{25}$$

and therefore Eq. (20) coincides with the left hand side of a Bernstein's inequality for H-valued random variables (Theorem 13). Consequently, it remains to bound the m-th moment of ξ_2, for $m \ge 2$,

$$\mathbb{E}\|\xi_2\|^m_{S_2(\mathcal{H}_X,\mathcal{Y})} = \int_{E_X}\left\|\left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1/2}\phi_X(x)\right\|^m_{\mathcal{H}_X}\int_{\mathcal{Y}}\|y - F_\lambda(x)\|^m_{\mathcal{Y}}\,p(x,\,\mathrm{d}y)\mathrm{d}\pi(x). \tag{26}$$

First, we consider the inner integral. Using the triangle inequality and (MOM),

$$\begin{aligned}
\int_{\mathcal{Y}}\|y - F_\lambda(x)\|^m_{\mathcal{Y}}\,p(x,\,\mathrm{d}y) &\le 2^{m-1}\left(\|\operatorname{Id}_{\mathcal{Y}} - F_*(x)\|^m_{L_m(p(x,\cdot);\mathcal{Y})} + \|F_*(x) - F_\lambda(x)\|^m_{\mathcal{Y}}\right) \\
&\le \frac{1}{2}m!(2R)^{m-2}2\sigma^2 + 2^{m-1}\|F_*(x) - F_\lambda(x)\|^m_{\mathcal{Y}}.
\end{aligned} \tag{27}$$

for π-almost all $x \in E_X$. If we plug this bound into the outer integral and use the abbreviation $h_x := \left(C_{XX} + \lambda \operatorname{Id}_{\mathcal{H}_X}\right)^{-1/2}\phi_X(x)$ we get

$$\mathbb{E}\|\xi_2\|^m_{S_2(\mathcal{H}_X,\mathcal{Y})} \le \frac{1}{2}m!(2R)^{m-2}2\sigma^2\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\,\mathrm{d}\pi(x) + 2^{m-1}\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\|F_*(x) - F_\lambda(x)\|^m_{\mathcal{Y}}\,\mathrm{d}\pi(x). \tag{28}$$

Using Lemma 10, we can bound the first term in Eq. (28) above by

$$\begin{aligned}\frac{1}{2}m!(2R)^{m-2}2\sigma^2\int_{E_X}\|h_x\|_{\mathcal{H}_X}^m\ \mathrm{d}\pi(x) &\le \frac{1}{2}m!(2R)^{m-2}2\sigma^2\left(\frac{A}{\lambda^{\alpha/2}}\right)^{m-2}\int_{E_X}\|h_x\|_{\mathcal{H}_X}^2\ \mathrm{d}\pi(x)\\ &= \frac{1}{2}m!\left(\frac{2RA}{\lambda^{\alpha/2}}\right)^{m-2}2\sigma^2\mathcal{N}(\lambda)\\ &\le \frac{1}{2}m!\left(\frac{2Q_\lambda A}{\lambda^{\alpha/2}}\right)^{m-2}2\sigma^2\mathcal{N}(\lambda)\end{aligned}$$

where we only used $R \le Q_\lambda$ in the last step. Again, using Lemma 10, the second term in Eq. (28) can be bounded by

$$\begin{aligned}&2^{m-1}\int_{E_X}\|h_x\|_{\mathcal{H}_X}^m\|F_*(x)-F_\lambda(x)\|_{\mathcal{Y}}^m\ \mathrm{d}\pi(x)\\ \le&\frac{1}{2}\left(\frac{2A}{\lambda^{\alpha/2}}\right)^m M(\lambda)^{m-2}\int_{E_X}\|F_*(x)-F_\lambda(x)\|_{\mathcal{Y}}^2\ \mathrm{d}\pi(x)\\ =&\frac{1}{2}\left(\frac{2AM(\lambda)}{\lambda^{\alpha/2}}\right)^{m-2}\|F_*-[F_\lambda]\|_{L_2(\pi;\mathcal{Y})}^2\frac{4A^2}{\lambda^\alpha}\\ \le&\frac{1}{2}m!\left(\frac{2Q_\lambda A}{\lambda^{\alpha/2}}\right)^{m-2}\|F_*-[F_\lambda]\|_{L_2(\pi;\mathcal{Y})}^2\frac{2A^2}{\lambda^\alpha},\end{aligned}$$

where we only used $M(\lambda) \le Q_\lambda$ and $2 \le m!$ in the last step. Finally, we get

$$\mathbb{E}\|\xi_2\|_{S_2(\mathcal{H}_X,\mathcal{Y})}^m \le \frac{1}{2}m!\left(\frac{2Q_\lambda A}{\lambda^{\alpha/2}}\right)^{m-2}2\left(\sigma^2\mathcal{N}(\lambda)+\|F_*-[F_\lambda]\|_{L_2(\pi;\mathcal{Y})}^2\frac{A^2}{\lambda^\alpha}\right) \tag{29}$$

and an application of Bernstein's inequality from Theorem 13 with $L = 2Q_\lambda A\lambda^{-\alpha/2}$ and $\sigma^2 = 2\left(\sigma^2\mathcal{N}(\lambda)+\|F_*-[F_\lambda]\|_{L_2(\pi;\mathcal{Y})}^2 A^2\lambda^{-\alpha}\right)$ yields the bound. Putting all the terms together, we obtain our result. ■

A.3 Learning Rates - Proof of Theorem 7

In this section, we aim to establish our upper bound on the learning rate for vector-valued regression by combining the learning rates obtained for the bias and variance.

Let us fix some $\tau \ge 1$ and a lower bound $0 < c \le 1$ with $c \le \|C_{XX}\|$. We first show that Theorem 8 is applicable. To this end, we prove that there is an index bound $n_0 \ge 1$ such that $n \ge 8A^2\tau g_{\lambda_n}\lambda_n^{-\alpha}$ is satisfied for all $n \ge n_0$. Since $\lambda_n \to 0$ we choose $n_0' \ge 1$ such that $\lambda_n \le c \le \min\{1, \|C_{XX}\|\}$ for all $n \ge n_0'$. We get for $n \ge n_0'$,

$$\begin{aligned}\frac{8A^2\tau g_{\lambda_n}\lambda_n^{-\alpha}}{n} &= \frac{8A^2\tau\lambda_n^{-\alpha}}{n}\cdot\log\left(2e\mathcal{N}(\lambda_n)\frac{\|C_{XX}\|+\lambda_n}{\|C_{XX}\|}\right)\\ &\le \frac{8A^2\tau\lambda_n^{-\alpha}}{n}\cdot\log\left(4De\lambda_n^{-p}\right)\\ &= 8A^2\tau\left(\frac{\log(4De)\lambda_n^{-\alpha}}{n}+\frac{p\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n}\right)\end{aligned}$$

where the second step uses Lemma 10. Hence, it is enough to show $\frac{\log(\lambda_n^{-1})}{n\lambda_n^{\alpha}} \to 0$. We consider the cases $\beta + p \leq \alpha$ and $\beta + p > \alpha$.

- $\beta + p \leq \alpha$. By substituting that $\lambda_n = \Theta\left(\left(\frac{n}{\log^{\theta} n}\right)^{-\frac{1}{\alpha}}\right)$ for some $\theta > 1$ we have
$$\frac{\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n} = \Theta\left(\frac{\log(n)}{n}\frac{n}{\log^{\theta}(n)}\right) = \Theta\left(\frac{1}{\log^{\theta-1}(n)}\right) \to 0, \text{ as } \ n \to \infty.$$
- $\beta + p > \alpha$. By substituting that $\lambda_n = \Theta\left(n^{-\frac{1}{\beta+p}}\right)$ and using $1 - \frac{\alpha}{\beta+p} > 0$ we have
$$\frac{\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n} = \Theta\left(\frac{\log(n)}{n}n^{\frac{\alpha}{\beta+p}}\right) = \Theta\left(\frac{\log(n)}{n^{1-\frac{\alpha}{\beta+p}}}\right) \to 0, \text{ as } \ n \to \infty.$$

Consequently, there is a $n_0 \geq n_0'$ with $n \geq 8A^2 \log \tau g_{\lambda_n}\lambda_n^{-\alpha}$ for all $n \geq n_0$. Moreover, n_0 just depends on λ_n, c, D, τ, A, and on the parameters α, p.

Let $n \geq n_0$ be fixed. By Theorem 8, we have
$$\left\|[\hat{C}_{\lambda_n} - C_{\lambda_n}]\right\|^2_{S_2([\mathcal{H}]^{\gamma}_X,\mathcal{Y})} \leq \frac{576\tau^2}{n\lambda_n^{\gamma}}\left(\sigma^2\mathcal{N}(\lambda_n) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}A^2}{\lambda_n^{\alpha}} + \frac{2Q^2_{\lambda_n}A^2}{n\lambda_n^{\alpha}}\right).$$

Using Lemma 10 and Lemma 3 with $\gamma = 0$, we have
$$\left\|[\hat{C}_{\lambda_n} - C_{\lambda_n}]\right\|^2_{S_2([\mathcal{H}]^{\gamma}_X,\mathcal{Y})} \leq \frac{576\tau^2}{n\lambda_n^{\gamma}}\left(\sigma^2 D\lambda_n^{-p} + A^2\|F_*\|^2_{\beta}\lambda_n^{\beta-\alpha} + \frac{2Q^2_{\lambda_n}A^2}{n\lambda_n^{\alpha}}\right)$$

For the last term, using the definition of Q_λ in Theorem 8 with Lemma 5 and $\lambda_n \leq 1$, we get
$$\begin{aligned} Q^2_{\lambda_n} &= \max\{R^2, \|[F_\lambda] - F_*\|^2_{\infty}\} \\ &\leq \max\left\{R^2, \left(\|F_*\|_{\infty} + A\|F_*\|_{\beta}\right)^2\lambda_n^{-(\alpha-\beta)}\right\} \\ &\leq K_0\lambda_n^{-(\alpha-\beta)_+}, \end{aligned}$$
where $K_0 := \max\left\{R^2, \left(B_\infty + A\|F_*\|_{\beta}\right)^2\right\}$. Thus,
$$\left\|[\hat{C}_{\lambda_n} - C_{\lambda_n}]\right\|^2_{S_2([\mathcal{H}]^{\gamma}_X,\mathcal{Y})} \leq \frac{576\tau^2}{n\lambda_n^{\gamma}}\left(\sigma^2 D\lambda_n^{-p} + A^2\|F_*\|^2_{\beta}\lambda_n^{\beta-\alpha} + 2A^2K_0\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+}}\right). \quad (30)$$

For the first and second terms in the bracket, we use again the fact that $\lambda_n \leq 1$, and get
$$D\sigma^2\lambda_n^{-p} + A^2\|F_*\|^2_{\beta}\lambda_n^{-(\alpha-\beta)} \leq \left(D\sigma^2 + A^2\|F_*\|^2_{\beta}\right)\max\{\lambda_n^{-p}, \lambda_n^{-(\alpha-\beta)}\} \leq K_1\lambda_n^{-\max\{p,\alpha-\beta\}}$$
with $K_1 := D\sigma^2 + A^2\|F_*\|^2_{\beta}$. We now have
$$\begin{aligned} \left\|[\hat{C}_{\lambda_n} - C_{\lambda_n}]\right\|^2_{S_2([\mathcal{H}]^{\gamma}_X,\mathcal{Y})} &\leq \frac{576\tau^2}{n\lambda_n^{\gamma}}\left(K_1\lambda_n^{-\max\{p,\alpha-\beta\}} + 2A^2K_0\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+}}\right) \\ &= \frac{576\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}}\left(K_1 + 2A^2K_0\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}\right). \end{aligned}$$

Again, we treat the cases $\beta + p \leq \alpha$ and $\beta + p > \alpha$ separately.

- $\beta + p \le \alpha$. In this case we have

  $$\alpha + (\alpha - \beta)_+ - \max\{p, \alpha - \beta\} = \alpha.$$

  Since $\lambda_n = \Theta\left(\left(\frac{n}{\log^\theta n}\right)^{-\frac{1}{\alpha}}\right)$, for some $\theta > 1$ we therefore have

  $$\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}} = \frac{1}{n\lambda^\alpha} = \Theta\left(\frac{1}{\log^\theta n}\right).$$

- $\beta + p > \alpha$. We have $p > \alpha - \beta$ and $\lambda_n = \Theta\left(n^{-\frac{1}{\beta+p}}\right)$, and hence

  $$\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}} = \frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-p}} = \Theta\left(\left(\frac{1}{n}\right)^{1-\frac{\alpha+(\alpha-\beta)_+-p}{\beta+p}}\right).$$

  Using $p > \alpha - \beta$ again gives us

  $$1 - \frac{\alpha + (\alpha-\beta)_+ - p}{\beta + p} = \frac{2p - (\alpha-\beta)_+ - (\alpha - \beta)}{\beta + p} > 0.$$

As such, there is a constant $K_2 > 0$ with

$$\left\|[\hat{F}_{\lambda_n} - F_{\lambda_n}]\right\|_\gamma^2 = \left\|[\hat{C}_{\lambda_n} - C_{\lambda_n}]\right\|_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})}^2 \le 576\frac{\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}}\left(K_1 + 2A^2K_0K_2\right)$$

for all $n \ge n_0$. Defining $K_3 := 576(K_1 + 2A^2K_0K_2)$, and using the bias-variance splitting from Eq. (14) and Lemma 1, we have

$$\begin{aligned}\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 &\le 2\|C_*\|_{S_2([\mathcal{H}]_X^\beta,\mathcal{Y})}^2\lambda_n^{\beta-\gamma} + 2K_3\frac{\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}} \\ &\le \tau^2\lambda_n^{\beta-\gamma}\left(2\|C_*\|_{S_2([\mathcal{H}]_X^\beta,\mathcal{Y})}^2 + 2K_3\frac{1}{n\lambda_n^{\max\{\beta+p,\alpha\}}}\right),\end{aligned}$$

where we used $\tau \ge 1$. Since in both cases $\beta + p \le \alpha$ and $\beta + p > \alpha$, $\lambda_n \succcurlyeq n^{-1/\max\{\alpha,\beta+p\}}$ there is some constant $J > 0$ such that

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 \le \tau^2 J\lambda_n^{\beta-\gamma}$$

for all $n \ge n_0$.

Appendix B. Proof of Theorem 3

In this section, we prove our main results, Theorem 3. The case where $\beta \ge \alpha$ is identical to Theorem 7. Hence we will only focus on the rate when $\beta < \alpha$ without assuming the boundedness of F_*. We adopt the same risk decomposition as in Eq. (14). The bias is upper bounded in the same way as in Section A.1. For the variance, we have the following results.

Theorem 9 *Let $\mathcal{H}_X$ be a RKHS on E_X with respect to a kernel k_X such that assumptions 1 to 3 hold. Let P be a probability distribution on $E_X \times \mathcal{Y}$ with $\pi := P_{E_X}$ (the marginal distribution on E_X). Furthermore, let* (EMB), (SRC) *and* (MOM) *be satisfied. Suppose that $F_* \in L_q(\pi;\mathcal{Y})$ and $\|F_*\|_{L_q} \leq C_q < \infty$ for some $q \geq 2$. Denote $\Omega_0 = \{x \in E_X : \|F_*(x)\|_{\mathcal{Y}} \leq t\}$ and*

$$Q(t,\lambda) := \max\{t + 2A\,\|F_*\|_\beta\,\lambda^{\frac{\beta-\alpha}{2}}, R\}.$$

Then, for $0 \leq \gamma \leq 1$, $\tau \geq 1$, $\lambda > 0$ and $n \geq 8A^2\tau g_\lambda \lambda^{-\alpha}$, with probability $1 - 4e^{-\tau} - \tau_n$ with $\tau_n = 1 - \left(1 - \frac{C_q^q}{t^q}\right)^n$:

$$\left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|^2_{S_2([\mathcal{H}]_X^\gamma,\mathcal{Y})} \leq \frac{1152\tau^2}{n\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2A^2}{n\lambda^\alpha}\right) + \frac{18}{\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right)\frac{C_q^q}{t^q},$$

where g_λ is defined in Theorem 8.

Proof The proof adopts the same strategy as in Theorem 8. The difference is that for Eq. (20), instead of using Lemma 6, we use Lemma 7 below. ■

Lemma 6 provides the upper bound under the assumption that $\|F_*\|_{L_\infty(\pi;\mathcal{Y})} < \infty$. For $\beta \geq \alpha$ this assumption is automatically satisfied by the condition (EMB), but for $\beta < \alpha$, this assumption remains crucial. However, as indicated in Zhang et al. (2023b), in the scalar-valued setting, when $\alpha - p < \beta < \alpha$, this assumption can be replaced by an assumption of the form $\|F_*\|_{L_q(\pi;\mathcal{Y})} < \infty$ for some $q \geq 2$. Below, we generalise their technique to the vector-valued setting.

Lemma 7 *Assume the conditions in Theorem 9 holds. Then for $\tau \geq 1$, $\lambda > 0$ and $n \geq 1$ with probability over $1 - 2e^{-\tau} - \tau_n$ with $\tau_n = 1 - \left(1 - \frac{C_q^q}{t^q}\right)^n$,*

$$\left\|\left(\hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})}$$

$$\leq \frac{384\tau^2}{n}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2A^2}{n\lambda^\alpha}\right) + 6\left(\sigma^2\mathcal{N}(\lambda) + \|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right)\frac{C_q^q}{t^q}. \tag{31}$$

Proof Denote $\Omega_0 := \{x \in E_X : \|F_*(x)\|_{\mathcal{Y}} \leq t\}$ and $\Omega_1 = E_X \backslash \Omega_0$. Since $\|F_*\|_{L_q} \leq C_q$, we have

$$\mathbb{P}(X \in \Omega_1) = \mathbb{P}(\|F_*(X)\|_{\mathcal{Y}} > t) \leq \frac{\mathbb{E}\left[\|F_*(X)\|^q_{\mathcal{Y}}\right]}{t^q} \leq \frac{C_q^q}{t^q}. \tag{32}$$

Let $\xi_2(x,y)$ be defined as in Eq. (24). As shown before in Eq. (25), we have

$$\left(\hat{C}_{YX} - C_{YX}\left(C_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\left(\hat{C}_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)\right)\left(C_{XX} + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-\frac{1}{2}} = \frac{1}{n}\sum_{i=1}^n\left(\xi_2\left(x_i,y_i\right) - \mathbb{E}\xi_2\right).$$

Decomposing $\xi_2 = \xi_2 \mathbb{1}_{x\in\Omega_0} + \xi_2 \mathbb{1}_{x\in\Omega_1}$, we have

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^n \xi_2\left(x_i, y_i\right) - \mathbb{E}\xi_2 \right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \leq & 3 \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \xi_2\left(x_i, y_i\right)\mathbb{1}_{x_i\in\Omega_0} - \mathbb{E}\left[\xi_2 \mathbb{1}_{X\in\Omega_0}\right] \right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})}}_{I} \\ & + 3 \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \xi_2\left(x_i, y_i\right)\mathbb{1}_{x_i\in\Omega_1} \right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})}}_{II} + 3 \underbrace{\left\|\mathbb{E}\left[\xi_2 \mathbb{1}_{X\in\Omega_1}\right]\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})}}_{III}.\end{aligned}$$

For term I, we employ Proposition 2 and obtain that for any $\tau \geq 1$, with probability over $1 - 2e^{-\tau}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n \left(\xi_2\left(x_i, y_i\right)\mathbb{1}_{x_i\in\Omega_0} - \mathbb{E}\xi_2 \mathbb{1}_{x\in\Omega_0}\right) \right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \leq \frac{64\tau^2}{n}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\left\|F_* - [F_\lambda]\right\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2 A^2}{n\lambda^\alpha}\right) := C_I.$$

For term II, we have

$$\begin{aligned}\tau_n := \mathbb{P}(II > C_I) \leq \mathbb{P}(\exists x_i, \text{ s.t. } x_i \in \Omega_1) &= 1 - \mathbb{P}(x_i \in \Omega_0, \forall i = [n]) \\ &= 1 - \mathbb{P}(X \in \Omega_0)^n \\ &= 1 - \mathbb{P}(\|F_*(X)\|_{\mathcal{Y}} \leq t)^n \\ &\leq 1 - \left(1 - \frac{C_q^q}{t^q}\right)^n.\end{aligned}$$

For term III, we have

$$\begin{aligned}\left\|\mathbb{E}\xi_2 \mathbb{1}_{X\in\Omega_1}\right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} &\leq \mathbb{E}\left[\|\xi_2\|_{S_2(\mathcal{H}_X\mathcal{Y})}\mathbb{1}_{X\in\Omega_1}\right]^2 \\ &\leq \mathbb{E}\left[\|\xi_2\|^2_{S_2(\mathcal{H}_X\mathcal{Y})}\right]\mathbb{P}(X \in \Omega_1) \\ &\leq 2\left(\sigma^2\mathcal{N}(\lambda) + \left\|F_* - [F_\lambda]\right\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right)C_q^q t^{-q}.\end{aligned}$$

where for the last inequality, we used Eq. (29). ■

Proposition 2 *Suppose that conditions* (EMB), (SRC) *and* (MOM) *are satisfied with* $0 \leq \alpha \leq 1$ *and* $\beta \in (0, 2]$. *Let* $\xi_2(x, y)$ *be defined as in Eq.* (24) *and* $\Omega_0 = \{x \in E_X : \|F_*(x)\|_{\mathcal{Y}} \leq t\}$. *We have for any* $\tau \geq 1$, *with probability over* $1 - 2e^{-\tau}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n \left(\xi_2\left(x_i, y_i\right)\mathbb{1}_{x_i\in\Omega_0} - \mathbb{E}\xi_2 \mathbb{1}_{X\in\Omega_0}\right) \right\|^2_{S_2(\mathcal{H}_X,\mathcal{Y})} \leq \frac{64\tau^2}{n}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\left\|F_* - [F_\lambda]\right\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2 A^2}{n\lambda^\alpha}\right).$$

Proof We would like to apply the Bernstein's inequality from Theorem 13 to obtain the desired bound. To this end, we bound the m-th moment of $\xi_2 \mathbb{1}_{X\in\Omega_0}$. Similar to Eq. (26), we have

$$\mathbb{E}\left\|\xi_2 \mathbb{1}_{X\in\Omega_0}\right\|^m_{S_2(\mathcal{H}_X,\mathcal{Y})} = \int_{E_X}\left\|\left(C_{XX} + \lambda \mathrm{Id}_{\mathcal{H}_X}\right)^{-1/2}\phi_X(x)\right\|^m_{\mathcal{H}_X}\mathbb{1}_{X\in\Omega_0}\int_{\mathcal{Y}}\|y - F_\lambda(x)\|^m_{\mathcal{Y}}\, p(x,\, \mathrm{d}y)\mathrm{d}\pi(x).$$

Apply Eq. (27) to the above inner integral, we obtain

$$\begin{aligned}\mathbb{E}\left\|\xi_2 \mathbb{1}_{X\in\Omega_0}\right\|^m_{S_2(\mathcal{H}_X,\mathcal{Y})} \leq& \frac{1}{2}m!(2R)^{m-2}2\sigma^2\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\mathbb{1}_{X\in\Omega_0}\,\mathrm{d}\pi(x)\\ &+2^{m-1}\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\|F_*(x)-F_\lambda(x)\|^m_{\mathcal{Y}}\mathbb{1}_{X\in\Omega_0}\,\mathrm{d}\pi(x).\end{aligned} \tag{33}$$

The first term in Eq. (33) can be bounded using Lemma 10 as below:

$$\frac{1}{2}m!(2R)^{m-2}2\sigma^2\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\mathbb{1}_{x\in\Omega_0}\,\mathrm{d}\pi(x)\leq\frac{1}{2}m!\left(\frac{2RA}{\lambda^{\alpha/2}}\right)^{m-2}2\sigma^2\mathcal{N}(\lambda).$$

For the second term, note that if $\beta\geq\alpha$, by Lemma 5 Eq. (18),

$$\left\|(F_*-[F_\lambda])\,\mathbb{1}_{X\in\Omega_0}\right\|_{L_\infty}\leq\|F_*-[F_\lambda]\|_{L_\infty}\leq\left(\|F_*\|_{L_\infty}+A\|F_*\|_\beta\right)\lambda^{\frac{\beta-\alpha}{2}}\leq 2A\|F_*\|_\beta\lambda^{\frac{\beta-\alpha}{2}},$$

and if $\beta<\alpha$, by Lemma 5 Eq. (19),

$$\left\|(F_*-[F_\lambda])\,\mathbb{1}_{X\in\Omega_0}\right\|_{L_\infty}\leq\left\|F_*\mathbb{1}_{X\in\Omega_0}\right\|_{L_\infty}+\|[F_\lambda]\|_{L_\infty}\leq t+A\,\|F_*\|_\beta\,\lambda^{\frac{\beta-\alpha}{2}}.$$

Therefore, for all $\beta\in(0,2]$,

$$\left\|(F_*-[F_\lambda])\,\mathbb{1}_{X\in\Omega_0}\right\|_{L_\infty}\leq t+2A\,\|F_*\|_\beta\,\lambda^{\frac{\beta-\alpha}{2}}\leq Q(t,\lambda).$$

We have, using Lemma 10 again,

$$\begin{aligned}&2^{m-1}\int_{E_X}\|h_x\|^m_{\mathcal{H}_X}\|F_*(x)-F_\lambda(x)\|^m_{\mathcal{Y}}\mathbb{1}_{x\in\Omega_0}\,\mathrm{d}\pi(x)\\ \leq&\frac{1}{2}\left(\frac{2A}{\lambda^{\alpha/2}}\right)^m Q(t,\lambda)^{m-2}\int_{E_X}\|F_*(x)-F_\lambda(x)\|^2_{\mathcal{Y}}\mathbb{1}_{x\in\Omega_0}\,\mathrm{d}\pi(x)\\ \leq&\frac{1}{2}m!\left(\frac{2Q(t,\lambda)A}{\lambda^{\alpha/2}}\right)^{m-2}\|F_*-[F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{2A^2}{\lambda^\alpha}.\end{aligned} \tag{34}$$

Finally, we get

$$\mathbb{E}\|\xi_2\|^m_{S_2(\mathcal{H}_X,\mathcal{Y})}\leq\frac{1}{2}m!\left(\frac{2Q(t,\lambda)A}{\lambda^{\alpha/2}}\right)^{m-2}2\left(\sigma^2\mathcal{N}(\lambda)+\|F_*-[F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right).$$

An application of Bernstein's inequality from Theorem 13 with

$$L=2Q(t,\lambda)A\lambda^{-\alpha/2},\qquad\sigma^2=2\left(\sigma^2\mathcal{N}(\lambda)+\|F_*-[F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}A^2\lambda^{-\alpha}\right)$$

yields the bound. ■

B.1 Learning Rates

In this section, we establish the upper bound on the learning rate for Theorem 3. To this end, we first look at Theorem 9, with probability over $1-4e^{-\tau}-\tau_n$ with $\tau_n = 1-\left(1-\frac{C_q^q}{t^q}\right)^n$,

$$\left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|^2_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} \le \frac{1152\tau^2}{n\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2A^2}{n\lambda^\alpha}\right)$$
$$+ \frac{18}{\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right)\frac{C_q^q}{t^q}.$$

We first notice that if $t > n^{1/q}$, the second term on the r.h.s is

$$\frac{18C_q^q}{n\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})}\frac{A^2}{\lambda^\alpha}\right).$$

Moreover, by Bernouilli's inequality, $\tau_n \le \frac{C_q^q n}{t^q}$. As such, given any $\tau \ge 1$, if $t > n^{1/q}$, there is a $n_0' \ge 1$ such that $\tau_n \le e^{-\tau}$ for all $n \ge n_0'$. We therefore have with probability greater than $1-5e^{-\tau}$, for all $n \ge n_0'$,

$$\left\|\left[\hat{C}_\lambda - C_\lambda\right]\right\|^2_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} \le \frac{c_0\tau^2}{n\lambda^\gamma}\left(\sigma^2\mathcal{N}(\lambda) + \frac{\|F_* - [F_\lambda]\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda^\alpha} + \frac{2Q(t,\lambda)^2A^2}{n\lambda^\alpha}\right). \tag{35}$$

where $c_0 = \max\{1152, 18C_q^q\}$. From now on, we denote λ as λ_n to indicate that λ is a function of n and we pick $n_1 \ge 1$ such that $\lambda_n \le 1$ for all $n \ge n_1$. For $n \ge n_1$, we have,

$$\begin{aligned}
Q(t,\lambda_n)^2 &= \max\{R^2, \left(t + 2A\,\|F_*\|_\beta\,\lambda_n^{\frac{\beta-\alpha}{2}}\right)^2\} \\
&\le \max\left\{R^2, 2t^2 + 8A^2B^2\lambda_n^{-(\alpha-\beta)}\right\} \\
&\le \max\left\{R^2, 8A^2B^2\lambda_n^{-(\alpha-\beta)}\right\} + 2t^2 \\
&\le C_0(\lambda_n^{-(\alpha-\beta)_+} + t^2),
\end{aligned}$$

where $C_0 := \max\left\{R^2, 8A^2B^2, 2\right\}$. As a result, Eq. (35) become

$$\left\|\left[\hat{C}_{\lambda_n} - C_{\lambda_n}\right]\right\|^2_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} \le \frac{c_0\tau^2}{n\lambda_n^\gamma}\left(\sigma^2\mathcal{N}(\lambda_n) + \frac{\|F_* - [F_{\lambda_n}]\|^2_{L_2(\pi;\mathcal{Y})} A^2}{\lambda_n^\alpha} + 2C_0A^2\frac{\lambda_n^{-(\alpha-\beta)_+}}{n\lambda_n^\alpha}\right)$$
$$+ 2C_0A^2c_0\tau^2\frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}. \tag{36}$$

Using Lemma 10 and Lemma 3 with $\gamma = 0$, we can see that the first term on the r.h.s coincides with Eq. (30) up to some constants. Hence, the analysis in Section A.3 carries on. In particular, the choice of $n_0 \ge 1$, such that $n \ge 8A^2\tau g_{\lambda_n}\lambda_n^{-\alpha}$ for all $n \ge n_0$ remains the same as in Section A.3 (see Eq. (30)). For $n \ge n_1$, using $\lambda_n \le 1$ and $C_1 := D\sigma^2 + A^2B^2$,

$$\begin{aligned}
\left\|\left[\hat{F}_{\lambda_n} - F_{\lambda_n}\right]\right\|^2_\gamma = \left\|\left[\hat{C}_{\lambda_n} - C_{\lambda_n}\right]\right\|^2_{S_2([\mathcal{H}]^\gamma_X,\mathcal{Y})} &\le \frac{c_0\tau^2}{n\lambda_n^\gamma}\left(C_1\lambda_n^{-\max\{p,\alpha-\beta\}} + 2C_0A^2\frac{\lambda_n^{-(\alpha-\beta)_+}}{n\lambda_n^\alpha}\right) + 2C_0A^2c_0\tau^2\frac{t^2}{n^2\lambda_n^{\alpha+\gamma}} \\
&\le \frac{c_0\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}}\left(C_1 + 2C_0A^2\frac{1}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}\right) + 2C_0A^2c_0\tau^2\frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}
\end{aligned}$$

There is a constant $C_2 > 0$ such that for all $n \geq n_2$, for $n_2 := \max\{n_0', n_0, n_1\}$,

$$\left\|[\hat{F}_{\lambda_n} - F_{\lambda_n}]\right\|_\gamma^2 \leq \frac{c_0\tau^2}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}}\left(C_1 + 2C_0A^2C_2\right) + 2C_0A^2c_0\tau^2\frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}.$$

Using the bias-variance splitting from Eq. (14) and Lemma 1, let $C_3 := \max\{c_0\left(C_1 + 2A^2C_0C_2\right), 2C_0A^2c_0\}$, we have

$$\begin{aligned}\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 &\leq 2\|C_*\|^2_{S_2([\mathcal{H}]^\beta_X,\mathcal{Y})}\lambda_n^{\beta-\gamma} + 2C_3\tau^2\left(\frac{1}{n\lambda_n^{\gamma+\max\{p,\alpha-\beta\}}} + \frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}\right)\\ &\leq \tau^2\lambda_n^{\beta-\gamma}\left(2\|C_*\|^2_{S_2([\mathcal{H}]^\beta_X,\mathcal{Y})} + 2C_3\frac{1}{n\lambda_n^{\max\{\beta+p,\alpha\}}}\right) + 2C_3\tau^2\frac{t^2}{n^2\lambda_n^{\alpha+\gamma}},\end{aligned}$$

where we used $\tau \geq 1$.

We now have two scenarios. We first assume that $\beta + p > \alpha$. If we choose $\lambda_n = n^{-1/(\beta+p)}$ there is some constant $J > 0$ such that

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 \leq \tau^2 J\left(\lambda_n^{\beta-\gamma} + \frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}\right)$$

for all $n \geq n_2$. In order for the second term to match the first term, we will need $\frac{t^2}{n^2\lambda_n^\alpha} \leq \lambda_n^\beta$ when $\lambda_n = n^{-\frac{1}{\beta+p}}$. This amounts to require that

$$t \leq n^{\frac{\beta+2p-\alpha}{2(\beta+p)}}.$$

Recall that we require $t > n^{1/q}$, combining with the above, we need

$$q > \frac{2(\beta+p)}{\beta+2p-\alpha} =: q_0.$$

Therefore we need $F_* \in L_q(\pi;\mathcal{Y})$ with $q > q_0$. By Theorem 4, $F_* \in L_{q_1}(\pi;\mathcal{Y})$ with $q_1 := \frac{2\alpha}{\alpha-\beta}$, and since $q_1 > q_0$, the requirement is automatically satisfied.

Secondly, we assume that $\beta + p \leq \alpha$. If we choose $\lambda_n = \left(n/\log^\theta n\right)^{-\frac{1}{\alpha}}$ with $\theta > 1$, we obtain, for some constant $J > 0$ that

$$\left\|[\hat{F}_{\lambda_n}] - F_*\right\|_\gamma^2 \leq \tau^2 J\left(\lambda_n^{\beta-\gamma} + \frac{t^2}{n^2\lambda_n^{\alpha+\gamma}}\right)$$

for all $n \geq n_2$. In order for the second term to match the first term, we will need $\frac{t^2}{n^2\lambda_n^\alpha} \leq \lambda_n^\beta$ when $\lambda_n = \left(n/\log^\theta n\right)^{-\frac{1}{\alpha}}$. This amounts to require that

$$t \leq n\left(\frac{\log^\theta n}{n}\right)^{\frac{\alpha+\beta}{2\alpha}}.$$

Merging with the constraint $t > n^{1/q}$, we require

$$q \geq \frac{2\alpha}{\alpha-\beta} = q_1.$$

Therefore we need $F_* \in L_{q_1}(\pi;\mathcal{Y})$ which is automatically satisfied by Theorem 4. This concludes the proof of Theorem 3.

B.2 Proof of Theorem 4

Let us drop the subscript α, β and use $q := q_{\alpha,\beta}$. Theorem 5 in Zhang et al. (2023a) proves the scalar-valued L_q–embedding property: the inclusion $[\mathcal{H}]^\beta_X \hookrightarrow L_q(\pi, \mathbb{R})$ is bounded with operator norm denoted as M.

Let $F \in [\mathcal{G}]^\beta$ be given by the representation $F(\cdot) = \sum_{i=1}^N f_i(\cdot) y_i \simeq \sum_{i=1}^N y_i \otimes f_i$ with $f_i \in [\mathcal{H}]^\beta$ and $y_i \in \mathcal{Y}$. We may assume without loss of generality that the y_i are orthogonal in $\mathcal{Y}$ (to obtain such representation we can apply the Gram-Schmidt process to any finite-rank expansion). Note in particular that functions of this form are dense in $[\mathcal{G}]^\beta$ by construction. We prove the claim by manually bounding the norm of F in $L_q(\pi; \mathbb{R})$ by a constant multiple of its norm in $[\mathcal{G}]^\beta$. We have

$$\begin{aligned}
\|F\|^2_{L_q(\pi;\mathcal{Y})} &= \left(\int \left\| \sum_{i=1}^N f_i(x) y_i \right\|^q_{\mathcal{Y}} \, \mathrm{d}\pi(x) \right)^{2/q} && \text{(definition of } \|\cdot\|_{L_q(\pi;\mathcal{Y})} \text{)} \\
&= \left(\int \left(\sum_{i,j=1}^N f_i(x) f_j(x) \langle y_i, y_j \rangle_{\mathcal{Y}} \right)^{q/2} \, \mathrm{d}\pi(x) \right)^{2/q} \\
&= \left(\int \left(\sum_{i=1}^N f_i(x)^2 \, \|y_i\|^2_{\mathcal{Y}} \right)^{q/2} \, \mathrm{d}\pi(x) \right)^{2/q} && (y_i \text{ orthogonal in } \mathcal{Y}) \\
&= \left\| \sum_{i=1}^N f_i(\cdot)^2 \, \|y_i\|^2_{\mathcal{Y}} \right\|_{L_{q/2}(\pi;\mathbb{R})} && \text{(definition of } \|\cdot\|_{L_{q/2}(\pi;\mathbb{R})}\text{)} \\
&\leq \sum_{i=1}^N \|y_i\|^2_{\mathcal{Y}} \, \|f_i^2\|_{L_{q/2}(\pi;\mathbb{R})} && \text{(triangle inequality)} \\
&= \sum_{i=1}^N \|y_i\|^2_{\mathcal{Y}} \, \|f_i\|^2_{L_q(\pi;\mathbb{R})} \\
&\leq M^2 \sum_{i=1}^N \|y_i\|^2_{\mathcal{Y}} \, \|f_i\|^2_{[\mathcal{H}]^\beta_X} && \text{(scalar-valued } L_q\text{-embedding property)} \\
&= M^2 \|F\|^2_{[\mathcal{G}]^\beta} && \text{(definition of } [\mathcal{G}]^\beta \text{ and } y_i \text{ orthogonal in } \mathcal{Y}\text{).}
\end{aligned}$$

The standard denseness argument proves this property for all $F \in [\mathcal{G}]^\beta$.

Appendix C. Proof of Theorem 5

The proof of the lower bound is performed by projecting the infinite-dimensional response variable Y onto a one-dimensional subspace of $\mathcal{Y}$ and applying arguments from the real-valued learning scenario.

We start by noticing that for any $F \in L_2(\pi; \mathcal{Y})$ and $a \in \mathcal{Y}$,

$$\begin{aligned}
\int_{E_X} \left(\langle F(x), a \rangle_{\mathcal{Y}} - \langle F_*(x), a \rangle_{\mathcal{Y}} \right)^2 d\pi(x) &\leq \int_{E_X} \|F(x) - F_*(x)\|^2_{\mathcal{Y}} \|a\|^2_{\mathcal{Y}} d\pi(x) \\
&= \|a\|^2_{\mathcal{Y}} \|F - F_*\|^2_{L_2(\pi;\mathcal{Y})}.
\end{aligned} \tag{37}$$

Moreover, by Lemma 8, the inequality holds for general γ-norm (which implies the previous equation, setting $\gamma = 0$),

$$\|\langle F(.), a\rangle_{\mathcal{Y}} - \langle F_*(.), a\rangle_{\mathcal{Y}}\|_\gamma \leq \|a\|_{\mathcal{Y}}\|F - F_*\|_\gamma. \tag{38}$$

Lemma 8 *Let $\gamma \geq 0$, for any $F \in [\mathcal{G}]^\gamma$ and $a \in \mathcal{Y}$, we have*

$$\|\langle F(.), a\rangle_{\mathcal{Y}}\|_\gamma \leq \|a\|_{\mathcal{Y}}\|F\|_\gamma.$$

Proof The case where $\gamma = 0$ is already proved in Eq. (37). We now let $\gamma > 0$. Recall $\{d_j\}_{j\in J}$ and $\{\mu_i^{\gamma/2}[e_i]\}_{i\in I}$ are the orthonormal basis of $\mathcal{Y}$ and $[\mathcal{H}]_X^\gamma$, since $F \in [\mathcal{G}]^\gamma$, we can write F as

$$F = \sum_{i\in I, j\in J} a_{ij} d_j \mu_i^{\gamma/2}[e_i].$$

Therefore, we have

$$\langle F(\cdot), a\rangle_{\mathcal{Y}} = \sum_{i\in I, j\in J} a_{ij}\langle d_j, a\rangle_{\mathcal{Y}} \mu_i^{\gamma/2}[e_i](\cdot).$$

$\langle F(\cdot), a\rangle_{\mathcal{Y}}$ is an element of $[\mathcal{H}]_X^\gamma$ as

$$\begin{aligned}\|\langle F(\cdot), a\rangle_{\mathcal{Y}}\|_\gamma^2 &= \sum_{i\in I}\left(\sum_{j\in J} a_{ij}\langle d_j, a\rangle_{\mathcal{Y}}\right)^2 \\ &\leq \sum_{i\in I}\sum_{j\in J} a_{ij}^2 \sum_{j\in J}\langle d_j, a\rangle_{\mathcal{Y}}^2 \\ &= \|a\|_{\mathcal{Y}}^2 \sum_{i\in I, j\in J} a_{ij}^2 \\ &= \|a\|_{\mathcal{Y}}^2 \|F\|_\gamma^2 < +\infty,\end{aligned}$$

where for the second step, we used Cauchy-Schwartz inequality and for the third step Parseval's identity. ■

We now express the l.h.s as the risk of a scalar-valued regression. Consider a distribution P on $E_X \times \mathcal{Y}$ that factorizes as $P(x, y) = p(y \mid x)\pi(x)$ for all $(x, y) \in E_X \times \mathcal{Y}$. For all $x \in E_X$, $p(\cdot \mid x)$ defines a probability distribution on $\mathcal{Y}$. We fix an element $a \in \mathcal{Y}$, $a \neq 0$ and define $\mathcal{Y}^a := \{y_a \in \mathbb{R} \mid y_a = \langle y, a\rangle_{\mathcal{Y}}, y \in \mathcal{Y}\}$. Since $\mathcal{Y}$ is a Hilbert space hence a vector space, we have $\mathcal{Y}^a = \mathbb{R}$. Consider the joint distribution P_a on $E_X \times \mathbb{R}$ such that

$$\begin{aligned} p_a(. \mid x) &:= (\langle \cdot, a\rangle_{\mathcal{Y}})_{\#}\, p(\cdot \mid x) \\ P_a(x, y_a) &:= p_a(y_a \mid x)\pi(x), \quad (x, y_a) \in E_X \times \mathbb{R} \end{aligned} \tag{39}$$

where # denotes the push-forward operation. For a dataset $D = \{(x_i, y_i)\}_{i=1}^n \in (E_X \times \mathcal{Y})^n$ where the data are i.i.d from P, the dataset $D_a = \{(x_i, y_{a.i})\}_{i=1}^n \in (E_X \times \mathbb{R})^n$ where $y_{a.i} := \langle y_i, a\rangle_{\mathcal{Y}}$ for all $i = 1, \ldots, n$ is i.i.d from P_a. Note that $p_a(\cdot \mid x)$ is a probability distribution on $\mathbb{R}$ for all x supported by π. By definition of the push-forward operator, the Bayes predictor associated to the joint distribution P_a is

$$\begin{aligned} f_{a,*}(x) = \int_{\mathbb{R}} y_a dp_a(y_a \mid x) &= \int_{\mathcal{Y}} \langle y, a\rangle_{\mathcal{Y}} dp(y \mid x) \\ &= \left\langle \int_{\mathcal{Y}} y dp(y \mid x), a \right\rangle_{\mathcal{Y}} \\ &= \langle F_*(x), a\rangle_{\mathcal{Y}} \end{aligned} \tag{40}$$

where F_* is the $\mathcal{Y}$-valued Bayes predictor associated to P. Therefore plugging Eq. (40) in Eq. (38) we obtain that for any learning method $D \to \hat{F}_D \in \mathcal{Y}^{E_X}$

$$\|\hat{F}_D - F_*\|_\gamma \geq \|a\|_{\mathcal{Y}}^{-1} \|\hat{f}_{D_a} - f_{a.*}\|_\gamma \tag{41}$$

where $\hat{f}_{D_a}(\cdot) \coloneqq \langle \hat{F}_D(\cdot), a\rangle_{\mathcal{Y}}$. The r.h.s is the error measured in (scalar) γ-norm of the learning method $D_a \to \hat{f}_{D_a} \in \mathbb{R}^{E_X}$ on the scalar-regression learning problem associated to D_a.

In what follows we fix $\{d_n\}_{n\geq 1}$ an orthonormal basis of $\mathcal{Y}$ and take $a = d_1$.

To derive a lower bound on the r.h.s in Eq. (41), the strategy is to define a conditional distribution $p_a(. \mid x)$ on $\mathbb{R}$, $x \in E_X$, that is difficult to learn. We offer to use Gaussian conditional distributions as in Fischer and Steinwart (2020). The additional difficulty in our setting is to show the existence of conditional distributions $p(. \mid x)$ on $\mathcal{Y}$, $x \in E_X$ such that Eq. (39) holds and such that $F_* \in [\mathcal{G}]^\beta$ and (MOM) are satisfied. We make use of the following Lemma that corresponds to Lemma 19, Lemma 23 and Equation (55) in Fischer and Steinwart (2020).

Lemma 9 *Let k_X be a kernel on E_X such that assumptions 1 to 3 hold and π be a probability distribution on E_X such that* (EVD+) *is satisfied for some $0 < p \leq 1$. Then, for all parameters $0 < \beta \leq 2, 0 \leq \gamma \leq 1$ with $\gamma < \beta$ and all constant $B > 0$, there exist constants $0 < \epsilon_0 \leq 1$ and $L_0, L > 0$ such that the following statement is satisfied: for all $0 < \epsilon \leq \epsilon_0$ there is an $M_\epsilon \geq 1$ with*

$$2^{L\epsilon^{-u}} \leq M_\epsilon \leq 2^{3L\epsilon^{-u}}$$

where $u \coloneqq \frac{p}{\beta-\gamma}$, and functions $f_0, \ldots, f_{M_\epsilon}$ such that $f_i \in [\mathcal{H}]_X^\beta$, $\|f_i\|_\beta \leq B$, and

$$\begin{aligned} &\|f_i - f_j\|_\gamma^2 \geq 4\epsilon \\ &\|f_i - f_j\|_{L_2(\pi)}^2 \leq 32 L_0^\gamma \epsilon m^{-\gamma/p}, \end{aligned}$$

for all $i, j \in \{0, \ldots, M_\varepsilon\}$ with $i \neq j$ where $m \leq U\epsilon^{-u}$ for some constant $U > 0$.

Recall that the Kullback-Leibler divergence of two probability measures P_1, P_2 on some measurable space $(\Omega, \mathcal{A})$ is given by

$$KL(P_1, P_2) \coloneqq \int_\Omega \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1$$

if $P_1 \ll P_2$ and otherwise $KL(P_1, P_2) \coloneqq \infty$.

Exploiting Lemma 9, given $0 < \beta \leq 2$ and $\sigma, R, B > 0$, we build distributions $P_{d_1,i}$ on $E_X \times \mathbb{R}$ and P_i on $E_X \times \mathcal{Y}$ for $i \in \{1, \ldots, M_\varepsilon\}$ such that,

- Step 1. $f_i = f_{*,i}^{d_1}$ where $f_{*,i}^{d_1}$ is the Bayes predictor associated to $P_{d_1,i}$
- Step 2. $KL(P_{d_1,i}, P_{d_1,j})$ is upper-bounded by a linear function of $\|f_i - f_j\|^2_{L_2(\pi)}$
- Step 3. Eq. (39) holds, i.e. $P_{d_1,i}$ is related to P_i through the projection along the line with direction d_1
- Step 4. The Bayes predictor $F_{*,i}$ associated to P_i is in $[\mathcal{G}]^\beta$ and $\|F_*\|_\beta \le B$
- Step 5. $Y \in L_2(\pi_i, \Omega, \mathbb{P}, \mathcal{Y})$
- Step 6. (MOM) is satisfied with P_i with parameters σ, R.

Gaussian conditional distributions as in Blanchard and Mücke (2018); Fischer and Steinwart (2020). Take $\bar\sigma = \min(\sigma, R)$. For all $i = 1, \ldots, M_\epsilon$ we define the joint distribution $P_{d_1,i}(x,y) = p_{d_1,i}(y \mid x)\pi(x)$ where

$$p_{d_1,i}(\cdot \mid x) = \mathcal{N}\left(f_i(x), \bar\sigma^2\right)$$

is a univariate Gaussian distribution.

Step 1. We automatically get $f_i = f_{*,i}^{d_1}$.

Step 2. The Kullback-Leibler divergence satisfies

$$\begin{aligned} KL(P_{a,i}, P_{a,j}) &= \int_{E_X} KL(p_{a,i}(\cdot \mid x), p_{a,j}(\cdot \mid x)) d\pi(x) \\ &= \frac{1}{2\bar\sigma^2} \int_{E_X} (f_i(x) - f_j(x))^2 d\pi(x) \\ &= \frac{1}{2\bar\sigma^2} \|f_i - f_j\|^2_{L_2(\pi)} \end{aligned}$$

Step 3. We consider the map $y_i : E_X \to \mathcal{Y}, x \mapsto f_i(x)d_1$ such that for all $x \in E_X$, $\langle y_i(x), d_1 \rangle_{\mathcal{Y}} = f_i(x)$, then we build a Gaussian measure on $\mathcal{Y}$ as follows, we fix $x \in E_X$ and consider Z a univariate Gaussian random variable, then

$$X(\omega) = y_i(x) + \bar\sigma Z(\omega) d_1, \qquad \omega \in \Omega$$

is such that (according to Definition 6) $X \sim \mathcal{N}_{\mathcal{Y}}(y_i(x), \bar\sigma^2 d_1 \otimes d_1)$. Therefore, we pick

$$p_i(\cdot \mid x) = \mathcal{N}_{\mathcal{Y}}\left(y_i(x), \bar\sigma^2 d_1 \otimes d_1\right)$$

and clearly $\langle X, a \rangle = f_i(x) + \bar\sigma Z$ so Eq. (39) holds.

Step 4. Note that $F_{*,i}(\cdot) = y_i(\cdot) = f_i(\cdot)d_1$, therefore we have

$$\|F_{*,i}\|_\beta = \|f_i\|_\beta \|d_1\|_{\mathcal{Y}} = \|f_i\|_\beta \le B$$

Step 5. Write $\Sigma := \bar{\sigma}^2 d_1 \otimes d_1$,

$$\begin{aligned}&\int_{E_X \times \mathcal{Y}} \|y\|_{\mathcal{Y}}^2 p_i(x, dy)\pi(dx)\\ &= \int_{E_X} \mathbb{V}(Y \sim \mathcal{N}(0, \Sigma)) + \|y_i(x)\|_{\mathcal{Y}}^2 \pi(dx)\\ &= \mathbb{V}(Y \sim \mathcal{N}(0, \Sigma)) + \|f_i\|_{L_2(\pi)}^2\\ &= \bar{\sigma}^2 + \|f_i\|_{L_2(\pi)}^2 < \infty,\end{aligned}$$

since

$$\mathbb{V}(Y \sim \mathcal{N}(0, \Sigma)) = \sum_{n \geq 1} \langle \Sigma d_n, d_n \rangle = \bar{\sigma}^2.$$

Step 6. We now show that P_i satisfies (MOM) with parameters $\sigma = R = \bar{\sigma}$. We recall that for all $x \in E_X$, we picked $p_i(\cdot \mid x) = \mathcal{N}_{\mathcal{Y}}\left(y_i(x), \bar{\sigma}^2 d_1 \otimes d_1\right)$. We consider the centered random variable $X(\omega) = \bar{\sigma} Z(\omega) d_1$ and (MOM) amounts to show $\mathbb{E}\left[\|X\|_{\mathcal{Y}}^q\right] \leq \frac{1}{2} q!(\bar{\sigma})^q$. But almost surely we have

$$\|X\|_{\mathcal{Y}}^q = \left(\sum_{n \geq 1} |\langle Y, d_n \rangle_{\mathcal{Y}}|^2\right)^{q/2} = \bar{\sigma}^q |Z|^q$$

which brings us back to the univariate case that it easy to demonstrate (see for example Lemma 21 by Fine and Scheinberg, 2002).

Proof of Theorem 5 Combining those 6 points with the proof of Lemma 19 and Theorem 2 Fischer and Steinwart (2020) gives us Theorem 5.

Appendix D. Interpolation theory and proof of Theorem 2

In the next two results, we consider H_1, H_2, two infinite dimensional separable Hilbert spaces such that $H_2 \hookrightarrow H_1$ and such that the inclusion operator $\mathcal{I} : H_2 \to H_1$ performing the change of norms $\mathcal{I}x = x$ for $x \in H_2$, is compact. By the spectral theorem (see e.g., Steinwart and Scovel (2012) Theorem A.3), $\mathcal{I}$ admits a decomposition

$$\mathcal{I} = \sum_{i \in I} \rho_i \mathcal{I}(h_i) \langle \sqrt{\rho_i} h_i, \cdot \rangle_{H_1},$$

where $\{\sqrt{\rho_i} h_i\}_{i \in I}$ is an ONB of $\ker(\mathcal{I})^{\perp}$, $\{\mathcal{I}(h_i)\}_{i \in I}$ is an ONB of $\overline{\operatorname{ran}(\mathcal{I})}$, and $\rho_1 \geq \rho_2 \geq \cdots > 0$ is a positive sequence of singular values, converging to 0. For any $f \in H_1$, we define $a_i := \langle f, \mathcal{I}(h_i) \rangle_{H_1}$, $i \in I$.

We recall that the symbol $\cong$ means that sets coincide and the corresponding norms are equivalent. For details on the theory of interpolation spaces of the real method, see Triebel (1995).

Proposition 3 *Let H_1, H_2 be two separable Hilbert spaces such that $H_2 \hookrightarrow H_1$ and such that the inclusion operator $\mathcal{I} : H_2 \to H_1$ is compact. Then, for all $\theta \in (0, 1)$,*

$$f \in [H_1, \mathcal{I}(H_2)]_{\theta,2} \iff (a_i)_{i \in I} \in \ell_2\left(\rho^{-\theta}\right) := \left\{(a_i)_{i \in I} \in \ell_2(I) \middle| \|(a_i)\|_{\ell_2(\rho^{-\theta})}^2 := \sum_{i \in I} \frac{a_i^2}{\rho_i^{\theta}} < +\infty\right\}.$$

Furthermore, there exist constants $c, C > 0$, such that for all $f \in H_1$,

$$c\|(a_i)\|_{\ell_2(\rho^{-\theta})} \leq \|f\|_{[H_1,\mathcal{I}(H_2)]_{\theta,2}} \leq C\|(a_i)\|_{\ell_2(\rho^{-\theta})}.$$

Proof The result is obtained as a direct consequence of the proof of Steinwart and Scovel (2012) Theorem 4.6, which uses the machinery of the so-called *K-functional* and focuses on a RKHS H_2 compactly embedded into $H_1 = L_2(\pi)$. ■

Theorem 10 *Let H_1, H_2 and E be three separable Hilbert spaces such that $H_2 \hookrightarrow H_1$ and such that the inclusion operator $\mathcal{I} : H_2 \to H_1$ is compact. Then, for all $\theta \in (0, 1)$,*

$$[E \otimes H_1, E \otimes \mathcal{I}(H_2)]_{\theta,2} \cong E \otimes [H_1, \mathcal{I}(H_2)]_{\theta,2}$$

Proof Using the notations of Proposition 3, we have $[H_1, \mathcal{I}(H_2)]_{\theta,2} \cong \ell_2\left(\rho^{-\theta}\right)$. Let $\{h_\ell\}_{\ell \in L}$ be an orthonormal basis for E, then

$$E \otimes [H_1, \mathcal{I}(H_2)]_{\theta,2} \cong \ell_2(L \times \rho^{-\theta}) := \left\{ (a_{i,\ell})_{i \in I, \ell \in L} \in \ell_2(I \times L) \middle| \sum_{i \in I, \ell \in L} \frac{a_{i,\ell}^2}{\rho_i^\theta} < +\infty \right\}.$$

To conclude, we require $[E \otimes H_1, E \otimes \mathcal{I}(H_2)]_{\theta,2} \cong \ell_2(L \times \rho^{-\theta})$. This can be obtained again by following the steps of Steinwart and Scovel (2012) Theorem 4.6 and introducing the orthonormal basis $\{h_\ell\}_{\ell \in L}$ for E. ■

Proof [Proof of Theorem 2] Recall that $I_\pi : \mathcal{H}_X \to L_2(\pi)$ is a compact inclusion. Therefore, using that $L_2(\pi; \mathcal{Y}) \simeq \mathcal{Y} \otimes L_2(\pi)$ and $[\mathcal{G}]^1 \simeq \mathcal{Y} \otimes I_\pi(\mathcal{H}_X)$, by Theorem 10 and Remark 4, we have

$$\left[L_2(\pi; \mathcal{Y}), [\mathcal{G}]^1\right]_{\alpha,2} \simeq [\mathcal{Y} \otimes L_2(\pi), \mathcal{Y} \otimes I_\pi(\mathcal{H}_X)]_{\alpha,2} \cong \mathcal{Y} \otimes [L_2(\pi), I_\pi(\mathcal{H}_X)]_{\alpha,2} \cong \mathcal{Y} \otimes [\mathcal{H}]_X^\alpha \simeq [\mathcal{G}]^\alpha.$$

■

Appendix E. Proofs of Section 6

We gather technical results related to Sobolev spaces. To this end, in the rest of this section we assume that $E_X \subseteq \mathbb{R}^d$ is a bounded domain with smooth boundary equipped with the Lebesgue measure μ and denote $L_2(E_X) := L_2(E_X, \mu)$ as the corresponding L_2 space. For $s > 0$, we denote $W^{s,2}(E_X)$ as the (fractional) Sobolev space with smoothness s (Adams and Fournier, 2003, Definition 7.32). Note that $W^{s,2}(E_X)$ is a subset of $L_2(E_X)$ and therefore not a space of functions, however we have the following well-known Sobolev embedding theorem. Let $C_0(E_X)$ be the space of bounded and continuous functions equipped with the norm $\|f\|_\infty := \sup_{x \in E_X} |f(x)|$, $f \in C_0(E_X)$.

Theorem 11 (Sobolev embedding theorem, Adams and Fournier (2003) Theorem 7.34 (c)) *If $s > d/2$, for each $f \in W^{s,2}(E_X)$, there exists a unique function in $C_0(E_X)$, denoted $\bar{f}$ such* *that $f = \bar{f}$ μ−almost everywhere. Furthermore, there is a constant $C_\infty \geq 0$ such that for all $f \in W^{s,2}(E_X)$,*

$$\|\bar{f}\|_\infty \leq C_\infty \|f\|_{W^{s,2}(E_X)}.$$

In short, if $s > d/2$, $W^{s,2}(E_X) \hookrightarrow C_0(E_X)$. As a consequence for $s > d/2$, we can build a RKHS from $W^{s,2}(E_X)$.

Theorem 12 *For $s > d/2$, define*

$$\bar{W}^{s,2}(E_X) \coloneqq \{\bar{f} \in C_0(E_X) : [\bar{f}]_\mu \in W^{s,2}(E_X)\}$$

equipped with the norm $\|\bar{f}\|_{\bar{W}^{s,2}(E_X)} \coloneqq \|[\bar{f}]_\mu\|_{W^{s,2}(E_X)}$. $\bar{W}^{s,2}(E_X)$ is a separable RKHS (with respect to a kernel k_s) that we call the Sobolev RKHS. Furthermore, k_s is bounded and measurable. Therefore assumptions 1 to 3 are satisfied for k_s and μ.

Proof For any $x \in E_X$ and $\bar{f} \in \bar{W}^{s,2}(E_X)$, by the Sobolev embedding theorem,

$$|\bar{f}(x)| \leq \|\bar{f}\|_\infty \leq C_\infty \|[\bar{f}]_\mu\|_{W^{s,2}(E_X)} = \|\bar{f}\|_{\bar{W}^{s,2}(E_X)}.$$

Therefore, the evaluation functional is continuous, proving that $\bar{W}^{s,2}(E_X)$ is a RKHS. k_s is bounded by (Steinwart and Christmann, 2008, Lemma 4.23) and measurable by (Steinwart and Christmann, 2008, Lemma 4.24). ■

We now characterise (EMB) and (EVD) for $\bar{W}^{s,2}(E_X)$.

Proposition 4 *For $s > d/2$ and $\mathcal{H}_X = \bar{W}^{s,2}(E_X)$,* (EVD) *is satisfied with $p = d/(2s)$ and* (EMB) *is satisfied for any $\alpha \in (p, 1]$.*

Proof For (EVD) see Edmunds and Triebel (1996). For $\alpha \in (0,1]$, it is shown in Fischer and Steinwart (2020) Eq. (14) that $[\bar{W}^{s,2}(E_X)]^\alpha_\mu \simeq W^{\alpha s,2}(E_X)$. Hence by Theorem 11, if $\alpha s > d/2$,

$$[\bar{W}^{s,2}(E_X)]^\alpha_\mu \simeq W^{\alpha s,2}(E_X) \hookrightarrow C_0(E_X) \hookrightarrow L_\infty(\mu).$$

■

We thank Haobo Zhang for bringing to our attention a sketch of proof for the next result. Up to our knowledge this result was only proved with $\theta \in (0,1]$.

Proposition 5 *For $s > d/2$ and $\mathcal{H}_X = \bar{W}^{s,2}(E_X)$, for all $\theta > 0$, $[\bar{W}^{s,2}(E_X)]^\theta_\mu \simeq W^{\theta s,2}(E_X)$.*

Proof For $\theta \in (0,1]$, see Fischer and Steinwart (2020) Eq. (14). For $\theta > 1$, since $\theta^{-1} \in (0,1)$, we have

$$[\bar{W}^{s\theta,2}(E_X)]^{\theta^{-1}}_\mu = W^{s,2}(E_X).$$

Since $s\theta > d/2$, by Theorem 12, $\bar{W}^{s\theta,2}(E_X)$ is a RKHS with a kernel satisfying assumptions 1 to 3 for μ. Therefore it is compactly embedded into $L_2(E_X)$ and the singular value decomposition of the inclusion $\bar{W}^{s\theta,2}(E_X) \hookrightarrow L_2(E_X)$ leads to a characterization of $\bar{W}^{s\theta,2}(E_X)$ as

$$\bar{W}^{s\theta,2}(E_X) = \left\{\sum_{i\in I} a_i \sqrt{\lambda_i} f_i : (a_i)_{i\in I} \in \ell_2(I)\right\}$$

with $\lambda_i > 0$ for all $i \in I$ and $\{\sqrt{\lambda_i} f_i\}_{i \in I}$ forming an ONB in $\bar{W}^{s\theta,2}(E_X)$ (Steinwart and Scovel, 2012, Lemma 2.6). We therefore have

$$W^{s,2}(E_X) = [\bar{W}^{s\theta,2}(E_X)]_{\mu}^{\theta^{-1}} = \left\{ \sum_{i \in I} a_i \lambda_i^{\frac{1}{2\theta}} [f_i]_{\nu} : (a_i)_{i \in I} \in \ell_2(I) \right\}$$

which further proves

$$\bar{W}^{s,2}(E_X) = \left\{ \sum_{i \in I} a_i \lambda_i^{\frac{1}{2\theta}} f_i : (a_i)_{i \in I} \in \ell_2(I) \right\}.$$

Since $s > d/2$, $\bar{W}^{s,2}(E_X)$ is a RKHS with bounded kernel k_s and by (Steinwart and Scovel, 2012, Lemma 2.6), we must have for all $x \in E_X$

$$k_s(x,x) = \sum_{i \in I} \lambda_i^{\frac{1}{\theta}} f_i(x)^2 < +\infty.$$

Therefore, using Proposition 6 with $\mathcal{H}_X = \bar{W}^{s\theta,2}(E_X)$, $\beta = \theta^{-1}$ and $\alpha = \theta$, we get

$$[\bar{W}^{s,2}(E_X)]_{\mu}^{\theta} = [\bar{W}^{s\theta,2}(E_X)]_{\mu}^{\theta^{-1} \cdot \theta} = [\bar{W}^{s\theta,2}(E_X)]_{\mu}^{1} = W^{s\theta,2}(E_X).$$

■

Proposition 6 *Let π be a probability measure on E_X and $\mathcal{H}_X$ be a RKHS on E_X with kernel k_X such that assumptions 1 to 3 are satisfied. Let $(\mu_i)_{i \in I} > 0$ be a non-increasing sequence, and let $(e_i)_{i \in I} \in \mathcal{H}_X$ be a family such that $([e_i])_{i \in I}$ is an orthonormal basis (ONB) of $\overline{\operatorname{ran} I_\pi} \subseteq L_2(\pi)$ and such that Eq. (9) holds. We denote the integral operator by $L_{k_X} := L_X$ to highlight the dependence on k_X. If $\beta > 0$ is such that*

$$\sum_{i \in I} \mu_i^{\beta} e_i(x)^2 < +\infty, \qquad x \in E_X,$$

then

$$\mathcal{H}_X^{\beta} := \left\{ \sum_{i \in I} a_i \mu_i^{\beta/2} e_i : (a_i)_{i \in I} \in \ell_2(I) \right\}$$

with the norm

$$\left\| \sum_{i \in I} a_i \mu_i^{\beta/2} [e_i] \right\|_{\mathcal{H}_X^{\beta}} := \|(a_i)_{i \in I}\|_{\ell_2(I)} = \left(\sum_{i \in I} a_i^2 \right)^{1/2}$$

is a RKHS compactly embedded into $L_2(\pi)$ and its kernel k_X^{β} is given by

$$k_X^{\beta}(x,z) = \sum_{i \in I} \mu_i^{\beta} e_i(x) e_i(z), \qquad x, z \in E_X.$$

Furthermore, we have $L_{k_X}^{\beta} = L_{k_X^{\beta}}$ which implies that for all $\alpha > 0$,

$$[\mathcal{H}_X^{\beta}]_{\pi}^{\alpha} = [\mathcal{H}_X]_{\pi}^{\beta \cdot \alpha}.$$

Proof For the proof that $\mathcal{H}_X^\beta$ is a RKHS with kernel k_X^β, see (Steinwart and Scovel, 2012, Definition 4.1 and Proposition 4.2) and for the proof that $L_{k_X}^\beta = L_{k_X^\beta}$ see (Steinwart and Scovel, 2012, Lemma 4.3). For the last point recall that interpolation spaces are defined such that $[\mathcal{H}_X^\beta]_\pi^\alpha = \operatorname{ran} L_{k_X^\beta}^{\alpha/2}$, and since $L_{k_X}^\beta = L_{k_X^\beta}$ we have

$$[\mathcal{H}_X^\beta]_\pi^\alpha = \operatorname{ran} L_{k_X^\beta}^{\alpha/2} = L_{k_X}^{(\beta\cdot\alpha)/2} = [\mathcal{H}_X]_\pi^{\beta\cdot\alpha}.$$

■

Proof [Proof of Proposition 1] Using Definition 4 combined with Theorem 6 and Theorem 10, we have for $r > 0$ and $m := \min\{s \in \mathbb{N} : s > r\}$,

$$\begin{aligned} W^{r,2}(E_X;\mathcal{Y}) &= \left[L_2(E_X;\mathcal{Y}), W^{m,2}(E_X;\mathcal{Y})\right]_{r/m,2} \\ &\cong \left[\mathcal{Y}\otimes L_2(E_X), \mathcal{Y}\otimes W^{m,2}(E_X)\right]_{r/m,2} \\ &\cong \mathcal{Y}\otimes\left[L_2(E_X), W^{m,2}(E_X)\right]_{r/m,2} \\ &= \mathcal{Y}\otimes W^{r,2}(E_X). \end{aligned}$$

Then, by Proposition 5, if k_X is a kernel on E_X such that $\mathcal{H}_X = \bar{W}^{m,2}(E_X)$ with $m > d/2$, then for all $r \geq 0$,

$$[\mathcal{G}]^{r/m} \simeq \mathcal{Y}\otimes[\mathcal{H}]_X^{r/m} \simeq \mathcal{Y}\otimes W^{r,2}(E_X) \cong W^{r,2}(E_X;\mathcal{Y}).$$

■

Appendix F. Auxiliary Results

The following lemma is from Lemma 11 and 13 Fischer and Steinwart (2020).

Lemma 10 *Under* (EMB), *we have*

$$\left\|(C_{XX} + \lambda Id_{\mathcal{H}_X})^{-\frac{1}{2}} k_X(X,\cdot)\right\|_{\mathcal{H}_X} \leq A\lambda^{-\frac{\alpha}{2}}.$$

Under (EVD), *there exists a constant $D > 0$ such that*

$$\mathcal{N}(\lambda) = Tr\left(C_{XX}\,(C_{XX} + \lambda Id_{\mathcal{H}_X})^{-1}\right) \leq D\lambda^{-p}.$$

The following Theorem is from Fischer and Steinwart (2020, Theorem 26).

Theorem 13 (Bernstein's inequality) . *Let $(\Omega,\mathcal{B},P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with*

$$\mathbb{E}[\|\xi\|_H^m] \leq \frac{1}{2}m!\sigma^2 L^{m-2}$$

for all $m \geq 2$. Then, for $\tau \geq 1$ and $n \geq 1$, the following concentration inequality is satisfied

$$P^n\left((\omega_1,\ldots,\omega_n) \in \Omega^n : \left\|\frac{1}{n}\sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P\xi\right\|_H^2 \geq 32\frac{\tau^2}{n}\left(\sigma^2 + \frac{L^2}{n}\right)\right) \leq 2e^{-\tau}.$$

Definition 6 (Gaussian Measures on separable Hilbert spaces Bogachev, 1998) *Let H be a separable Hilbert space, $\mathcal{F}_H$ be the Borel σ-algebra defined on H and let H^*, the topological dual space, be identified with H by means of the Riesz representation. A probability measure γ on $(H, \mathcal{F}_H)$ is said to be Gaussian if $\gamma \circ f^{-1}$ is a Gaussian measure in $\mathbb{R}$ for every $f \in H^*$. For the Gaussian measure γ and any $f \in H^*$, we define the mean and covariance as*

$$\begin{aligned}\mu_\gamma(f) &:= \int_H f(x)\gamma(dx)\\ Q_\gamma(f,g) &:= \int_H [f(x) - \mu_\gamma(f)]\,[g(x) - \mu_\gamma(g)]\,\gamma(dx).\end{aligned}$$

By the Riesz representation theorem, for any $f \in H^*$ there is a unique $v_f \in H$ such that $f(x) = \langle x, v_f \rangle_H$ for all $x \in H$. It is then straightforward to see that $\mu_\gamma(f) = \langle \mu, v_f \rangle_H$ for all $f \in H^*$ where $\mu := \int_H x\gamma(dx)$ and $Q_\gamma(f,g) = \langle Q v_f, v_g \rangle_H$ for all $f, g \in H^*$ where $Q := \int_H (x - \mu) \otimes (x - \mu)\,\gamma(dx)$. This justifies the notation $\mathcal{N}_H(\mu, Q)$ for the Gaussian measure γ and we we write the variance as $\mathbb{V}(\gamma) := Q$.