SUPPLEMENTARY MATERIAL.

This supplementary material is organized as follows. Section A provides a reminder about operator-valued kernels and vector-valued RKHSs. In Section B, we detail the proofs of the propositions from Section 4 of the main paper. In Section C, we introduce key concepts from learning theory using integral operators. Section D is dedicated to supporting results for the the theoretical proofs. The proofs of the two propositions from Section 5 of the main paper are detailed in Section E. In Section F, some additional results on projection learning and kernel-based projection learning are presented. Section G is dedicated to a detailed description of related work. Eventually, in section H, experimental details supplements are laid out.

A OVKs AND VV-RKHSs

First, we give the definition of an operator-valued kernel (OVK) and of its associated reproducing kernel Hilbert space (RKHS).

Definition A.1. Let $\mathcal{X}$ be a space on which a kernel can be defined and let $\mathcal{U}$ be a Hilbert space. An operator-valued kernel on $\mathcal{X} \times \mathcal{X}$ is a function $\mathsf{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{U})$ satisfying the two following conditions:

- Symmetry: for all $x, x' \in \mathcal{X}$, $\mathsf{K}(x, x') = \mathsf{K}(x', x)^{\#}$.
- Positivity: for all $n \in \mathbb{N}^*$, for all $(x_1, ..., x_n) \in \mathcal{X}^n$, for all $(u_1, ..., u_n) \in \mathcal{U}^n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\langle u_i, \mathsf{K}(x_i, x_j)u_j\rangle_{\mathcal{U}} \geq 0 \ .$$

The following theorem shows that given an OVK, it is possible to build a unique RKHS associated to it.

Theorem A.1. *(Senkene and Templeman, 1973; Carmeli et al., 2010) Let K be a given operator-valued kernel $\mathsf{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{U})$. For any $x \in \mathcal{X}$, we define K_x as*

$$\mathsf{K}_x : u \longmapsto \mathsf{K}_x u, \quad \textit{with} \quad \mathsf{K}_x u : x' \longmapsto \mathsf{K}(x', x)u. \tag{1}$$

There exists a unique Hilbert space $\mathcal{H}_{\mathsf{K}}$ of functions $h : \mathcal{X} \to \mathcal{U}$ satisfying the two conditions:

- *For all $x \in \mathcal{X}$, $\mathsf{K}_x \in \mathcal{L}(\mathcal{U}, \mathcal{H}_{\mathsf{K}})$.*
- *For all $h \in \mathcal{H}_{\mathsf{K}}$, $h(x) = \mathsf{K}_x^{\#} h$.*

The second condition is called the reproducing property; it implies that for all $x \in \mathcal{X}$, for all $u \in \mathcal{U}$ and for all $h \in \mathcal{H}_{\mathsf{K}}$,

$$\langle \mathsf{K}_x u, h\rangle_{\mathcal{H}_{\mathsf{K}}} = \langle u, h(x)\rangle_{\mathcal{U}}. \tag{2}$$

The Hilbert space $\mathcal{H}_{\mathsf{K}}$ is the RKHS associated to the kernel K.

The scalar product on $\mathcal{H}_{\mathsf{K}}$ between two functions $h_0 = \sum_{i=1}^{n} \mathsf{K}_{x_i} u_i$ and $h_1 = \sum_{j=1}^{n'} \mathsf{K}_{x'_j} u'_j$ with $x_i, x'_j \in \mathcal{X}$, $u_i, u'_j \in \mathcal{U}$, is defined as:

$$\langle h_0, h_1\rangle_{\mathcal{H}_{\mathsf{K}}} = \sum_{i=1}^{n}\sum_{j=1}^{n'}\langle u_i, \mathsf{K}(x_i, x'_j)u_j\rangle_{\mathcal{U}}.$$

The corresponding norm $\|\cdot\|_{\mathcal{H}_{\mathsf{K}}}$ is defined by $\|h\|^2_{\mathcal{H}_{\mathsf{K}}} = \langle h, h\rangle_{\mathcal{H}_{\mathsf{K}}}$.

This RKHS $\mathcal{H}_{\mathsf{K}}$ can be built by taking the closure of the set $\{\mathsf{K}_x u \,|x \in \mathcal{X},\, u \in \mathcal{U}\}$ with respect to the topology induced by $\|\cdot\|_{\mathcal{H}_{\mathsf{K}}}$.

Finally, we state the following Lemma which we use in the subsequent proofs. We now take $\mathcal{U} = \mathbb{R}^d$ in accordance with the use we make of vector-valued RKHSs (vv-RKHS) in the main paper.

Lemma A.1. *(Micchelli and Pontil, 2005) Let $\mathcal{H}_\mathsf{K} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ a vv-RKHS associated to a positive matrix-valued kernel K. Then we have for all $x \in \mathcal{X}$:*

$$\|h(x)\|_{\mathbb{R}^d} \leq \|h\|_{\mathcal{H}_\mathsf{K}} \, \|\mathsf{K}(x,x)\|^{1/2}_{\mathcal{L}(\mathbb{R}^d)}.$$

Additionally, since for all $x \in \mathcal{X}$, $h(x) = \mathsf{K}_x^\# h$, this implies that

$$\|\mathsf{K}_x\|_{\mathcal{L}(\mathbb{R}^d, \mathcal{H}_\mathsf{K})} = \|\mathsf{K}_x^\#\|_{\mathcal{L}(\mathcal{H}_\mathsf{K}, \mathbb{R}^d)} \leq \|\mathsf{K}(x,x)\|^{1/2}_{\mathcal{L}(\mathbb{R}^d)} \, . \tag{3}$$

B PROOFS FOR SECTION 4

B.1 Proof of Proposition 4.1 from the main paper

We recall first the proposition which corresponds to Proposition 4.1 of the main paper. Given $\mathsf{K} : \mathcal{X} \times \mathcal{X} \longmapsto \mathcal{L}(\mathbb{R}^d)$ an OVK with $\mathcal{H}_\mathsf{K} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ its associated vv-RKHS, we want to solve the following optimization problem

$$\min_{h \in \mathcal{H}_\mathsf{K}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i)) + \lambda \|h\|^2_{\mathcal{H}_\mathsf{K}}. \tag{4}$$

Proposition B.1. ***(Representer theorem.)*** *For ℓ continuous and convex with respect to its second argument, Problem (4) admits a unique minimizer $h^\lambda_\mathbf{z}$. Moreover there exists $\alpha \in \mathbb{R}^{d \times n}$ such that $h^\lambda_\mathbf{z} = \sum_{j=1}^n \mathsf{K}_{x_j} \alpha_j$.*

Proof. Since the loss is assumed to be continuous and convex with respect to the second argument, the objective $h \longmapsto \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda \|h\|^2_{\mathcal{H}_\mathsf{K}}$ is thus a continuous and strictly convex function on $\mathcal{H}_\mathsf{K}$ (strictly because $\lambda > 0$). As a consequence, it admits a unique minimizer on $\mathcal{H}_\mathsf{K}$ (Bauschke and Combettes, 2017), which we denote by $h^\lambda_\mathbf{z}$.

Let $\mathcal{U} := \left\{ h |\ h = \sum_{j=1}^n \mathsf{K}_{x_j} \alpha_j,\ \alpha \in \mathbb{R}^{d \times n} \right\}$. Since it is a closed subspace of $\mathcal{H}_\mathsf{K}$, $\mathcal{H}_\mathsf{K} = \mathcal{U} \oplus \mathcal{U}^\perp$ and we can decompose $h^\lambda_\mathbf{z}$ as $h^\lambda_\mathbf{z} = h^\lambda_{\mathbf{z},\mathcal{U}} + h^\lambda_{\mathbf{z},\mathcal{U}^\perp}$ with $(h^\lambda_{\mathbf{z},\mathcal{U}}, h^\lambda_{\mathbf{z},\mathcal{U}^\perp}) \in \mathcal{U} \times \mathcal{U}^\perp$. We recall that $\phi \in \mathsf{L}^2(\Theta)^d = (\phi_l)_{l=1}^d$ is the dictionary associated to Φ (see Definition 2.1 of the main paper) and we take the convention that for $\theta \in \Theta$, $\phi(\theta) = (\phi_l(\theta))_{l=1}^d \in \mathbb{R}^d$. Now, for all $i \in [n]$ and $\theta \in \Theta$, from Theorem A.1, we have:

$$(\Phi h^\lambda_\mathbf{z}(x_i))(\theta) = \langle \phi(\theta), h^\lambda_\mathbf{z}(x_i) \rangle_{\mathbb{R}^d} = \langle \mathsf{K}_{x_i} \phi(\theta), h^\lambda_\mathbf{z} \rangle_{\mathcal{H}_\mathsf{K}}.$$

Since $\mathsf{K}_{x_i} \phi(\theta) \in \mathcal{U}$, we get that

$$(\Phi h^\lambda_\mathbf{z}(x_i))(\theta) = \langle \mathsf{K}_{x_i} \phi(\theta), h^\lambda_{\mathbf{z},\mathcal{U}} \rangle_{\mathcal{H}_\mathsf{K}} = \langle \phi(\theta), h^\lambda_{\mathbf{z},\mathcal{U}}(x_i) \rangle_{\mathbb{R}^d} = (\Phi h^\lambda_{\mathbf{z},\mathcal{U}}(x_i))(\theta) \, .$$

Then, on the one hand the data-attach term in the criterion to minimize is unchanged when replacing $h^\lambda_\mathbf{z}$ by its projection $h^\lambda_{\mathbf{z},\mathcal{U}}$ onto $\mathcal{U}$. On the other hand, the penalty $\|h^\lambda_\mathbf{z}\|^2_{\mathcal{H}_\mathsf{K}}$ decreases if we replace $h^\lambda_\mathbf{z}$ by $h^\lambda_{\mathbf{z},\mathcal{U}}$, hence we must have $h^\lambda_\mathbf{z} = h^\lambda_{\mathbf{z},\mathcal{U}}$. □

B.2 Proof of Proposition 4.2 from the main paper

First, we recall the proposition which corresponds to Proposition 4.2 of the main paper. We want to solve the following (Problem (8) from the main paper):

$$\min_{\alpha \in \mathbb{R}^{d \times n}} \frac{1}{n} \left\| \mathbf{y} - \Phi_{(n)} \mathbf{K} \mathsf{vec}(\alpha) \right\|^2_{\mathsf{L}^2(\Theta)^n} + \lambda \langle \mathsf{vec}(\alpha), \mathbf{K} \mathsf{vec}(\alpha) \rangle_{\mathbb{R}^{dn}}. \tag{5}$$

Proposition B.2. ***(Ridge solution)*** *The minimum in Problem (5) is achieved by any $\alpha^* \in \mathbb{R}^{d \times n}$ verifying*

$$\left(\mathbf{K} (\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{K} \right) \mathsf{vec}(\alpha^*) := \mathbf{K} \Phi^\#_{(n)} \mathbf{y}. \tag{6}$$

Such α^* exists. Moreover if K is full rank then $\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{I}\right)$ is invertible and α^* is such that

$$\mathsf{vec}(\alpha^*) = \left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{I}\right)^{-1} \Phi^{\#}_{(n)}\mathbf{y}. \tag{7}$$

We then define the ridge estimator as $h^{\lambda}_{\mathbf{z}} := \sum_{j=1}^{n} \mathsf{K}_{x_j}\alpha^*_j$.

Proof. For $\boldsymbol{\alpha} \in \mathbb{R}^{dn}$ we consider the objective function

$$\frac{1}{n}\left\|\Phi_{(n)}\mathsf{K}\boldsymbol{\alpha}\right\|^2_{\mathsf{L}^2(\Theta)^n} - \frac{2}{n}\langle \mathbf{y}, \Phi_{(n)}\mathsf{K}\boldsymbol{\alpha}\rangle_{\mathsf{L}^2(\Theta)^n} + \lambda\langle\boldsymbol{\alpha}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} .$$

Up to an additional term not dependant on $\boldsymbol{\alpha}$, this corresponds to the objective function in Problem (5) where we have set $\boldsymbol{\alpha} = \mathsf{vec}(\alpha)$ to simplify the exposition.

Using that $(\Phi_{(n)})^{\#}\Phi_{(n)} = \Phi^{\#}_{(n)}\Phi_{(n)} = (\Phi^{\#}\Phi)_{(n)}$, that $\mathsf{K}^{\#} = \mathsf{K}$ and multiplying by n, we can consider as objective function

$$\begin{aligned} V(\boldsymbol{\alpha}) &:= \langle\boldsymbol{\alpha}, \mathsf{K}(\Phi^{\#}\Phi)_{(n)}\mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} - 2\langle\Phi^{\#}_{(n)}\mathbf{y}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} + n\lambda\langle\boldsymbol{\alpha}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} \\ &= \langle\boldsymbol{\alpha}, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} - 2\langle\Phi^{\#}_{(n)}\mathbf{y}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} . \end{aligned}$$

Let $\boldsymbol{\alpha}^* \in \mathbb{R}^{dn}$ be such that

$$\left(\mathsf{K}(\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{K}\right)\boldsymbol{\alpha}^* = \mathsf{K}\Phi^{\#}_{(n)}\mathbf{y} .$$

We want to prove that $\boldsymbol{\alpha}^*$ is then a solution to Problem (5). Observe now that

$$\begin{aligned} \langle\boldsymbol{\alpha}^*, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} &= \langle\boldsymbol{\alpha}, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)\boldsymbol{\alpha}^*\rangle_{\mathbb{R}^{dn}} \\ &= \langle\boldsymbol{\alpha}, \mathsf{K}\Phi^{\#}_{(n)}\mathbf{y}\rangle_{\mathbb{R}^{dn}} \\ &= \langle\Phi^{\#}_{(n)}\mathbf{y}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} . \end{aligned} \tag{8}$$

Using Equation (8), we deduce that

$$\begin{aligned} V(\boldsymbol{\alpha}) &= \langle\boldsymbol{\alpha}, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} - 2\langle\Phi^{\#}_{(n)}\mathbf{y}, \mathsf{K}\boldsymbol{\alpha}\rangle_{\mathbb{R}^{dn}} \\ &= \langle\boldsymbol{\alpha} - \boldsymbol{\alpha}^*, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\rangle_{\mathbb{R}^{dn}} \\ &\quad + \langle\boldsymbol{\alpha}^*, \mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)\boldsymbol{\alpha}^*\rangle_{\mathbb{R}^{dn}} . \end{aligned}$$

Since $\mathsf{K}\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\,\mathsf{I}\right)$ is a non-negative symmetric matrix, we conclude that $V(\boldsymbol{\alpha})$ is minimal at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

We now show that Equation (6) always has a solution $\boldsymbol{\alpha}^*$ in $\mathbb{R}^{dn}$ and conclude with the special case where K is full rank. Note that $\left(\mathsf{K}(\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{K}\right)$ is a positive symmetric matrix and its null space is exactly that of K. Hence it is bijective on the image of K, which shows that Equation (6) always has a solution. If K is moreover full rank then

$$\left((\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{I}\right) = \mathsf{K}^{-1}\left(\mathsf{K}(\Phi^{\#}\Phi)_{(n)}\mathsf{K} + n\lambda\mathsf{K}\right)$$

is also invertible and we can simplify by K on both sides of Equation (6) and obtain the claimed formula for $\boldsymbol{\alpha}^*$. Taking $\alpha^* \in \mathbb{R}^{d\times n}$ such that $\mathsf{vec}(\alpha^*) = \boldsymbol{\alpha}^*$ yields the desired results. □

C LEARNING THEORY AND INTEGRAL OPERATORS

This section is devoted to the study of Problem (4) for the functional square loss in the framework of integral operators (Caponnetto and De Vito, 2005, 2007; Smale and Zhou, 2007). In Section C.1 the expected risk and the excess risk are reformulated in terms of two operators of interest. In Section C.2, we introduce empirical approximations of those operators. From there we can reformulate the minimizer of the regularized empirical risk in terms of those empirical operators.

C.1 Excess risk reformulation

The first goal is to characterize the minimizer of the expected risk using two operators of interest as in (Caponnetto and De Vito, 2007). Using this characterization, a closed form for the excess risk of any regressor $\Phi \circ h$ is derived.

Considering the functional square loss, we recall the definition of the expected risk $\mathcal{R}$ of a regressor $f \in \mathcal{F}(\mathcal{X}, \mathsf{L}^2(\Theta))$

$$\mathcal{R}(f) := \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\rho}\left[\|\mathsf{Y} - f(\mathsf{X})\|^2_{\mathsf{L}^2(\Theta)}\right], \tag{9}$$

as well as that of its empirical risk on a sample $\mathbf{z}$

$$\widehat{\mathcal{R}}(f, \mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n} \|y_i - f(x_i)\|^2_{\mathsf{L}^2(\Theta)}. \tag{10}$$

Let us introduce $\mathsf{L}^2(\mathcal{Z}, \rho, \mathsf{L}^2(\Theta))$ the space of square integrable functions from $\mathcal{Z}$ to $\mathsf{L}^2(\Theta)$ with respect to the measure ρ endowed with the scalar product

$$\langle \psi_0, \psi_1 \rangle_\rho = \int_{\mathcal{Z}} \langle \psi_0(x,y), \psi_1(x,y)\rangle_{\mathsf{L}^2(\Theta)} \, \mathrm{d}\rho(x,y),$$

and its associated norm $\|.\|_\rho$. Then, the expected risk in Equation (9) of a regressor f can then be equivalently formulated as

$$\mathcal{R}(f) = \|f \circ X - Y\|^2_\rho, \tag{11}$$

where we have defined $X : (x,y) \in \mathcal{Z} \longmapsto x \in \mathcal{X}$ and $Y \in \mathsf{L}^2(\mathcal{Z}, \rho, \mathsf{L}^2(\Theta))$ as $Y : (x,y) \in \mathcal{Z} \longmapsto y \in \mathsf{L}^2(\Theta)$.

We wish to study the excess risk of any regressor of the form $f = \Phi \circ h$. To that end, we define the operator $\mathsf{A}_\Phi : \mathcal{H}_\mathsf{K} \longrightarrow \mathsf{L}^2(\mathcal{Z}, \rho, \mathsf{L}^2(\Theta))$ as

$$\mathsf{A}_\Phi : h \longmapsto \mathsf{A}_\Phi h \quad \text{with} \quad (\mathsf{A}_\Phi h) : (x,y) \in \mathcal{Z} \longmapsto \Phi \mathsf{K}_x^\# h. \tag{12}$$

We can reformulate the expected risk in terms of A_Φ for any $h \in \mathcal{H}_\mathsf{K}$,

$$\|\mathsf{A}_\Phi h - Y\|^2_\rho = \int_{\mathcal{Z}} \|\Phi \mathsf{K}_x^\# h - y\|^2_{\mathsf{L}^2(\Theta)} \, \mathrm{d}\rho(x,y) = \int_{\mathcal{Z}} \|\Phi h(x) - y\|^2_{\mathsf{L}^2(\Theta)} \, \mathrm{d}\rho(x,y) = \mathcal{R}(\Phi \circ h). \tag{13}$$

We now define T_Φ as $\mathsf{T}_\Phi := \mathsf{A}_\Phi^\# \mathsf{A}_\Phi$.

Lemma C.1. *Assume that there exists $h_{\mathcal{H}_\mathsf{K}} \in \mathcal{H}_\mathsf{K}$ such that*

$$h_{\mathcal{H}_\mathsf{K}} := \inf_{h \in \mathcal{H}_\mathsf{K}} \mathcal{R}(\Phi \circ h).$$

Then, for all $h \in \mathcal{H}_\mathsf{K}$,

$$\langle h, \mathsf{T}_\Phi h_{\mathcal{H}_\mathsf{K}} - \mathsf{A}_\Phi^\# Y \rangle_{\mathcal{H}_\mathsf{K}} = 0; \tag{14}$$

or equivalently:

$$\mathsf{T}_\Phi h_{\mathcal{H}_\mathsf{K}} = \mathsf{A}_\Phi^\# Y, \tag{15}$$

with $Y \in \mathsf{L}^2(\mathcal{Z}, \rho, \mathsf{L}^2(\Theta))$ denoting the function $Y : (x,y) \longmapsto y$.

Proof. We use the formulation of the expected risk from Equation (13). The function $h \longmapsto \mathcal{R}(\Phi \circ h) = \|\mathsf{A}_\Phi h - Y\|^2_\rho$ is convex as a convex function composed with an affine mapping. Its differential is given by

$$D\mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}})(h) = 2\langle \mathsf{A}_\Phi h, \mathsf{A}_\Phi h_{\mathcal{H}_\mathsf{K}} - Y\rangle_\rho = 2\langle h, \mathsf{A}_\Phi^\# \mathsf{A}_\Phi h_{\mathcal{H}_\mathsf{K}} - \mathsf{A}_\Phi^\# Y\rangle_{\mathcal{H}_\mathsf{K}} = 2\langle h, \mathsf{T}_\Phi h_{\mathcal{H}_\mathsf{K}} - \mathsf{A}_\Phi^\# Y\rangle_{\mathcal{H}_\mathsf{K}}.$$

We then must have for all $h \in \mathcal{H}_\mathsf{K}$,

$$\langle h, \mathsf{T}_\Phi h_{\mathcal{H}_\mathsf{K}} - \mathsf{A}_\Phi^\# Y \rangle_{\mathcal{H}_\mathsf{K}} = 0.$$

□

Using the formulation of the expected risk from Equation (13) as well as the characterization of $h_{\mathcal{H}_\mathsf{K}}$ in Equation (14), for any $h \in \mathcal{H}_\mathsf{K}$, we can then reformulate the excess risk of h as a distance in $\mathcal{H}_\mathsf{K}$ between h and $h_{\mathcal{H}_\mathsf{K}}$ taken through the operator T_Φ.

Lemma C.2. *We have that for any $h \in \mathcal{H}_\mathsf{K}$,*

$$\mathcal{R}(\Phi \circ h) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) = \|\sqrt{\mathsf{T}_\Phi}(h - h_{\mathcal{H}_\mathsf{K}})\|^2_{\mathcal{H}_\mathsf{K}}. \tag{16}$$

Proof.

$$\begin{aligned}\mathcal{R}(\Phi \circ h) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) &= \|\mathsf{A}_\Phi h - Y\|^2_\rho - \|\mathsf{A}_\Phi h_{\mathcal{H}_\mathsf{K}} - Y\|^2_\rho \\ &= \|\mathsf{A}_\Phi(h - h_{\mathcal{H}_\mathsf{K}})\|^2_\rho + 2\langle \mathsf{A}_\Phi(h - h_{\mathcal{H}_\mathsf{K}}), \mathsf{A}_\Phi h_{\mathcal{H}_\mathsf{K}} - Y\rangle_\rho \\ &= \|\mathsf{A}_\Phi(h - h_{\mathcal{H}_\mathsf{K}})\|^2_\rho,\end{aligned}$$

where we have used Equation (14). Since we have the following polar decomposition $\mathsf{A}_\Phi = \mathsf{U}\sqrt{\mathsf{A}^\#_\Phi \mathsf{A}_\Phi} = \mathsf{U}\sqrt{\mathsf{T}_\Phi}$ with U a partial isometry from the closure of $\mathrm{Im}(\sqrt{\mathsf{T}_\Phi})$ onto the closure of $\mathrm{Im}(\mathsf{A}_\Phi)$,

$$\|\mathsf{A}_\Phi(h - h_{\mathcal{H}_\mathsf{K}})\|_\rho = \|\mathsf{U}\sqrt{\mathsf{T}_\Phi}(h - h_{\mathcal{H}_\mathsf{K}})\|_\rho = \|\sqrt{\mathsf{T}_\Phi}(h - h_{\mathcal{H}_\mathsf{K}})\|_{\mathcal{H}_\mathsf{K}}.$$

□

Such reformulation enables us to decompose the excess risk in terms that we can easily control using concentration inequalities in Hilbert spaces.

C.2 Empirical approximations and closed form solutions

We now define empirical approximations of the operators A_Φ and T_Φ. Using those approximations, we can derive a closed-form for the minimizer of the regularized expected risk. We utilize that closed-form to bound the excess risk in the subsequent proof.

To define those approximations, we need to precise the integral expressions of $\mathsf{A}^\#_\Phi$ and T_Φ. This is the object of the following lemma, which is almost a restatement of Proposition 1 from Caponnetto and De Vito (2005), as a consequence, we do not re-write the proof here.

Let us define for all $x \in \mathcal{X}$ the operators $\mathsf{K}_{x,\Phi} := \mathsf{K}_x\Phi^\#$ and $\mathsf{T}_{x,\Phi} := \mathsf{K}_{x,\Phi}\mathsf{K}^\#_{x,\Phi}$.

Lemma C.3. *For $\psi \in \mathsf{L}^2(\mathcal{Z}, \rho, \mathsf{L}^2(\Theta))$, the adjoint of A_Φ applied to ψ is given by*

$$\mathsf{A}^\#_\Phi \psi = \int_\mathcal{Z} \mathsf{K}_{x,\Phi}\psi(x,y)\ \mathrm{d}\rho(x,y), \tag{17}$$

with the integral converging in $\mathcal{H}_\mathsf{K}$. And $\mathsf{A}^\#_\Phi\mathsf{A}_\Phi$ is the Hilbert Schmidt operator on $\mathcal{H}_\mathsf{K}$ given by

$$\mathsf{A}^\#_\Phi\mathsf{A}_\Phi = \mathsf{T}_\Phi = \int_\mathcal{X} \mathsf{T}_{x,\Phi}\ \mathrm{d}\rho_\mathsf{X}(x), \tag{18}$$

with the integral converging in $\mathcal{L}_2(\mathcal{H}_\mathsf{K})$.

Empirical approximations of the operators A_Φ and T_Φ can then straightforwardly be set as

$$\begin{aligned}\mathsf{A}^\#_{\mathbf{x},\Phi}\mathbf{w} &= \frac{1}{n}\sum_{i=1}^n \mathsf{K}_{x_i,\Phi}w_i, \quad \mathbf{w} = (w_i)_{i=1}^n \in \mathsf{L}^2(\Theta)^n. \\ (\mathsf{A}_{\mathbf{x},\Phi}h)_i &= \mathsf{K}^\#_{x_i,\Phi}h = \Phi h(x_i), \quad h \in \mathcal{H}_\mathsf{K}, \quad \forall i \in [n]. \\ \mathsf{T}_{\mathbf{x},\Phi} &= \mathsf{A}^\#_{\mathbf{x},\Phi}\mathsf{A}_{\mathbf{x},\Phi} = \frac{1}{n}\sum_{i=1}^n \mathsf{T}_{x_i,\Phi}.\end{aligned}$$

Defining the regularized empirical risk of $\Phi \circ h$ for any $h \in \mathcal{H}_{\mathsf{K}}$ as

$$\widehat{\mathcal{R}}^{\lambda}(\Phi \circ h, \mathbf{z}) := \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda \|h\|^2_{\mathcal{H}_{\mathsf{K}}} = \frac{1}{n}\sum_{i=1}^{n} \|\mathsf{K}^{\#}_{x_i,\Phi} h - y_i\|^2_{\mathsf{L}^2(\Theta)} + \lambda \|h\|^2_{\mathcal{H}_{\mathsf{K}}},$$

the following closed form for its minimizer can be derived.

Lemma C.4. *There exists a unique minimizer $h^{\lambda}_{\mathbf{z}}$ of $h \in \mathcal{H}_{\mathsf{K}} \longmapsto \widehat{\mathcal{R}}^{\lambda}(\Phi \circ h, \mathbf{z})$ which is given by*

$$h^{\lambda}_{\mathbf{z}} := (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1} \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y} \ . \tag{19}$$

Proof. Since $\lambda > 0$, $h \longmapsto \widehat{\mathcal{R}}^{\lambda}(\Phi \circ h, \mathbf{z})$ is strictly convex. As it is continuous, there exist a unique minimizer which can be found by setting the differential to zero.

$$\begin{aligned} D\widehat{\mathcal{R}}^{\lambda}(\Phi \circ h_0, \mathbf{z})(h_1) &= \frac{2}{n}\sum_{i=1}^{n} \langle \mathsf{K}^{\#}_{x_i,\Phi} h_0 - y_i, \mathsf{K}^{\#}_{x_i,\Phi} h_1 \rangle_{\mathsf{L}^2(\Theta)} + 2\lambda \langle h_0, h_1 \rangle_{\mathcal{H}_{\mathsf{K}}} \\ &= 2\left\langle \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{T}_{x_i,\Phi} + \lambda \right) h_0 - \frac{1}{n}\sum_{i=1}^{n} \mathsf{K}_{x_i,\Phi} y_i, h_1 \right\rangle_{\mathcal{H}_{\mathsf{K}}} \\ &= 2\langle (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I}) h_0 - \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y}, h_1 \rangle_{\mathcal{H}_{\mathsf{K}}}. \end{aligned}$$

As a consequence, $h^{\lambda}_{\mathbf{z}}$ is characterized by

$$(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I}) h^{\lambda}_{\mathbf{z}} - \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y} = 0.$$

Since $\mathsf{T}_{\mathbf{x},\Phi}$ is positive and $\lambda > 0$, $(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})$ is invertible and thus

$$h^{\lambda}_{\mathbf{z}} = (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1} \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y}.$$

□

Importantly, $h^{\lambda}_{\mathbf{z}}$ is the same object as the ridge estimator from Proposition B.2 which is why we have used the same notation. The representation in terms of operators introduced above is however needed to carry out an excess risk analysis.

D SUPPORTING RESULTS FOR SECTION E

This section is dedicated to technical results on which the proofs in Section E rely.

D.1 Riesz families and projection operator

The proofs in the next section strongly relies on general inequalities on Riesz families and on the associated projection operator Φ, that we state and prove in this section.

Using the definition of a Riesz family we have

Lemma D.1. *Let $\phi := (\phi_1, ..., \phi_d)$ be a Riesz family, let Φ be its associated projection operator (see Definition 2.1 from the main paper). Then*

$$\|\Phi\|_{\mathcal{L}(\mathbb{R}^d, \mathsf{L}^2(\Theta))} \leq C_{\phi} \tag{20}$$

$$\|\Phi^{\#}\|_{\mathcal{L}(\mathsf{L}^2(\Theta), \mathbb{R}^d)} \leq C_{\phi} \tag{21}$$

$$\|\Phi^{\#}\Phi\|_{\mathcal{L}(\mathbb{R}^d)} \leq C^2_{\phi} \ . \tag{22}$$

Proof. Equation (20) is a direct consequence of the definition of a Riesz family (Definition 5.1 from the main paper). Since the operator Φ is bounded, $\|\Phi^{\#}\|_{\mathcal{L}(\mathsf{L}^2(\Theta), \mathbb{R}^d)} = \|\Phi\|_{\mathcal{L}(\mathbb{R}^d, \mathsf{L}^2(\Theta))}$ implying Equation (21). Finally combining the two inequalities yields Equation (22). □

D.2 Bound on Hilbert-Schmidt norm of $\mathsf{T}_{x,\Phi}$

In the subsequent proof, we need to derive concentration results on $\mathsf{T}_{\mathbf{x},\Phi}$. To that end, we need to bound the Hilbert-Schimdt norm of $\mathsf{T}_{x,\Phi}$.

For all $x \in \mathcal{X}$, we recall the definition of the following operators

- $\mathsf{K}_{x,\Phi} : \mathsf{L}^2(\Theta) \longrightarrow \mathcal{H}_{\mathsf{K}}$ is defined by $\mathsf{K}_{x,\Phi} := \mathsf{K}_x \Phi^{\#}$ with K_x as defined in Equation (1).
- $\mathsf{T}_{x,\Phi} : \mathcal{H}_{\mathsf{K}} \longrightarrow \mathcal{H}_{\mathsf{K}}$ is defined as $\mathsf{T}_{x,\Phi} := \mathsf{K}_{x,\Phi}\mathsf{K}_{x,\Phi}^{\#}$.

Observe that $\mathsf{T}_{x,\Phi}$ is of finite rank and positive. We can then deduce the following bound on its Hilbert-Schmidt norm.

Lemma D.2. *Assume that there exists $\kappa \geq 0$ such that for all $x \in \mathcal{X}$,*

$$\|\mathsf{K}(x,x)\|_{\mathcal{L}(\mathbb{R}^d)} \leq \kappa, \tag{23}$$

then for all $x \in \mathcal{X}$,

$$\|\mathsf{T}_{x,\Phi}\|_{\mathcal{L}_2(\mathcal{H}_{\mathsf{K}})} \leq \sqrt{d}\kappa C_\phi^2. \tag{24}$$

Proof. For all $x \in \mathcal{X}$, $\mathrm{Rank}(\mathsf{T}_{x,\Phi}) \leq d$. Let $(e_l)_{l=1}^{\mathrm{Rank}(\mathsf{T}_{x,\Phi})}$ be an orthonormal basis of $\mathrm{Im}(\mathsf{T}_{x,\Phi})$. We complete it to $(e_l)_{l\in\mathbb{N}^*}$ to be an orthonormal basis of $\mathcal{H}_{\mathsf{K}}$. Since $\mathrm{Im}(\mathsf{T}_{x,\Phi})$ is a finite dimensional subspace of $\mathcal{H}_{\mathsf{K}}$ and $\mathsf{T}_{x,\Phi}$ is self adjoint, we have that $\mathrm{Im}(\mathsf{T}_{x,\Phi}) = \mathrm{Ker}(\mathsf{T}_{x,\Phi})^{\perp}$. As a consequence, for all $l > \mathrm{Rank}(\mathsf{T}_{x,\Phi})$, $\mathsf{T}_{x,\Phi}e_l = 0$, which implies

$$\|\mathsf{T}_{x,\Phi}\|^2_{\mathcal{L}_2(\mathcal{H}_{\mathsf{K}})} = \sum_{l=1}^{\mathrm{Rank}(\mathsf{T}_{x,\Phi})} \langle \mathsf{T}_{x,\Phi}e_l, \mathsf{T}_{x,\Phi}e_l\rangle_{\mathcal{H}_{\mathsf{K}}} = \sum_{l=1}^{\mathrm{Rank}(\mathsf{T}_{x,\Phi})} \langle \mathsf{K}_x^{\#}e_l, \Phi^{\#}\Phi\mathsf{K}(x,x)\Phi^{\#}\Phi\mathsf{K}_x^{\#}e_l\rangle_{\mathbb{R}^d}.$$

Using Cauchy-Schwartz in the previous expression along with Equation (22), Equation (23) and Equation (3) we have that

$$\|\mathsf{T}_{x,\Phi}\|^2_{\mathcal{L}_2(\mathcal{H}_{\mathsf{K}})} \leq C_\phi^4\kappa \sum_{l=1}^{\mathrm{Rank}(\mathsf{T}_{x,\Phi})} \|\mathsf{K}_x^{\#}e_l\|^2_{\mathbb{R}^d} \leq C_\phi^4\kappa^2\mathrm{Rank}(\mathsf{T}_{x,\Phi}) \leq dC_\phi^4\kappa^2,$$

which achieves the proof. □

D.3 Concentration results

We now state two concentration inequalities that we use to control the different terms in our decomposition of the excess risk in Section E. We also introduce Lemma D.5 which we use to deduce concentration properties of $\sqrt{\mathsf{T}_{\mathbf{x},\Phi}}$ from concentration properties of $\mathsf{T}_{\mathbf{x},\Phi}$.

The following is a direct consequence of a Bernstein inequality for independent random variables in a separable Hilbert space—see Proposition 3.3.1 in (Yurinsky, 1995) or Theorem 3 in (Pinelis and Sakhanenko, 1986). It corresponds to Proposition 2 in (Caponnetto and De Vito, 2007).

Lemma D.3. *Let ξ be a random variable taking its values in a real separable Hilbert space $\mathcal{K}$ such that there exist $H \geq 0$ and $\sigma \geq 0$ such that*

$$\|\xi\|_{\mathcal{K}} \leq \frac{H}{2} \text{ almost surely, and}$$
$$\mathbb{E}[\|\xi\|^2_{\mathcal{K}}] \leq \sigma^2.$$

Let $n \in \mathbb{N}$ and $(\xi_1, ..., \xi_n)$ be i.i.d. realizations of ξ. Let $0 < \eta < 1$, then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi]\right\|_{\mathcal{K}} \leq 2\left(\frac{H}{n} + \frac{\sigma}{\sqrt{n}}\right)\log\frac{2}{\eta}\right] \geq 1 - \eta.$$

We introduce a variant of the previous Lemma for independent variables that are not necessarily identically distributed. It stems from the same Bernstein inequality (Pinelis and Sakhanenko, 1986; Yurinsky, 1995). We need it to treat the case where the output functions are partially observed in Section E. The proof is almost similar to that of Lemma D.3 which can be found in Caponnetto and De Vito (2007), so we do not rewrite it here.

Lemma D.4. *Let $(\mathsf{U}_i)_{i=1}^n$ be independent random variables taking their values in a real separable Hilbert space $\mathcal{K}$ such that for all $i \in [n]$*

$$\mathbb{E}[\mathsf{U}_i] = 0,$$

and there exist $H \geq 0$ and $\sigma \geq 0$ such that for all $i \in [n]$

$$\|\mathsf{U}_i\|_{\mathcal{K}} \leq \frac{H}{2} \text{ almost surely, and}$$
$$\mathbb{E}[\|\mathsf{U}_i\|_{\mathcal{K}}^2] \leq \sigma^2.$$

Let $0 < \eta < 1$, then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathsf{U}_i\right\|_{\mathcal{K}} \leq 2\left(\frac{H}{n} + \frac{\sigma}{\sqrt{n}}\right)\log\frac{2}{\eta}\right] \geq 1-\eta.$$

Finally, we need the following result to state concentration results on the square root of Hilbert-Schmidt operators. It corresponds to Theorem X.1.1 in Bhatia (1997) where it is stated for positive symmetric matrices. Their proof remains however fully valid for positive bounded operators defined on real separable Hilbert spaces.

Lemma D.5. *Let $\mathcal{K}$ be a real separable Hilbert space, let $\mathsf{A}, \mathsf{B} \in \mathcal{L}(\mathcal{K})$ be two positive operators. Then, we have*

$$\|\sqrt{\mathsf{A}} - \sqrt{\mathsf{B}}\|_{\mathcal{L}(\mathcal{K})} \leq \sqrt{\|\mathsf{A}-\mathsf{B}\|_{\mathcal{L}(\mathcal{K})}}.$$

E PROOFS FOR SECTION 5

E.1 Proof of Proposition 5.1 from the main paper

We recall the assumptions, as well as the proposition itself which corresponds to Proposition 5.1 of the main paper.

Assumption E.1. *K is a vector-valued continuous kernel and there exists $\kappa > 0$ such that for $x \in \mathcal{X}$, $\|\mathsf{K}(x,x)\|_{\mathcal{L}(\mathbb{R}^d)} \leq \kappa$.*

Remark. We suppose that κ is independant from d. This is for instance the case if for $x \in \mathcal{X}$, $\mathsf{K}(x,x)$ is diagonal or block diagonal with bounded coefficients. More generally, we can rely on the fact that κ is bounded by the maximal $\|\cdot\|_1$-norm of the columns of $\mathsf{K}(x,x)$, which can easily be imposed to be be independent of d.

Assumption E.2. *The dictionary ϕ is a normed Riesz family in $\mathsf{L}^2(\Theta)$ with upper constant C_ϕ.*

Remark. We do not use the lower constant c_ϕ.

Assumption E.3. *There exist $h_{\mathcal{H}_\mathsf{K}} \in \mathcal{H}_\mathsf{K}$ such that $h_{\mathcal{H}_\mathsf{K}} = \inf_{h \in \mathcal{H}_\mathsf{K}} \mathcal{R}(\Phi \circ h)$.*

Remark. This is a standard assumption (Caponnetto and De Vito, 2007; Baldassarre et al., 2012; Li et al., 2019), it implies the existence of a ball of radius $R > 0$ in $\mathcal{H}_\mathsf{K}$ containing $h_{\mathcal{H}_\mathsf{K}}$, as a consequence

$$\|h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} \leq R. \tag{25}$$

Assumption E.4. *There exists $L \geq 0$ such that for all $\theta \in \Theta$, almost surely $|\mathsf{Y}(\theta)| \leq L$.*

Remark. This implies that almost surely $\|\mathsf{Y}\|_{\mathsf{L}^2(\Theta)} \leq L$.

We now state Proposition 5.1 of the main paper.

Proposition E.1. *Let $0 < \eta < 1$, taking*

$$\lambda = \lambda_n^*(\eta/2) := 6\kappa C_\phi^2 \frac{\log(4/\eta)\sqrt{d}}{\sqrt{n}},$$

with probability at least $1-\eta$

$$\mathcal{R}(\Phi \circ h_{\mathbf{z}}^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq 27\left(\frac{B_0}{\sqrt{d}} + B_1\sqrt{d}\right)\frac{\log(4/\eta)}{\sqrt{n}},$$

with $B_0 := (L + \sqrt{\kappa}C_\phi R)^2$ and $B_1 := \kappa C_\phi^2 R^2$.

E.1.1 Concentration results

Lemma E.1. *Let $0 < \eta < 1$, then with probability at least $1 - \eta$*

$$\|\mathsf{A}^{\#}_{\mathbf{x},\Phi}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} \leq \delta_1(n,\eta),$$

with δ_1 defined as

$$\delta_1(n,\eta) := 6(\sqrt{\kappa} C_\phi L + \kappa C_\phi^2 R)\frac{\log(2/\eta)}{\sqrt{n}}. \tag{26}$$

Proof. Let us define the function $\xi_1 : \mathcal{Z} \longrightarrow \mathcal{H}_\mathsf{K}$ as $\xi_1 : (x,y) \longmapsto \mathsf{K}_{x,\Phi}(y - \Phi h_{\mathcal{H}_\mathsf{K}}(x)) = \mathsf{K}_{x,\Phi}(y - \mathsf{K}^{\#}_{x,\Phi} h_{\mathcal{H}_\mathsf{K}})$.

Observe that

$$\frac{1}{n}\sum_{i=1}^{n} \xi_1(x_i, y_i) = \mathsf{A}^{\#}_{\mathbf{x},\Phi}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}},$$

and using Equation (15), that

$$\mathbb{E}_{\mathsf{X},\mathsf{Y}\sim\rho}\left[\xi_1(\mathsf{X},\mathsf{Y})\right] = \int_{\mathcal{Z}} \mathsf{K}_{x,\Phi} y \; \mathrm{d}\rho(x,y) - \left(\int_{\mathcal{Z}} \mathsf{K}_{x,\Phi}\mathsf{K}^{\#}_{x,\Phi} \; \mathrm{d}\rho(x,y)\right) h_{\mathcal{H}_{\mathcal{H}_\mathsf{K}}} = \mathsf{A}^{\#}_{\Phi} Y - \mathsf{T}_\Phi h_{\mathcal{H}_\mathsf{K}} = 0.$$

The aim is now to apply the Bernstein inequality of Lemma D.3 to the random variable (RV) $\xi_1(\mathsf{X},\mathsf{Y})$. First, we have almost surely

$$\begin{aligned}
\|\xi_1(\mathsf{X},\mathsf{Y})\|_{\mathcal{H}_\mathsf{K}} = \|\mathsf{K}_{\mathsf{X},\Phi}(\mathsf{Y} - \Phi h_{\mathcal{H}_\mathsf{K}}(\mathsf{X}))\|_{\mathcal{H}_\mathsf{K}} &\leq \|\mathsf{K}_{\mathsf{X},\Phi}\|_{\mathcal{L}(\mathsf{L}^2(\Theta),\mathcal{H}_\mathsf{K})}\|\mathsf{Y} - \Phi h_{\mathcal{H}_\mathsf{K}}(\mathsf{X}))\|_{\mathsf{L}^2(\Theta)} \\
&\leq \sqrt{\kappa} C_\phi(\|\mathsf{Y}\|_{\mathsf{L}^2(\Theta)} + \|\mathsf{K}^{\#}_{\mathsf{X},\Phi} h\|_{\mathsf{L}^2(\Theta)}) \\
&\leq \sqrt{\kappa} C_\phi(L + \sqrt{\kappa} C_\phi R),
\end{aligned} \tag{27}$$

where we have used the inequality $\|\mathsf{K}_{x,\Phi}\|_{\mathcal{L}(\mathsf{L}^2(\Theta),\mathcal{H}_\mathsf{K})} = \|\mathsf{K}^{\#}_{x,\Phi}\|_{\mathcal{L}(\mathsf{L}^2(\Theta),\mathcal{H}_\mathsf{K})} \leq \sqrt{\kappa} C_\phi$ (immediate consequence of Equations (20) and (3)), as well as Assumptions E.4 and E.3.

Equation (27) also implies

$$\mathbb{E}_{\mathsf{X},\mathsf{Y}\sim\rho}[\|\xi_1(\mathsf{X},\mathsf{Y})\|^2_{\mathcal{H}_\mathsf{K}}] \leq \kappa C_\phi(L + \sqrt{\kappa} C_\phi R)^2.$$

Hence we can apply Lemma D.3, yielding that with probability at least $1 - \eta$,

$$\begin{aligned}
\|\mathsf{A}^{\#}_{\mathbf{x},\Phi}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} &\leq (\sqrt{\kappa} C_\phi L + \kappa C_\phi^2 R)\log(2/\eta)\left(\frac{4}{n} + \frac{2}{\sqrt{n}}\right) \\
&\leq 6(\sqrt{\kappa} C_\phi L + \kappa C_\phi^2 R)\frac{\log(2/\eta)}{\sqrt{n}}.
\end{aligned}$$

□

Lemma E.2. *Let $0 < \eta < 1$, then with probability at least $1 - \eta$*

$$\|\mathsf{T}_{\mathbf{x},\Phi} - \mathsf{T}_\Phi\|_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})} \leq \delta_2(n,d,\eta),$$

with δ_2 defined as

$$\delta_2(n,d,\eta) := 6\kappa C_\phi^2 \frac{\log(2/\eta)\sqrt{d}}{\sqrt{n}}. \tag{28}$$

Proof. We introduce the $\xi_2 : \mathcal{Z} \longrightarrow \mathcal{L}_2(\mathcal{H}_\mathsf{K})$ as $\xi_2 : x, y \longmapsto \mathsf{T}_{x,\Phi}$.

We have that

$$\mathbb{E}_{\mathsf{X},\mathsf{Y}\sim\rho}[\xi_2(\mathsf{X},\mathsf{Y})] = \int_{\mathcal{X}} \mathsf{T}_{x,\Phi}\ \mathrm{d}\rho_\mathsf{X}(x) = \mathsf{T}_\Phi.$$

And from Equation (24), we have almost surely

$$\|\xi_2(\mathsf{X},\mathsf{Y})\|_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})} \leq \kappa C_\phi^2 \sqrt{d},$$

which implies as well

$$\mathbb{E}_{\mathsf{X},\mathsf{Y}\sim\rho}[\|\xi_2(\mathsf{X},\mathsf{Y})\|^2_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})}] \leq \kappa^2 C_\phi^4 d.$$

Since K is continuous and $\mathcal{X}$ is separable, $\mathcal{H}_\mathsf{K}$ is separable. As a consequence the space $\mathcal{L}_2(\mathcal{H}_\mathsf{K})$ is also separable, we can thus apply Lemma D.3, yielding that with probability at least $1-\eta$,

$$\begin{aligned}\|\mathsf{T}_{\mathbf{x},\Phi} - \mathsf{T}_\Phi\|_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})} &\leq \kappa C_\phi^2 \sqrt{d} \log(4/\eta)\left(\frac{4}{n} + \frac{2}{\sqrt{n}}\right) \\ &\leq 6\kappa C_\phi^2 \sqrt{d}\frac{\log(2/\eta)}{\sqrt{n}}.\end{aligned}$$

□

Lemma E.3. *Let $0 < \eta < 1$, then with probability at least $1-\eta$ the two following inequalities hold:*

$$\begin{aligned}\|\mathsf{A}^{\#}_{\mathbf{x},\Phi}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} &\leq \delta_1(n, \eta/2) \\ \|\mathsf{T}_{\mathbf{x},\Phi} - \mathsf{T}_\Phi\|_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})} &\leq \delta_2(n, d, \eta/2),\end{aligned}$$

with δ_1 and δ_2 defined respectively in Equations (26) *and* (28)*.*

Proof. This is a union bound using Lemma E.1 and Lemma E.2. □

E.1.2 Proof

We are now ready to prove Proposition E.1. We follow the same proof strategy as (Baldassarre et al., 2012). To that end, we first prove the following intermediate proposition of which Proposition E.1 is a direct consequence.

Proposition E.2. *Let $0 < \eta < 1$, provided λ is taken such that*

$$\lambda \geq 6\kappa C_\phi^2 \frac{\log(4/\eta)\sqrt{d}}{\sqrt{n}} = \delta_2(n, d, \eta/2), \tag{29}$$

we have with probability at least $1-\eta$ that

$$\mathcal{R}(\Phi \circ h^\lambda_\mathbf{z}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq \frac{9}{2}\left(\frac{36(\sqrt{\kappa}C_\phi L + \kappa C_\phi^2 R)^2 \log(4/\eta)^2}{\lambda n} + \lambda R^2\right). \tag{30}$$

Proof. We introduce h^λ as

$$h^\lambda := (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1} \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}. \tag{31}$$

We consider the following decomposition of the risk using Equation (16),

$$\begin{aligned}\mathcal{R}(\Phi \circ h^\lambda_\mathbf{z}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) &= \|\sqrt{\mathsf{T}_\Phi}(h^\lambda_\mathbf{z} - h_{\mathcal{H}_\mathsf{K}})\|^2_{\mathcal{H}_\mathsf{K}} \\ &\leq 2\|\sqrt{\mathsf{T}_\Phi}(h^\lambda_\mathbf{z} - h^\lambda)\|^2_{\mathcal{H}_\mathsf{K}} + 2\|\sqrt{\mathsf{T}_\Phi}(h^\lambda - h_{\mathcal{H}_\mathsf{K}})\|^2_{\mathcal{H}_\mathsf{K}}.\end{aligned} \tag{32}$$

We first bound the term $\|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_\mathsf{K}}$. Using the expression of $h_{\mathbf{z}}^\lambda$ from Lemma C.4, we have that

$$\begin{aligned}\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda) &= \sqrt{\mathsf{T}_{\mathbf{x},\Phi}}(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}(\mathsf{A}_{\mathbf{x},\Phi}^{\#}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}) \\ &\quad + (\sqrt{\mathsf{T}_\Phi} - \sqrt{\mathsf{T}_{\mathbf{x},\Phi}})(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}(\mathsf{A}_{\mathbf{x},\Phi}^{\#}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}).\end{aligned} \tag{33}$$

Since for all $a \geq 0$, $\frac{\sqrt{a}}{a+\lambda} \leq \frac{1}{2\sqrt{\lambda}}$, since $\mathsf{T}_{\mathbf{x},\Phi}$ is positive, by spectral theorem we have that

$$\|\sqrt{\mathsf{T}_{\mathbf{x},\Phi}}(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})} \leq \max_{a \in \mathrm{Sp}(\mathsf{T}_{\mathbf{x},\Phi})} \frac{\sqrt{a}}{a+\lambda} \leq \max_{a \in \mathbb{R}_+} \frac{\sqrt{a}}{a+\lambda} \leq \frac{1}{2\sqrt{\lambda}}, \tag{34}$$

where $\mathrm{Sp}(\mathsf{T}_{\mathbf{x},\Phi})$ denotes the spectrum of $\mathsf{T}_{\mathbf{x},\Phi}$.

Similarily, since for all $a \geq 0$, $\frac{1}{a+\lambda} \leq \frac{1}{\lambda}$, we have as well

$$\|(\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})} \leq \frac{1}{\lambda}.$$

Taking the norm in Equation (33), applying Minkowski's inequality and using Lemma D.5 as well as the last two displays yields

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} \leq \|\mathsf{A}_{\mathbf{x},\Phi}^{\#}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi} h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} \left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\|\mathsf{T}_\Phi - \mathsf{T}_{\mathbf{x},\Phi}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})}}}{\lambda} \right). \tag{35}$$

Now dealing with the term on the right-hand side in Equation (32), using the definition of h^λ in Equation (31), we have that

$$\begin{aligned}\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^\lambda) &= \sqrt{\mathsf{T}_\Phi}(\mathsf{I} - (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\mathsf{T}_{\mathbf{x},\Phi}) h_{\mathcal{H}_\mathsf{K}} \\ &= (\sqrt{\mathsf{T}_\Phi} - \sqrt{\mathsf{T}_{\mathbf{x},\Phi}})(\mathsf{I} - (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\mathsf{T}_{\mathbf{x},\Phi}) h_{\mathcal{H}_\mathsf{K}} \\ &\quad + \sqrt{\mathsf{T}_{\mathbf{x},\Phi}}(\mathsf{I} - (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\mathsf{T}_{\mathbf{x},\Phi}) h_{\mathcal{H}_\mathsf{K}}.\end{aligned} \tag{36}$$

Since for all $a \geq 0$, $\sqrt{a}\left(1 - \frac{a}{a+\lambda}\right) = \frac{\sqrt{a}\lambda}{a+\lambda} \leq \frac{1}{2}\sqrt{\lambda}$, using the same arguments as in Equation (34) yields

$$\|\sqrt{\mathsf{T}_{\mathbf{x},\Phi}}(\mathsf{I} - (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\mathsf{T}_{\mathbf{x},\Phi})\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})} \leq \frac{1}{2}\sqrt{\lambda}.$$

Moreover, since for all $a \geq 0$, $1 - \frac{a}{a+\lambda} = \frac{\lambda}{a+\lambda} \leq 1$, similarly we have that

$$\|\mathsf{I} - (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1}\mathsf{T}_{\mathbf{x},\Phi}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})} \leq 1.$$

Thus, taking the norm in Equation (36), using Minkowski's inequality, Lemma D.5 and Equation (25) yields

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} \leq R\sqrt{\|\mathsf{T}_\Phi - \mathsf{T}_{\mathbf{x},\Phi}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})}} + \frac{R}{2}\sqrt{\lambda}. \tag{37}$$

Combining Equations (35) and (37) with Lemma E.3, for $0 < \eta < 1$, we have with probability at least $1 - \eta$

$$\begin{aligned}\|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} &\leq \delta_1(n, \eta/2)\left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\delta_2(n,d,\eta/2)}}{\lambda} \right) \\ \|\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} &\leq R\sqrt{\delta_2(n,d,\eta/2)} + \frac{R}{2}\sqrt{\lambda}.\end{aligned}$$

Using the condition on λ given by Equation (29), still with probability at least $1-\eta$, we have

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} \leq \frac{3}{2\sqrt{\lambda}}\delta_1(n, \eta/2), \tag{38}$$

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^\lambda)\|_{\mathcal{H}_\mathsf{K}} \leq \frac{3R}{2}\sqrt{\lambda}. \tag{39}$$

Combining Equations (38) and (39) into Equation (32) yields that with probability at least $1-\eta$,

$$\mathcal{R}(\Phi \circ h_{\mathbf{z}}^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq \frac{9}{2}\left(\frac{\delta_1(n, \eta/2)^2}{\lambda} + R^2\lambda\right).$$

□

In Proposition E.2, we have a compromise in λ in the two terms. Taking $\lambda = \mathcal{O}(\sqrt{n})$ yields the best compromise. So as to satisfy the condition from Equation (29), we take $\lambda = 6\kappa C_\phi^2 \frac{\log(4/\eta)\sqrt{d}}{\sqrt{n}}$, which after simplifications in the constants yields Proposition E.1.

E.2 Proof of Proposition 5.2 from the main paper

We recall the additional assumption made on the dictionary, as well as the proposition itself which corresponds to Proposition 5.2 from the main paper.

Assumption E.5. *There exists $M(d) \geq 0$ such that for all $\theta \in \Theta$ and for all $l \in [d]$, $|\phi_l(\theta)| \leq M(d)$.*

Remark. The dependence in d is specific to the family to which ϕ belongs. For instance for wavelets, we have $M(d) = 2^{r(\Theta,d)/2}\max_{\theta\in\Theta}|\psi(\theta)|$ with ψ the mother wavelet and $r(\Theta, d) \in \mathbb{N}$ the number of dilatations that are included in ϕ, whereas for a Fourier dictionary we have $M(d) = 1$.

Proposition E.3. *Let $0 < \eta < 1$, taking*

$$\lambda = \lambda_n^*(\eta/3) := 6\kappa C_\phi^2 \frac{\log(6/\eta)\sqrt{d}}{\sqrt{n}},$$

with probability at least $1-\eta$,

$$\mathcal{R}(\Phi \circ \widetilde{h}_{\mathbf{z}}^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq \left(\frac{B_2(d)\sqrt{n}}{m^2} + \frac{B_3(d)}{m^{3/2}} + \frac{9C(d)^2}{2\sqrt{n}m} + \frac{B_4(d)}{\sqrt{n}}\right)\log(6/\eta),$$

with $C(d) := \frac{LM(d)}{C_\phi}$, $B_2(d) := 18\sqrt{d}\left(C(d) + \frac{R}{\sqrt{d}}\right)^2$, $B_3(d) := B_2(d) - 18\frac{R^2}{\sqrt{d}}$, $B_4(d) := \frac{81}{2}\left(\frac{B_0}{\sqrt{d}} + B_1\sqrt{d}\right)$ and B_0 and B_1 are defined as in Proposition E.1.

E.2.1 Approximated solution for partially observed functions

We recall the notion of partially observed functional output sample:

$$\widetilde{\mathbf{z}} := (x_i, (\theta_i, \widetilde{y}_i))_{i=1}^n,$$

where for all $i \in [n]$, $\theta_i \in \Theta^{m_i}$, $\widetilde{y}_i \in \mathbb{R}^{m_i}$ with $m_i \in \mathbb{N}^*$ the number of observations available for the i-th function, and for all $p \in [m_i]$, $\theta_{ip} \in \Theta$ and $\widetilde{y}_{ip} \in \mathbb{R}$. We remind the reader as well that to simplify, we have supposed in Section 5 from the main paper that for all $i \in [n]$, $m_i = m$.

We introduce the notation $\widetilde{\mathbf{y}} := (\widetilde{y}_i)_{i=1}^n$ and highlight that since there is no added noise, we have for all $i \in [n]$

$$\widetilde{y}_i = (y_i(\theta_{ip}))_{p=1}^m.$$

We recall that μ is the uniform probability measure over Θ which governs the draws of the locations of sampling.

For $i \in [n]$, we define $\widetilde{\Phi}_i \in \mathbb{R}^{m \times d}$ the approximation of Φ using the locations θ_i as

$$\widetilde{\Phi}_i := (\phi_1(\theta_i), .., \phi_d(\theta_i)),$$

where for $i \in [n]$ and for $l \in [d]$, $\phi_l(\theta_i) = (\phi_l(\theta_{ip}))_{p=1}^m \in \mathbb{R}^m$.

Let us recall that the solution when the output functions are fully observed (Equation (19)) reads:

$$h_{\mathbf{z}}^{\lambda} = (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1} \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y},$$

with

$$\mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{w} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{x_i} \Phi^{\#} w_i \quad \text{for} \quad \mathbf{w} \in \mathsf{L}^2(\Theta)^n.$$

We now consider of partially observed output functions with observed locations $(\theta_i)_{i=1}^n$ and define an estimator in this setting. We first define

$$\mathsf{A}^{\#}_{\mathbf{x},\widetilde{\Phi}} \widetilde{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{x_i} \frac{\widetilde{\Phi}_i^{\#}}{m} \widetilde{w}_i \quad \text{with} \quad \widetilde{\mathbf{w}} \in \mathbb{R}^{n \times m},$$

The solution we consider when dealing with partially observed functions is then the following

$$\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} := (\mathsf{T}_{\mathbf{x},\Phi} + \lambda \mathsf{I})^{-1} \mathsf{A}^{\#}_{\mathbf{x},\widetilde{\Phi}} \widetilde{\mathbf{y}}.$$

It is another equivalent expression for the plug-in ridge estimator from Definition 4.1 from the main paper.

E.2.2 Concentration results

Lemma E.4. *Let $0 < \eta < 1$, then with probability at least $1 - \eta$*

$$\|\mathsf{A}^{\#}_{\mathbf{x},\widetilde{\Phi}} \widetilde{\mathbf{y}} - \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y}\|_{\mathcal{H}_{\mathsf{K}}} \leq \delta_3(n, m, d, \eta),$$

with δ_3 defined as

$$\delta_3(n, m, d, \eta) := \left(\frac{4(L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_{\phi}R)}{m} + \frac{2L\sqrt{\kappa}\sqrt{d}M(d)}{\sqrt{n}\sqrt{m}} \right) \log\left(2/\eta\right). \tag{40}$$

Proof. Let us define the function $\xi_3 : \mathcal{X} \times \mathsf{L}^2(\Theta) \times \Theta \longrightarrow \mathcal{H}_{\mathsf{K}}$ as $\xi_3 : (x, y, \theta) \longmapsto y(\theta)\mathsf{K}_x\phi(\theta) - \mathsf{K}_x\Phi^{\#}y$

The proof relies on the fact that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{p=1}^{m} \xi_3(x_i, y_i, \theta_{ip}) &= \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{x_i} \frac{\widetilde{\Phi}_i^{\#}}{m} \widetilde{y}_i - \mathsf{K}_{x_i} \Phi^{\#} y_i \\ &= \mathsf{A}^{\#}_{\mathbf{x},\widetilde{\Phi}} \widetilde{\mathbf{y}} - \mathsf{A}^{\#}_{\mathbf{x},\Phi} \mathbf{y}. \end{aligned}$$

Let $(\mathsf{X}_i, \mathsf{Y}_i)_{i=1}^n$ be n i.i.d. RVs distributed according to the distribution ρ. Let $(\vartheta_{ip})_{i=1,p=1}^{n,m}$ be nm i.i.d. RVs distributed according to the distribution μ. For all $i \in [n]$ and for all $p \in [m]$ we then define the RVs W_{ip} as

$$\begin{aligned} \mathsf{W}_{ip} :&= \xi_3(\mathsf{X}_i, \mathsf{Y}_i, \vartheta_{ip}) \\ &= \mathsf{Y}_i(\vartheta_{ip})\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta_{ip}) - \mathsf{K}_{\mathsf{X}_i}\Phi^{\#}\mathsf{Y}_i \\ &= \mathsf{Y}_i(\vartheta_{ip})\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta_{ip}) - \mathbb{E}[\mathsf{Y}_i(\vartheta)\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta)|\mathsf{X}_i, \mathsf{Y}_i], \end{aligned} \tag{41}$$

where the last line holds because μ is the uniform distribution and because we have assumed that $|\Theta| = \int_\Theta 1 \mathrm{d}\theta = 1$ (see the notation and context paragraph at the end of Section 1 from the main paper).

We denote by $\mathbb{P}[.|\mathbf{z}]$ the probability conditional on the realization of the sample $\mathbf{z}$, thus

$$\mathbb{P}[.|\mathbf{z}] = \mathbb{P}[.|\mathsf{X}_i = x_i, \mathsf{Y}_i = y_i, \quad i \in [n]]$$

Then, Equation (41) implies that $\mathbb{E}[\mathsf{W}_{ip}|\mathbf{z}] = 0$.

We define as well for all $p \in [m]$, $\overline{\mathsf{W}}_p := \frac{1}{n}\sum_{i=1}^n \mathsf{W}_{ip}$.

We have almost surely that

$$\begin{aligned}\|\overline{\mathsf{W}}_p\|_{\mathcal{H}_\mathsf{K}} \leq \frac{1}{n}\sum_{i=1}^n \|\mathsf{W}_{ip}\|_{\mathcal{H}_\mathsf{K}} &\leq \frac{1}{n}\sum_{i=1}^n \left(|\mathsf{Y}_i(\vartheta_{ip})|\|\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta_{ip})\|_{\mathcal{H}_\mathsf{K}} + \|\mathsf{K}_{\mathsf{X}_i}\Phi^\#\mathsf{Y}_i\|_{\mathcal{H}_\mathsf{K}}\right) \\ &\leq L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_\phi R.\end{aligned}$$

We have used Assumptions E.4 and E.5 as well as Equation (21).

Since for all $p \in [m]$, the RVs $(\mathsf{W}_{ip})_{i=1}^n$ are independent conditionally on $\mathbf{z}$, we have that

$$\mathbb{E}[\|\overline{\mathsf{W}}_p\|^2_{\mathcal{H}_\mathsf{K}}|\mathbf{z}] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|\mathsf{W}_{ip}\|^2_{\mathcal{H}_\mathsf{K}}|\mathbf{z}]. \tag{42}$$

Using the fact that $\mathbb{E}[\mathsf{Y}_i(\vartheta_{ip})\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta_{ip})|\mathbf{z}] = \mathsf{K}_{x_i}\Phi^\# y_i$, the identity $\mathbb{E}[\|\mathsf{U} - \mathbb{E}[\mathsf{U}]\|^2_{\mathcal{H}_\mathsf{K}}] = \mathbb{E}[\|\mathsf{U}\|^2_{\mathcal{H}_\mathsf{K}}]$ gives us

$$\mathbb{E}[\|\mathsf{W}_{ip}\|^2_{\mathcal{H}_\mathsf{K}}|\mathbf{z}] = \mathbb{E}[\|\mathsf{Y}_i(\vartheta_{ip})\mathsf{K}_{\mathsf{X}_i}\phi(\vartheta_{ip})\|^2_{\mathcal{H}_\mathsf{K}}|\mathbf{z}]. \tag{43}$$

Then using Equation (43) into Equation (42) along with Assumptions E.4 and E.5 yields

$$\mathbb{E}[\|\overline{\mathsf{W}}_p\|^2_{\mathcal{H}_\mathsf{K}}|\mathbf{z}] \leq \frac{1}{n}L^2\kappa d M(d)^2.$$

We can then apply Lemma D.4 to obtain that

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{p=1}^m \overline{\mathsf{W}}_p\right\|_{\mathcal{H}_\mathsf{K}} \leq \left(\frac{4(L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_\phi R)}{m} + \frac{2L\sqrt{\kappa}\sqrt{d}M(d)}{\sqrt{n}\sqrt{m}}\right)\log(2/\eta)\middle|\mathbf{z}\right] \geq 1-\eta.$$

Multiplying the above inequality by $\mathbb{P}[\mathbf{z}]$ and integrating over $\mathbf{z} \in \mathcal{Z}^n$, yields that

$$\mathbb{P}\left[\left\|\mathsf{A}^\#_{\mathbf{x},\widetilde{\Phi}}\widetilde{\mathbf{y}} - \mathsf{A}^\#_{\mathbf{x},\Phi}\mathbf{y}\right\|_{\mathcal{H}_\mathsf{K}} \leq \left(\frac{4(L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_\phi R)}{m} + \frac{2L\sqrt{\kappa}\sqrt{d}M(d)}{\sqrt{n}\sqrt{m}}\right)\log(2/\eta)\right] \geq 1-\eta.$$

□

Lemma E.5. *Let $0 < \eta < 1$, then with probability at least $1-\eta$ the three following inequalities hold:*

$$\|\mathsf{A}^\#_{\mathbf{x},\Phi}\mathbf{y} - \mathsf{T}_{\mathbf{x},\Phi}h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} \leq \delta_1(n, \eta/3) \tag{44}$$

$$\|\mathsf{T}_{\mathbf{x},\Phi} - \mathsf{T}_\Phi\|_{\mathcal{L}_2(\mathcal{H}_\mathsf{K})} \leq \delta_2(n, d, \eta/3) \tag{45}$$

$$\|\mathsf{A}^\#_{\mathbf{x},\widetilde{\Phi}}\widetilde{\mathbf{y}} - \mathsf{A}^\#_{\mathbf{x},\Phi}\mathbf{y}\|_{\mathcal{H}_\mathsf{K}} \leq \delta_3(n, m, d, \eta/3), \tag{46}$$

with δ_1, δ_2 and δ_3 respectively defined as in Equations (26), (28) *and* (40).

Proof. This Lemma is an union bound using Lemma E.1, Lemma E.2 and Lemma E.4. □

E.2.3 Proof

We are now ready to prove Proposition E.3. To do so we prove the following intermediate result of which Proposition E.3 is a direct consequence.

Proposition E.4. *Let $0 < \eta < 1$, provided λ is taken such that*

$$\lambda \geq 6\kappa C_\phi^2 \frac{\log(6/\eta)\sqrt{d}}{\sqrt{n}} = \delta_2(n, d, \eta/3), \tag{47}$$

we have with probability at least $1-\eta$ that

$$\mathcal{R}(\Phi \circ \widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq \frac{27}{4}\left(\left(\frac{A_0(d)^2}{\lambda m^2} + \frac{2A_0(d)A_1(d)}{\lambda\sqrt{n}m^{3/2}} + \frac{A_1(d)^2}{\lambda nm} + \frac{A_2^2}{\lambda n}\right)\log(6/\eta)^2 + \lambda R^2\right), \tag{48}$$

with

$$\begin{aligned} A_0(d) &:= 4(L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_\phi R) \\ A_1(d) &:= 2L\sqrt{\kappa}\sqrt{d}M(d) \\ A_2 &:= 6(\sqrt{\kappa}C_\phi L + \kappa C_\phi^2 R). \end{aligned}$$

Proof. Taking h^λ as in Equation (31), we consider the following decomposition of the risk using Equation (16)

$$\begin{aligned} \mathcal{R}(\Phi \circ \widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) &= \|\sqrt{T_\Phi}(\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} - h_{\mathcal{H}_\mathsf{K}})\|_{\mathcal{H}_\mathsf{K}}^2 \\ &\leq 3\|\sqrt{T_\Phi}(\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} - h_{\mathbf{z}}^{\lambda})\|_{\mathcal{H}_\mathsf{K}}^2 + 3\|\sqrt{T_\Phi}(h_{\mathbf{z}}^{\lambda} - h^{\lambda})\|_{\mathcal{H}_\mathsf{K}}^2 + 3\|\sqrt{T_\Phi}(h^{\lambda} - h_{\mathcal{H}_\mathsf{K}})\|_{\mathcal{H}_\mathsf{K}}^2. \end{aligned} \tag{49}$$

We focus on the term on the left as we have already controlled the two others in the proof of Lemma E.2 . Using the same strategy as for proving Equation (35), we get that

$$\|\sqrt{T_\Phi}(\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} - h_{\mathbf{z}}^{\lambda})\|_{\mathcal{H}_\mathsf{K}} \leq \|\mathsf{A}^{\#}_{\mathbf{x},\widetilde{\Phi}}\widetilde{\mathbf{y}} - \mathsf{A}^{\#}_{\mathbf{x},\Phi}\mathbf{y}\|_{\mathcal{H}_\mathsf{K}}\left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\|\mathsf{T}_\Phi - \mathsf{T}_{\mathbf{x},\Phi}\|_{\mathcal{L}(\mathcal{H}_\mathsf{K})}}}{\lambda}\right). \tag{50}$$

Combining Equations (35) , (37) and (50) with Lemma E.5, for $0 < \eta < 1$, the three following inequalities are verified with probability at least $1-\eta$

$$\begin{aligned} \|\sqrt{T_\Phi}(\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} - h_{\mathbf{z}}^{\lambda})\|_{\mathcal{H}_\mathsf{K}} &\leq \delta_3(n, m, d, \eta/3)\left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\delta_2(n, d, \eta/3)}}{\lambda}\right) \\ \|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^{\lambda} - h^{\lambda})\|_{\mathcal{H}_\mathsf{K}} &\leq \delta_1(n, \eta/3)\left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\delta_2(n, d, \eta/3)}}{\lambda}\right) \\ \|\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^{\lambda})\|_{\mathcal{H}_\mathsf{K}} &\leq R\sqrt{\delta_2(n, d, \eta/3)} + \frac{R}{2}\sqrt{\lambda}. \end{aligned}$$

Using the condition on λ given by Equation (47), still with probability at least $1-\eta$, we have

$$\|\sqrt{T_\Phi}(\widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda} - h_{\mathbf{z}}^{\lambda})\|_{\mathcal{H}_\mathsf{K}} \leq \frac{3}{2\sqrt{\lambda}}\delta_3(n, m, d, \eta/3) \tag{51}$$

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathbf{z}}^{\lambda} - h^{\lambda})\|_{\mathcal{H}_\mathsf{K}} \leq \frac{3}{2\sqrt{\lambda}}\delta_1(n, \eta/3) \tag{52}$$

$$\|\sqrt{\mathsf{T}_\Phi}(h_{\mathcal{H}_\mathsf{K}} - h^{\lambda})\|_{\mathcal{H}_\mathsf{K}} \leq \frac{3R}{2}\sqrt{\lambda}. \tag{53}$$

Combining Equation (51), (52) and (53) into Equation (49) yields that with probability at least $1-\eta$,

$$\mathcal{R}(\Phi \circ \widetilde{h}^{\lambda}_{\widetilde{\mathbf{z}}}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_\mathsf{K}}) \leq \frac{27}{4}\left(\frac{\delta_3(n,m,d,\eta/3)^2}{\lambda} + \frac{\delta_1(n,\eta/3)^2}{\lambda} + R^2\lambda\right).$$

In Proposition E.4, we have a compromise in λ. Taking $\lambda = \mathcal{O}(\sqrt{n})$ yields the best one. So as to satisfy the condition on λ (Equation (47)), we take $\lambda = 6\kappa C_\phi^2 \frac{\log(6/\eta)\sqrt{d}}{\sqrt{n}}$. After simplifications in the constants we get Proposition E.3. □

F ADDITIONAL PL AND KPL RESULTS

F.1 Gradient-based optimization for partially observed functions in the general case

An interesting property of PL (not only when considering vv-RKHSs as hypothesis class as in Section 4 of the main paper) is that the gradient of the data-fitting term can be estimated straightforwardly from partially observed functions. Let us consider the general PL problem (Problem (5) from the main paper):

$$\min_{h\in\mathcal{H}} \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda\Omega_{\mathcal{H}}(h), \tag{54}$$

We recall the definition of a partially observed functional output sample (Equation (3) from the main paper):

$$\widetilde{\mathbf{z}} := (x_i, (\theta_i, \widetilde{y}_i))_{i=1}^n,$$

Let us now compute the gradient for the data-fitting term considering a parametric hypothesis class of the form $\{h_\mathbf{w}, \mathbf{w}\in\mathbb{R}^p\}$; such that for $x\in\mathcal{X}$, $\mathbf{w}\longmapsto h_\mathbf{w}$ is differentiable. The gradient is given by

$$\sum_{i=1}^n (\nabla h_\mathbf{w}(x_i))^{\mathrm{T}} \Phi^{\#} \nabla \ell_{y_i}(\Phi h_\mathbf{w}(x_i)),$$

with $\nabla h_\mathbf{w}(x_i) \in \mathbb{R}^{d\times p}$ the Jacobian of $h_\mathbf{w}(x)$ and $\nabla\ell(y_i, \Phi h_\mathbf{w}(x_i)) \in \mathsf{L}^2(\Theta)$ the gradient of the loss ℓ with respect to its second argument. For integral losses (Equation (1) from the main paper), this gradient is $\nabla\ell(y_i,.) : v \longmapsto (\theta \longmapsto l(y_i(\theta), v(\theta)))$. We can estimate the vectors $\Phi^{\#}\nabla\ell(y_i, \Phi h_\mathbf{w}(x_i))$ from the partially observed functions $((\theta_i, \widetilde{y}_i))_{i=1}^n$:

$$\frac{1}{m_i}\sum_{p=1}^{m_i} l\left(y_i(\theta_{ip}), \phi(\theta_{ip})^{\mathrm{T}} h_\mathbf{w}(x_i)\right)\phi(\theta_{ip}),$$

Then replacing $h_\mathbf{w}$ by the regressor corresponding to the vv-RKHS hypothesis class with separable kernel: $x\longmapsto \mathsf{B}k(x)$, we obtain Equation (12) from the main paper.

Using those estimated gradient is unsurprisingly equivalent to minimizing the problem based on a formulation of an empirical risk using the partially observed functional output sample $\widetilde{\mathbf{z}}$:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{p=1}^{m_i} l\left(y_i(\theta_{ip}), \phi(\theta_{ip})^{\mathrm{T}} h_\mathbf{w}(x_i)\right). \tag{55}$$

F.2 Plug-in ridge estimator and iterative optimization solution for the square loss.

For $i\in[n]$, we recall the definition of $\widetilde{\Phi}_i \in \mathbb{R}^{m_i\times d}$ the discrete approximation of Φ using the locations θ_i:

$$\widetilde{\Phi}_i := (\phi_1(\theta_i), .., \phi_d(\theta_i)),$$

Then in the case of the square loss, Problem (7) from the main paper can be rewritten as

$$\min_{h\in\mathcal{H}_{\mathsf{K}}} \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\widetilde{y}_i}{\sqrt{m_i}} - \frac{\widetilde{\Phi}_i}{\sqrt{m_i}}h(x_i)\right\|^2_{\mathbb{R}^{m_i}} + \lambda\|h\|^2_{\mathcal{H}_{\mathsf{K}}} \tag{56}$$

Let us define $\widetilde{\Phi}\in\mathcal{L}\left(\mathbb{R}^{dn},\mathbb{R}^{\overline{m}}\right)$ as $\widetilde{\Phi}:(u_i)_{i=1}^n \longmapsto \mathsf{vec}\left(\left(\frac{\widetilde{\Phi}_i}{\sqrt{m_i}}u_i\right)_{i=1}^n\right)$ where we have set $\overline{m}:=\sum_{i=1}^n m_i$.

Then using Proposition 4.1 from the main paper, we can rewrite Problem (55) as

$$\min_{\alpha\in\mathbb{R}^{d\times n}} \frac{1}{n}\|\mathsf{vec}(\widetilde{\mathbf{y}}) - \widetilde{\Phi}\mathbf{K}\mathsf{vec}(\alpha)\|^2_{\mathbb{R}^{\overline{m}}} + \lambda\langle\mathsf{vec}(\alpha),\mathbf{K}\mathsf{vec}(\alpha)\rangle_{\mathbb{R}^{dn}}.$$

Carrying the same steps as in the proof of Proposition B.2 yields that α^* is such that

$$\mathsf{vec}(\alpha^*) = ((\widetilde{\Phi}^{\#}\widetilde{\Phi})\mathbf{K} + n\lambda\mathbf{I})^{-1}\widetilde{\Phi}^{\#}\mathsf{vec}(\widetilde{\mathbf{y}}). \tag{57}$$

We remark that $\widetilde{\Phi}^{\#}\mathsf{vec}(\widetilde{\mathbf{y}})\in\mathbb{R}^{dn}$ corresponds to the estimations of the scalar products that we use in the plug-in ridge estimator. Using the same notations as in Definition 4.1 from the main paper, we have $\widetilde{\Phi}^{\#}\mathsf{vec}(\widetilde{\mathbf{y}}) = \mathsf{vec}(\widetilde{\nu})$. Then the only difference with the plug-in ridge estimator is that the matrix $(\Phi^{\#}\Phi)_{(n)}$ is replaced by the matrix $(\widetilde{\Phi}^{\#}\widetilde{\Phi})$ which is block-diagonal with the matrices $\left(\frac{1}{m_i}\widetilde{\Phi}_i^{\#}\widetilde{\Phi}_i\right)_{i=1}^n$ as diagonal blocks. In other words, instead of using the true Gram matrix of the dictionary $\Phi^{\#}\Phi$ for all the observations, we use for the i-th observation an estimated Gram matrix using the locations of observation of the output function y_i.

G RELATED WORKS

We give more details on the methods presented briefly in Section 6.1 from the main paper. Two of them (Reimherr et al., 2018; Oliva et al., 2015) are specific to functional input data. While we propose a straightforward extension of the latter for non-functional input data, such extension is not possible for the former.

G.1 Functional kernel ridge regression (FKRR)

Kadri et al. (2010, 2016) solve a functional KRR problem in the framework of function-valued-RKHSs (fv-RKHSs). To that end, they pose the following empirical risk minimization problem:

$$\min_{f\in\mathcal{H}_{\mathsf{K}^{\mathrm{fun}}}} \frac{1}{n}\sum_{i=1}^{n}\|y_i - f(x_i)\|^2_{\mathcal{Y}} + \lambda\|f\|^2_{\mathcal{H}_{\mathsf{K}^{\mathrm{fun}}}},$$

with $\mathcal{H}_{\mathsf{K}^{\mathrm{fun}}}$ the fv-RKHS associated to some OVK $\mathsf{K}^{\mathrm{fun}}:\mathcal{X}\times\mathcal{X}\longrightarrow\mathcal{L}(\mathcal{Y})$, and $\mathcal{Y}$ a Hilbert space.

Through a representer theorem, the problem can be reformulated using n variables in $\mathcal{Y}$. The optimal representer coefficients can be found by solving the infinite dimensional system:

$$(\mathbf{K}^{\mathrm{fun}} + \lambda\mathbf{I})\alpha^{\mathsf{fun}} = \mathbf{y},$$

with $\alpha^{\mathsf{fun}}\in\mathcal{Y}^n$, $(\mathbf{K}^{\mathrm{fun}}+\lambda\mathbf{I})^{-1}\in\mathcal{L}(\mathcal{Y})^{n\times n}$ and $\mathbf{y}\in\mathcal{Y}^n$.

We now focus on the case of the separable kernel $\mathsf{K}^{\mathrm{fun}}(x,x') = k^{\mathrm{in}}(x,x')\mathsf{L}$. k^{in} is a scalar-valued kernel and $\mathsf{L}\in\mathcal{L}(\mathcal{Y})$ is an integral operator characterized by a scalar-valued kernel k^{out} on Θ^2 and a measure on Θ.

As an example of such kernel, in the experiments we take k^{in} a scalar Gaussian kernel, k^{out} a Laplace kernel and use the Lebesgue measure on $\Theta=[0,1]$ to define the operator L:

$$\mathsf{L}y:\theta'\longmapsto\int_{\theta\in\Theta}\exp\left(-\frac{|\theta'-\theta|}{\sigma_{k^{\mathrm{out}}}}\right)\mathrm{d}\theta. \tag{58}$$

For such separable kernel, the Kronecker product structure $(\mathsf{K}^{\text{fun}} + \lambda \mathsf{I}) = (\mathsf{K}^{\text{fun}}_{\mathcal{X}} \otimes \mathsf{L} + \lambda \mathsf{I})$ can greatly improve the computational complexity; two approaches are possible.

1. An eigendecomposition can be performed. If such decomposition of L is known in closed-form, the Kronecker product can be exploited to solve the system in $\mathcal{O}(n^3 + n^2 J m)$ time, with J the number of eigenfunctions considered and m the size of the discrete grid used to approximate functions in $\mathcal{Y}$. Unfortunately, such closed-forms are rarely known (Rasmussen and Williams, 2006, Section 4.3). We know that one exists if $k^{\text{out}}(\theta_0, \theta_1) = \exp(-|\theta_0 - \theta_1|)$, $\Theta = [0, 1]$ and μ is the Lebesgue measure (Hawkins, 1989), or if k^{out} is a Gaussian kernel, $\Theta = \mathbb{R}^q$ and μ is a Gaussian measure (Zhu et al., 1997). Otherwise, an approximate eigendecomposition can be performed which adds a $\mathcal{O}(m^3)$ term to the above time complexity.
2. The problem can be discretized on a regular grid (Kadri et al., 2010) and solved in $\mathcal{O}(n^3 + m^3 + n^2 m + n m^2)$ time using a Sylvester solver or in $\mathcal{O}(n^3 + t^3)$ time using an eigen decomposition (with higher constants). To compare the above time complexities to that of KPL, we highlight that typically $m \gg d$ and t is at least of the same order as n.

We compare both approaches numerically in Section H.4.4.

G.2 Triple basis estimator (3BE)

Oliva et al. (2015) firstly represent separately the input and output functions on truncated orthonormal bases obtaining a set of input and output decomposition coefficients: $(\beta^{\mathsf{in}}, \beta^{\mathsf{out}})$ with $\beta^{\mathsf{in}} \in \mathbb{R}^{n \times c}$ and $\beta^{\mathsf{out}} \in \mathbb{R}^{n \times d}$; $c \in \mathbb{N}^*$ being the cardinality of the input basis and $d \in \mathbb{N}^*$ that of the output basis. Then, each set of output coefficient (β^{out}_l for $l \in [d]$) is regressed on the input coefficients β^{in} using KRRs approximated with RFFs (Rahimi and Recht, 2008). Denoting by $\mathsf{R}(\beta^{\mathsf{in}}) \in \mathbb{R}^{n \times J}$ the matrix of RFFs evaluated on the input coefficients β^{in}, for all $l \in [d]$, the following (scalar-valued) sub-problem is solved:

$$\min_{c_l \in \mathbb{R}^J} \|\beta^{\mathsf{out}}_l - \mathsf{R}(\beta^{\mathsf{in}}) c\|^2_{\mathbb{R}^n} + \lambda \|c_l\|^2_{\mathbb{R}^J}.$$

All those sub-problems require the inversion of the same matrix $(\mathsf{R}(\beta^{\mathsf{in}})^{\mathrm{T}} \mathsf{R}(\beta^{\mathsf{in}}) + \lambda \mathsf{I})$, which can thus be carried out only once. Putting aside the computations of the decomposition coefficients, solving 3BE then has time complexity $\mathcal{O}(J^3 + J^2 d)$.

Nevertheless, 3BE as proposed in (Oliva et al., 2015) is specific to function-to-function regression. As a consequence, when the input data are not functional (as in Section 6.5 from the main paper), we propose to directly deal with them through a kernel; we call this extension **one basis estimator (1BE)**. We highlight that 1BE is in fact a particular case of the KPL plug-in ridge estimator with ϕ orthonormal and $\mathsf{K} = k\mathsf{I}$. In that case, the time complexity is $\mathcal{O}(n^3 + n^2 d)$ (we solve solve d scalar-valued KRRs problems sharing the same kernel matrix and the same regularization parameter).

G.3 Kernel additive model (KAM)

In this section only, we consider that the input data consist of functions and that $[0, 1]$ is the domain of both input and output functions. In the function-to-function additive linear model (Ramsay and Silverman, 2005), the following empirical risk is minimized:

$$\sum_{i=1}^{n} \int_0^1 \left(y_i(\theta) - a(\theta) - \int_0^1 b(\zeta, \theta) x_i(\zeta) \, \mathrm{d}\zeta \right)^2 \mathrm{d}\theta. \tag{59}$$

The functions $a : [0, 1] \longrightarrow \mathbb{R}$ and $b : [0, 1] \times [0, 1] \longrightarrow \mathbb{R}$ are the functions we want to learn. To define an hypothesis class for them, two truncated bases of $\mathsf{L}^2([0, 1])$ are chosen, one for the input space $(e^{\mathsf{in}}_l)^c_{l=1}$ and one for the output space $(e^{\mathsf{out}}_l)^d_{l=1}$. With the convention that for $\zeta \in [0, 1]$ and $\theta \in [0, 1]$, $e^{\mathsf{in}}(\zeta) = (e^{\mathsf{in}}_l(\zeta))^c_{l=1}$ and $e^{\mathsf{out}}(\theta) = (e^{\mathsf{out}}_l(\theta))^d_{l=1}$, the functions a and b are specified as

$$a(\theta) = \mathsf{A} e^{\mathsf{out}}(\theta)$$
$$b(\zeta, \theta) = (e^{\mathsf{in}}(\zeta))^{\mathrm{T}} \mathsf{B} e^{\mathsf{out}}(\theta).$$

Then, we use those expressions for a and b and minimize the objective from Equation (59) in the variables $\mathsf{A} \in \mathbb{R}^{1\times d}$ and $\mathsf{B} \in \mathbb{R}^{c\times d}$. Importantly, there is not explicit regularization penalty in the problem, however some regularization is achieved implicitly through the choice of the size of the bases c and d.

Reimherr et al. (2018) build on this model using RKHSs. The following empirical risk minimization problem is considered

$$\min_{h\in\mathcal{H}_{k^{\text{add}}}} \sum_{i=1}^{n} \int_0^1 \left(y_i(\theta) - \int_0^1 h(\zeta,\theta,x_i(\zeta))\mathrm{d}\zeta \right)^2 \mathrm{d}\theta + \lambda \|h\|^2_{\mathcal{H}_{k^{\text{add}}}},$$

where $\mathcal{H}_{k^{\text{add}}}$ is the RKHS of a scalar-valued kernel $k^{\text{add}} : ([0,1]\times[0,1]\times\mathbb{R})^2 \longrightarrow \mathbb{R}$ and $\lambda > 0$. A representer theorem leads to a closed-from solution. To alleviate the computations, a truncated basis of $J < n$ of empirical functional principal components of $(y_i)_{i=1}^n$ is used. A matrix of size $nJ \times nJ$ must then be inverted yielding a time complexity of $\mathcal{O}(n^3J^3)$. However, if k^{add} is chosen as a product of three kernels, the separability property can be exploited to solve the problem in $\mathcal{O}(n^3 + J^3 + n^2J + nJ^2)$ time using a Sylvester Solver. Note that this possibility to exploit the Kronecker structure of the matrix A—page 6 of (Reimherr et al., 2018)—is not highlighted nor exploited by the authors. However the main bottleneck of the method is the computation of this matrix A in itself; even when exploiting the product of kernels, $n^2 + J^2$ double integrals must be computed yielding a time complexity of $\mathcal{O}(n^2t^2 + J^2m^2)$ with t the size of the input discretization grid and m that of the output one. Even for medium n, t and m this becomes a challenge, especially as this matrix must be computed many times so as to tune the multiple kernel parameters.

As an example of a product of kernels used for KAM, in the experiments on the toy dataset and on the DTI dataset, we use a product of three Gaussian kernels:

$$k^{\text{add}} : ((\zeta,\theta,s),(\zeta',\theta',s)) \longmapsto \exp\left(\frac{-(\zeta-\zeta')^2}{\sigma_1^2}\right)\exp\left(\frac{-(\theta-\theta')^2}{\sigma_2^2}\right)\exp\left(\frac{-(s-s')^2}{\sigma_3^2}\right). \tag{60}$$

Reimherr et al. (2018) present the model for one functional covariate. However, it is straightforward to extend it to the case where there are several ones. Equivalently, consider the input functions are vector-valued with values in $\mathbb{R}^o$. Then we can consider a kernel defined on the adapted domain $k^{\text{add}} : ([0,1]\times[0,1]\times\mathbb{R}^o)^2 \longrightarrow \mathbb{R}$ and no further adaptations are required.

G.4 Kernel Estimator (KE)

Finally, the functional Nadaraya-Watson kernel estimator has been studied in Ferraty et al. (2011) in the general setting of Banach spaces. Considering a kernel function $K : \mathbb{R} \longmapsto \mathbb{R}$ combined with a given semi-metric S on $\mathcal{X}$, for all $x \in \mathcal{X}$, they use the following estimator:

$$\frac{\sum_{i=1}^n K \circ S(x,x_i) y_i}{\sum_{i=1}^n K \circ S(x,x_i)}.$$

This method is very fast as fitting it boils down to memorizing the training data, however it can lack precision.

H EXPERIMENTAL DETAILS AND SUPPLEMENTS

In this Section we give more insights into the numerical experiments. We introduce a toy function-to-function data to test several robustness properties of our method while two real worlds datasets have been gathered from different publications about functional regression. This collection of dataset could be used in the future for benchmarking.

To avoid mentioning it repeatedly, we highlight that when performing cross-validation, we use 5 folds in all the experiments; and when several values are given for a same parameters, all configurations generated by combining the described parameters/dictionaries are included in the cross-validation.

H.1 Parametrized logcosh loss

We consider the following logcosh loss in 1d:

$$a \in \mathbb{R} \longmapsto \frac{1}{\gamma}\log(\cosh(\gamma a)).$$

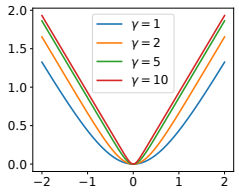

Figure 1: Logcosh loss on $\mathbb{R}$.

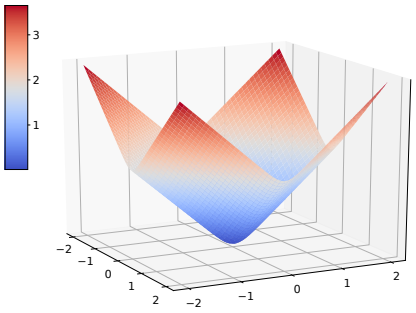

Figure 2: Logcosh loss on $\mathbb{R}^2$ ($\gamma = 5$).

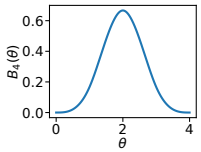

Figure 3: Cubic B-spline.

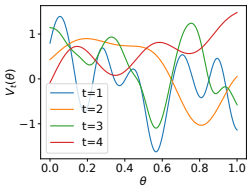

Figure 4: GP draws.

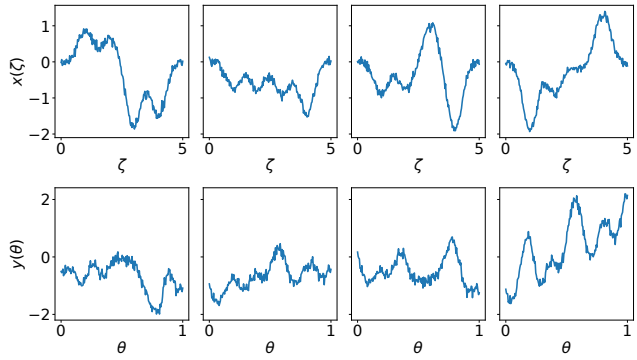

Figure 5: Examples of generated toy data.

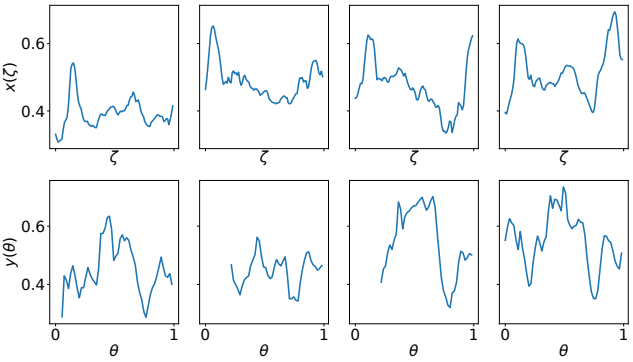

Figure 6: Examples from the DTI dataset.

It corresponds to the loss $l^{(\gamma)}_{\text{lch}}$ defined in Section 6.2 from the main paper. We illustrate the effect of the parameter γ in Figure 1.

As we cannot plot the integral version of this loss, we consider the loss defined on $\mathbb{R}^2$ as follows:

$$(a_0, a_1) \longmapsto \frac{1}{\gamma} \left(\log(\cosh(\gamma a_0)) + \log(\cosh(\gamma a_1)) \right).$$

We plot this loss for $\gamma = 5$ in Figure 2.

H.2 Toy dataset

H.2.1 Generating process

We consider a functional toy dataset. To generate it, we draw $r \in \mathbb{N}$ indepent zero mean Gaussian processes (GP) with Gaussian covariance functions. More precisly, for $t \in [r]$ the Gaussian process V_t has covariance $(\theta_1, \theta_2) \longmapsto \exp\left(-\frac{(\theta_2 - \theta_1)^2}{b_t^2}\right)$. We then keep those Gaussian processes fixed. In practice in those experiments we take $r = 4$ and $b_1 = 0.1$, $b_2 = 0.25$, $b_3 = 0.1$ and $b_4 = 0.25$. An example of a draw of such GPs is displayed in Figure 4. To generate an input/output pair, we draw r coefficients $a \in \mathbb{R}^r$ i.i.d according to a uniform distribution $\mathcal{U}([-1, 1])$ Let B_4 denote the cardinal cubic spline (de Boor, 2001); it is symmetric around $\zeta = 2$ and of width 4 (see Figure 3). Let then $\bar{B}_4 : \zeta \longmapsto B_4(4\zeta + 2)$ (a centered version of B_4 rescaled to have width 1). We consider the input function $x(\zeta) := \sum_{t=1}^{r} a_t \bar{B}_4(\zeta - t)$ with $\zeta \in [0, 5]$. To it we associate the output function $y(\theta) = \sum_{t=1}^{r} a_t V_t(\theta)$ with $\theta \in [0, 1]$. In practice, we observe x and y on regular grids of size 200. For the experiments with missing data, we remove sampling points from those grids. Finally we add Gaussian noise on the input observations with standard deviation $\sigma_x = 0.07$ in all experiments. Examples of data generated that way with a Gaussian noise with standard deviation $\sigma_y = 0.1$ added on the output observations are shown in Figure 5.

H.2.2 Experimental details

We compute the means over 10 runs with different train/test split for all experiments. For all the methods, λ is taken in a geometric grid of size 20 ranging from 10^{-9} to 10^{-4}. Moreover, we consider the following specific parameters.

- **KPL**. We take a truncated Fourier dictionary including 15 frequencies and use the separable kernel $\mathsf{K}(x, x') := k(x, x')\mathsf{I}$ with k a scalar-valued Gaussian kernel with standard deviation $\sigma_k = 20$ and $\mathsf{I} \in \mathbb{R}^{d\times d}$ the identity matrix. When using the logcosh loss, the parameter γ is set to $\gamma = 25$ for the in two experiments related to outliers (so as to approach the absolute loss) and to $\gamma = 10$ for the two other experiments.
- **3BE**. We use k a Gaussian kernel with standard deviation $\sigma_k = 3$. We use truncated Fourier bases as dictionaries, we include 10 and 15 frequencies respectively for the input dictionary and the output one.
- **KAM**. We use the kernel defined in Equation (60) taking $\sigma_1 = 0.2$, $\sigma_2 = 0.1$ and $\sigma_3 = 2.5$ and use $J = 20$ functional principal components.
- **FKRR**. We take a Gaussian kernel as input kernel with standard deviation parameter set as $\sigma_{k^{\text{in}}} = 20$. We use the output kernel defined in Equation (58) setting its parameter to $\sigma_{k^{\text{out}}} = 0.5$.

H.3 DTI dataset

H.3.1 Extensive description of the dataset

The diffusion tensor imaging (DTI) dataset [1] consists of 382 Fractional anisotropy (FA) profiles inferred from DTI scans along two tracts—corpus callosum (CCA) and right corticospinal (RCS). The scans were performed on 142 subjects; 100 multiple sclerosis (MS) patients and 42 healthy controls. MS is an auto-immune disease which causes the immune system to gradually destroy myelin (the substance which isolates and protects the axons of nerve cells), resulting in brain lesions and severe disability. FA profiles are frequently used as an indicator for demyelification which causes a degradation of the diffusivity of the nerve tissues. The latter process is however not well understood and does not occur uniformly in all regions of the brain. We thus propose here to use our method to try to predict FA profiles along the RCS tract from FA profiles along the CCA tract. So as to remain in an i.i.d. framework, we consider only the first scans of MS patients resulting in $n = 100$ pairs of functions. The functions are observed on regular grids of sizes 93 and 54 respectively for the CCA and RCS tracts. However, significant parts of the FA profiles along the RCS tract are missing, we are thus dealing with sparsely sampled functions. Examples of instances from this dataset are shown in Figure 6.

H.3.2 Tuning details for Table 1 of the main paper

The reported means and standard deviations are computer over 20 runs with different train/test split. For all methods (except KE) we center the output functions using the training examples and add back the corresponding mean to the predictions; and we consider values of λ in a geometric grid of size 25 ranging from 10^{-6} to 10^{-2}.

- **KE**. We use a Gaussian kernel with standard deviation in a regular grid ranging from 0.05 to 2 with 200 points.
- **KPL**. For the dictionary, we consider several families of Daubechies wavelets (Daubechies and Heil, 1992) with 2 or 3 vanishing moments and 4 or 5 dilatation levels. We use a separable kernel of the form $\mathsf{K}(x, x') = k(x, x')\mathsf{D}$ with k a Gaussian kernel with fixed standard deviation parameter $\sigma_k = 0.9$. The matrix D is a diagonal matrix of weights decreasing geometrically with the scale of the wavelet at the rate $\frac{1}{b}$ (meaning for instance that at the j-th scale, the corresponding coefficients in the matrix are set to $\frac{1}{b^j}$). b is chosen in a grid ranging from 1 to 2 with granularity 0.1. When using the logcosh loss, we consider values of the parameter γ in $\{0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 10\}$.
- **3BE**. We test the same dictionaries of wavelets as for KPL for both the input and the output functions. We use 200 RFFs for the approximated KRRs; and consider standard deviation for the corresponding approximated Gaussian kernel in the grid $\{7.5, 10, 12.5, 15, 17.5, 20\}$.

[1] This dataset was collected at Johns Hopkins University and the Kennedy-Krieger Institute and is freely available as a part of the *Refund* R package

- **KAM**. We use the product of Gaussian kernels defined in Equation (60) fixing $\sigma_1 = \sigma_2 = \sigma_3 = 0.1$. We consider including $J = 20$ and $J = 30$ principal components for the approximation.
- **FKRR**. We take a Gaussian kernel as input kernel with standard deviation parameter set as $\sigma_{k^{\text{in}}} = 0.9$. We use the output kernel defined in Equation (58) choosing its parameter in $\sigma_{k^{\text{out}}} \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4, 5, 7.5, 10\}$.

H.4 Speech dataset

H.4.1 More on the experimental setting

To match words of varying lengths, we extend symmetrically both the input sounds and the VT functions so as to match the longest word. We represent the sounds using 13 mel-frequency cepstral coefficients (MFCC) acquired each 5ms with a window duration of 10ms. We split the data as $n_{\text{train}} = 300$ and $n_{\text{test}} = 113$. Finally, we normalize the domain of the output functions to $[0, 1]$, and normalize as well their range of values to $[-1, 1]$ so that the scores are of the same magnitude for the different vocal tracts.

The input data consist of matrices in $\mathbb{R}^{m\times 13}$ (here the number of discretization points is the same for the input and for the output functions, so we have $t = m$ discretization points for the MFCCs). Those correspond to discrete observations from $\mathbb{R}^{13}$-valued functions. For ridge-DL-KPL, 1BE/ridge-Four-KPL and FKRR, we wish to use the following integral kernel based on a Gaussian kernel:

$$(x_0, x_1) \longmapsto \int_{[0,1]} \exp\left(\frac{-\|x_1(\zeta) - x_0(\zeta)\|_2^2}{\sigma^2}\right) d\zeta.$$

In practice, we approximate it using the discretized datapoints as:

$$(\widetilde{x}_0, \widetilde{x}_1) \longmapsto \frac{1}{m}\sum_{p=1}^{m} \exp\left(\frac{-\|\widetilde{x}_{1p} - \widetilde{x}_{0p}\|_2^2}{\sigma^2}\right). \tag{61}$$

For KAM, we use the kernel defined on $([0, 1] \times [0, 1] \times \mathbb{R}^{13})^2$ by:

$$((\zeta, \theta, w), (\zeta', \theta', w')) \longmapsto \exp\left(\frac{-(|\zeta - \zeta'|}{\sigma_1}\right) \exp\left(\frac{-|\theta - \theta'|}{\sigma_2}\right) \exp\left(\frac{-\|w - w'\|_2^2}{\sigma_3^2}\right). \tag{62}$$

In practice there are magnitude differences between the MFCCs. So as to avoid biasing the norms to be over-representative of the larger ones, before applying the above describe kernels, we standardize the MFCCs using the training data. For the r-th MFCC, we set $\text{avg}^{(r)} := \frac{1}{n_{\text{train}}m}\sum_{i=1}^{n_{\text{train}}}\sum_{p=1}^{m}\widetilde{x}_{ip}^{(r)}$ and $\text{std}^{(r)} := \sqrt{\frac{1}{n_{\text{train}}m-1}\sum_{i=1}^{n_{\text{train}}}\sum_{p=1}^{m}(\widetilde{x}_{ip}^{(r)} - \text{avg}^{(r)})^2}$, and use as input data $\left(\left(\frac{x_i^{(r)}}{\text{std}^{(r)}}\right)_{r=1}^{13}\right)_{i=1}^{n_{\text{train}}}$.

H.4.2 Details for the MSEs part of Figure 2 from the main paper

The reported means and standard deviations are computed over 10 runs with different train/test split. For all methods, we consider values of λ in a geometric grid ranging from 10^{-12} to 10^{-5} of size 30 and try both centering and not centering the output functions. For ridge-DL-KPL, 1BE/ridge-Four-KPL and FKRR, we use the kernel from Equation (61) as input kernel taking $\sigma \in \{3, 4, 5, 7.5, 10\}$.

- **ridge-DL-KPL**. The dictionary ϕ is learnt by solving Problem (6) from the main paper with $\mathcal{C}$ and $\Omega_{\mathbb{R}^d}$ as introduced in Section 3.2 from the main paper. The number of atoms is fixed at 30.
- **1BE/ridge-Four-KPL**. We use a truncated Fourier basis as dictionary with included number of frequencies in the grid $\{20, 30, 40, 50\}$.
- **FKRR**. We use the kernel from Equation (58) as output kernel. We consider the following values for its parameter: $\sigma_{k^{\text{out}}} \in \{0.005, 0.01, 0.05, 0.1, 0.125, 0.15\}$.

- **KAM**. We use the kernel defined above in Equation (62) for which we consider the following parameters values $\sigma_1 \in \{0.01, 0.05, 0.1, 0.5\}$, $\sigma_2 \in \{0.0005, 0.001, 0.005, 0.01\}$ and $\sigma_3 \in \{0.05, 0.1, 0.5, 1, 5\}$. We consider also $J \in \{30, 40, 50\}$ functional PCAs.

H.4.3 Details for the fitting times part of Figure 2 from the main paper

Infrastructure and measurements details. So as to get better control over execution, we perform those experiments on a laptop rather than on the computing cluster used for the other experiments. This laptop is equipped with a 8th Generation Intel Core i7-8665U processor and 16 Gb of RAM. In Python, using the *multiprocessing* package, we execute the tasks in parallel, each on exactly one core of the CPU. We measure the corresponding CPU time using the *process_ time()* function from the *time* package.

Parameters. Computation times necessarily depend on the choice of parameters. This dependence can be explicit for parameters determining the complexity of the problems (for instance the size of a dictionary or the size of an approximation grid). For such parameters, we use fixed values for each method which correspond either to the fixed values used or to those elected by cross-validation in the MSEs experiments; we detail those values below. Other parameters can influence the computational times through the conditioning of the problem. To account for this, we consider several values which we give below as well. The means and standard deviations of the obtained fitting times are reported in the right panel of Figure 2 from the main paper.

The computation times are averaged over 10 runs of the experiments with different shuffling of the dataset and over the VTs. For all methods, we consider values of λ in a geometric grid ranging from 10^{-12} to 10^{-5} of size 30 and center the output functions. For ridge-DL-KPL, 1BE/ridge-Four-KPL and FKRR, we use the kernel from Equation (61) as input kernel taking $\sigma = 3$.

- **ridge-DL-KPL**. The dictionary ϕ is learnt by solving Problem (6) from the main paper with $\mathcal{C}$ and $\Omega_{\mathbb{R}^d}$ as introduced in Section 3.2 from the main paper. The number of atoms is fixed at 30.

- **1BE/ridge-Four-KPL**. We use a truncated Fourier basis as dictionary with 50 included frequencies, thus the size of the dictionary is $d = 99$ (cosinuses and sinuses are included plus a constant function).

- **FKRR**. We use the kernel from Equation (58) as output kernel. We consider the following values for its parameter: $\sigma_{k^{\text{out}}} \in \{0.05, 0.1\}$.

- **KAM**. We use the kernel defined above in Equation (62) for which we use the following parameters values: $\sigma_1 = 0.1$, $\sigma_2 = 0.05$ and $\sigma_3 = 1$. We take $J = 40$ functional PCAs.

H.4.4 Comparison of solvers for FKRR

As highlighted in Section G, there are two possible ways of solving FKRR with a separable kernel. We compare the two approaches on the speech dataset in Figure 7. FKRR Eigapprox corresponds to the eigendecomposition solver and FKRR Syl to the Sylvester solver. Let J be the number of eigenfunctions considered for the output operator L. The difference in computational cost is mostly imputable to the need in FKRR Eigapprox to instantiate and compute nJ functions which correspond to Kronecker products between eigenvectors of the kernel matrix and eigenfunctions of the output operator. However, since those vectors, are functions, so as to be manipulated, they need to be discretized. Considering a discretization grid of size m, those vectors are of size $n \times m$ (see Algorithm 1 in Kadri et al. (2016) for more details) which can be heavy (there are nJ of them).

To obtain Figure 7, we consider the following parameters for the two solvers.

- **FKRR Eigapprox.** We use $J = 20$ eigenfunctions to approximate the output operator, a grid of size $t = 300$ to approximate functions. We take the output kernel parameters in $\sigma_{k^{\text{out}}} \in \{0.02, 0.05, 0.1, 0.15\}$ and λ in a geometric grid of size 30 ranging from 10^{-12} to 10^{-5}.

- **FKRR Syl.** The plots correspond to the experiments already performed and described previously.

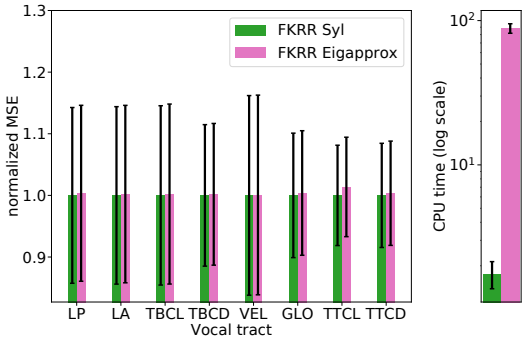

Figure 7: Comparison of two solvers for FKRR on speech dataset

References

L. Baldassarre, L. Rosasco, and A. Barla. Multi-output learning via spectral filtering. *Machine Learning*, 87: 259–301, 2012.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

R. Bhatia. *Matrix analysis*. Springer, 1997.

A. Caponnetto and E. De Vito. Risk bounds for regularized least-squares algorithm with operator-valued kernels. Technical report, MIT, Computer Science and Artificial Intelligence Laboratory, 2005.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, pages 331–368, 2007.

C. Carmeli, E. De Vito, and V. Umanita. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

I. Daubechies and C. Heil. *Ten Lectures on Wavelets*. American Institute of Physics, 1992.

C. de Boor. *A practical guide to Splines - Revised Edition*. Springer, 2001.

F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu. Kernel regression with functional response. *Electron. J. Statist.*, 5: 159–171, 2011.

D. L. Hawkins. Some practical problems in implementing a certain sieve estimator of the gaussian mean function. *Communications in Statistics- Simulationas and Computations*, 18, 1989.

H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional RKHS approach. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 374–380, 2010.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17:1–54, 2016.

Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random Fourier features. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 3905–3914, 2019.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

J. Oliva, W. Neiswanger, B. Poczos, E. Xing, H. Trac, S. Ho, and J. Schneider. Fast function to function regression. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38, pages 717–725, 2015.

I. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability and Its Applications*, 30:143–148, 1986.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 20*, pages 1177–1184. Curran Associates, Inc., 2008.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, 2005.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning, 2006.

M. Reimherr, B. Sriperumbudur, and B. Taoufik. Optimal prediction for additive function on function regression. *Electronic Journal of Statistics*, 12:4571–4601, 2018.

E. Senkene and A. Templeman. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 1973.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, pages 153–172, 2007.

V. Yurinsky. *Sums and Gaussian Vectors*. Springer, 1995.

H. Zhu, C. K. I. Williams, R. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. In *Neural Networks and Machine Learning*, pages 167–184. Springer-Verlag, 1997.