

Nonlinear Functional Output Regression: A Dictionary Approach

Dimitri Bouche¹

¹LTCI, Télécom Paris, Institut Polytechnique de Paris

Marianne Clausel²

François Roueff¹

Florence d'Alché-Buc¹

²Université de Lorraine, CNRS, IECL

Abstract

To address functional-output regression, we introduce projection learning (PL), a novel dictionary-based approach that learns to predict a function that is expanded on a dictionary while minimizing an empirical risk based on a functional loss. PL makes it possible to use non orthogonal dictionaries and can then be combined with dictionary learning; it is thus much more flexible than expansion-based approaches relying on vectorial losses. This general method is instantiated with reproducing kernel Hilbert spaces of vector-valued functions as kernel-based projection learning (KPL). For the functional square loss, two closed-form estimators are proposed, one for fully observed output functions and the other for partially observed ones. Both are backed theoretically by an excess risk analysis. Then, in the more general setting of integral losses based on differentiable ground losses, KPL is implemented using first-order optimization for both fully and partially observed output functions. Eventually, several robustness aspects of the proposed algorithms are highlighted on a toy dataset; and a study on two real datasets shows that they are competitive compared to other nonlinear approaches. Notably, using the square loss and a learnt dictionary, KPL enjoys a particularly attractive trade-off between computational cost and performances.

1 INTRODUCTION

In a large number of fields such as Biomedical Signal Processing, Epidemiology Monitoring, Speech and Acoustics, Climate Science, each data instance consists in a high number of measurements of a common

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

underlying phenomenon. Such high-dimensional data generally enjoys strong smoothness across features. To exploit that structure, it can be interesting to model the underlying functions rather than the vectors of discrete measurements we observe, opening the door to functional data analysis (FDA; Ramsay and Silverman, 2005). In practice, FDA relies on the assumption that the sampling rate of the observations is high enough to consider them as functions. Of special interest is the general problem of functional output regression (FOR) in which the output variable is a function and the input variable can be of any type, including a function.

While functional linear models have received a great deal of attention—see the additive linear model and its variations (Ramsay and Silverman, 2005; Morris, 2015, and references therein)—, nonlinear ones have been less studied. Reimherr et al. (2018) extend the function-to-function additive linear model by considering a tri-variate regression function in a reproducing kernel Hilbert space (RKHS). In non-parametric statistics, Ferraty and Vieu (2006) introduce variations of the Nadaraya-Watson kernel estimator for outputs in a Banach space. Oliva et al. (2015) rather project both input and output functions on orthogonal bases and regress the obtained output coefficients separately on the input ones using approximate kernel ridge regressions (KRR). Finally, extending kernel methods to functional data, Lian (2007) introduces a function-valued KRR. In that context Kadri et al. (2010, 2016) propose a solution based on the approximate inversion of an infinite-dimensional linear operator and studies richer kernels. We give more details on those methods and compare them with our approach in Section 6.1.

In this paper we introduce a novel dictionary-based approach to FOR. We learn to predict a function that is expanded on a dictionary while minimizing an empirical risk based on a functional loss. We call this approach *projection learning* (PL). It can be instantiated with any machine learning algorithm outputting vectors using a wide range of functional losses. PL also makes it possible to use non-orthonormal dictionaries. It represents a crucial advantage as complex functions generally cannot be well represented using

few vectors in conventional bases. They can however be compressed very efficiently using learnt redundant dictionaries (Mallat, 2008). Then, to solve FOR problems with complex output functions, PL combined with dictionary learning (DL) algorithms (Dumitrescu and Irofti, 2018) can be both fast and accurate. In practice functions are not fully observed; discrete observations are rather available. PL can accommodate such realistic case without making any assumptions on the sampling grids, either by learning with an estimated gradient or by plugging in an estimator in a closed-form functional solution.

Then, considering vector-valued RKHSs (vv-RKHS, Micchelli and Pontil, 2005), we introduce *kernel-based projection learning* (KPL). Vv-RKHSs extend the scope of kernel methods to vector-valued functions by means of operator-valued kernels (OVK)—see Section A of the Supplement for an introduction. They constitute a principled way of performing vector-valued nonlinear regression considering any type of input data for which a kernel can be defined (Shawe-Taylor and Cristianini, 2004). Learning typically relies on a representer theorem which remains valid for the KPL problem.

Contributions. We introduce PL, a novel dictionary-based approach to FOR. It can handle non orthonormal dictionaries and can thus be combined with dictionary learning. Then, we focus on KPL, an instantiation based on vv-RKHSs. For the functional square loss, we propose two estimators, one for fully observed output functions and another for partially observed ones. Both are backed with an excess risk bound. For an integral loss based on a differentiable ground loss, we solve KPL using first-order optimization and show that the gradient can easily be estimated from partially observed functions. Eventually, we study different robustness aspects of the proposed algorithms on a toy dataset; and demonstrate on two real datasets that they can be competitive with other nonlinear FOR methods while keeping the computational cost significantly lower.

Notations and context. We assimilate the spaces $(\mathbb{R}^d)^n$ and \mathbb{R}^{dn} . The concatenation of vectors $(u_i)_{i=1}^n \in \mathbb{R}^{dn}$ is denoted $\text{vec}((u_i)_{i=1}^n) \in \mathbb{R}^{dn}$. For $n \in \mathbb{N}^*$, we use the shorthand $[n]$ for the set $\{1, \dots, n\}$. We denote by $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ the space of functions from \mathcal{X} to \mathcal{Y} . For two Hilbert spaces \mathcal{U} and \mathcal{Y} , $\mathcal{L}(\mathcal{U}, \mathcal{Y})$ is the set of bounded linear operators from \mathcal{U} to \mathcal{Y} and $\mathcal{L}(\mathcal{U}) := \mathcal{L}(\mathcal{U}, \mathcal{U})$. The adjoint of a linear operator A is denoted $A^\#$. For $\mathcal{U} = \mathbb{R}^d$, we introduce $A_{(n)} \in \mathcal{L}(\mathbb{R}^{dn}, \mathcal{Y}^n)$ as $A_{(n)} : \text{vec}((u_i)_{i=1}^n) \mapsto (Au_1, \dots, Au_n)$ and $A_{\text{mat},(n)} \in \mathcal{L}(\mathbb{R}^{dn}, \mathcal{Y}^n)$ as $A_{\text{mat},(n)} : (u_i)_{i=1}^n \mapsto (Au_1, \dots, Au_n)$. For $B \in \mathbb{R}^{p \times q}$, $C \in \mathbb{R}^{d \times n}$, $B \otimes C \in \mathbb{R}^{pd \times qn}$ denotes the Kronecker product. Finally $L^2(\Theta)$ stands for the Hilbert space of real-valued square integrable functions on a given compact subset $\Theta \subset \mathbb{R}^q$; without loss of

generality we suppose that $|\Theta| := \int_\Theta 1 d\theta = 1$.

2 PROJECTION LEARNING

2.1 Functional output regression

Let \mathcal{X} be a measurable space and (X, Y) be a couple of random variables on $\mathcal{Z} := \mathcal{X} \times L^2(\Theta)$ with joint probability distribution ρ . To introduce the FOR problem, we define a functional loss ℓ as a real-valued function over $L^2(\Theta) \times L^2(\Theta)$. Examples of functional losses include the functional square loss and more generally, any integral of a ground loss $l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. Particularly, given such ground loss l , for $(y_0, y_1) \in L^2(\Theta) \times L^2(\Theta)$, a functional loss ℓ can be defined as:

$$\ell(y_0, y_1) = \int_\Theta l(y_0(\theta), y_1(\theta)) d\theta. \quad (1)$$

Specifically, taking the square loss as ground loss $l(y_0(\theta), y_1(\theta)) = (y_0(\theta) - y_1(\theta))^2$ we obtain the functional square loss $\ell_2(y_0, y_1) := \|y_0 - y_1\|_{L^2(\Theta)}^2$, widely used in the literature (Kadri et al., 2010).

Given such functional loss ℓ and a hypothesis class $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, L^2(\Theta))$, we now define the FOR problem as

$$\min_{f \in \mathcal{G}} \mathcal{R}(f) := \mathbb{E}_{(X, Y) \sim \rho} [\ell(Y, f(X))]. \quad (2)$$

However, we have access to the joint probability distribution ρ only through an observed sample. The aim is then to approximately solve the above problem using the available data. We study two possible settings.

In the first one, the output functions are *fully observed*. Our sample $\mathbf{z} := (x_i, y_i)_{i=1}^n$ then consists of $n \in \mathbb{N}$ i.i.d. realizations drawn from ρ , this setting coincides with the so-called *dense* one described in FDA (Kokoska and Reimherr, 2017). By contrast, in the *partially observed* setting (also referred to as the *sparse* one, described and studied in Kokoska and Reimherr (2017); Li and Hsing (2010); Cai and Yuan (2011)), the output functions are observed on grids which may be irregular, subject to randomness and potentially different for each function. Even though the former scenario is relatively frequent in theoretical works, the latter can be more realistic.

In the partially observed setting, we suppose that we only observe each y_i on a random sample of locations, $\theta_i := (\theta_{ip})_{p=1}^{m_i} \in \Theta^{m_i}$, drawn from a probability distribution μ . For the sake of simplicity, μ is chosen as the uniform distribution on Θ and the draws of locations are supposed to be independent. The learning problem depicted in Equation (2) has now to be solved using a partially observed functional output sample:

$$\tilde{\mathbf{z}} := (x_i, (\theta_i, \tilde{y}_i))_{i=1}^n, \quad (3)$$

where for all $i \in [n]$, $\theta_i \in \Theta^{m_i}$, $\tilde{y}_i \in \mathbb{R}^{m_i}$ 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 $\tilde{y}_{ip} \in \mathbb{R}$.

In this paper, we propose a novel angle to address the FOR problem using both types of samples.

2.2 Approximated FOR

To tackle Problem (2), we propose to learn to predict expansion coefficients on a dictionary of functions $\phi := (\phi_l)_{l=1}^d \in L^2(\Theta)^d$ with $d \in \mathbb{N}^*$ (considerations on the choice of this dictionary are postponed to Section 3). We then introduce the following linear operator:

Definition 2.1. (Projection operator) For a dictionary ϕ , the associated projection operator Φ is defined by $\Phi : u \in \mathbb{R}^d \mapsto \sum_{l=1}^d u_l \phi_l \in L^2(\Theta)$.

We can give an explicit expression of $\Phi^\#$ as well as a matrix representation of $\Phi^\# \Phi$.

Lemma 2.1. *The adjoint of Φ is given by $\Phi^\# : g \in L^2(\Theta) \mapsto (\langle \phi_l, g \rangle_{L^2(\Theta)})_{l=1}^d \in \mathbb{R}^d$. Thus we have $\Phi^\# \Phi = (\langle \phi_l, \phi_s \rangle_{L^2(\Theta)})_{l,s=1}^d$.*

The core idea of PL is to define a simpler model $f(x) = \Phi h(x)$ in Problem (2), where $h : \mathcal{X} \mapsto \mathbb{R}^d$ is a vector-valued function. This yields the problem

$$\min_{h \in \mathcal{H}} \mathcal{R}(\Phi \circ h), \quad (4)$$

that we can solve using a sample from one or the other of the two observation settings previously defined.

In the fully observed setting, we can minimize over $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ the empirical counterpart of the true risk based on \mathbf{z} , $\widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i))$, with some additional penalty $\Omega_{\mathcal{H}} : \mathcal{H} \rightarrow \mathbb{R}$ to control the model complexity:

$$\min_{h \in \mathcal{H}} \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda \Omega_{\mathcal{H}}(h), \quad (5)$$

with $\lambda > 0$. In other words, we search a solution in the hypothesis space $\{f : x \mapsto \Phi h(x), h \in \mathcal{H}\}$ and solve a function-valued problem at the price of solving a vector-valued one in \mathcal{H} . Even though a vector-valued function is learned, the loss remains a functional one. Moreover, any predictive model devoted to vectorial output regression (e. g. neural networks, random forests, kernel methods etc.) is eligible. We regularize our model through the vector-valued function h .

To tackle the partially observed setting, rather than formulating an empirical counterpart of the true risk based on $\widetilde{\mathbf{z}}$, we exploit specific properties of the learning algorithms proposed in Section 4. Namely in our closed form ridge estimator (Proposition 4.2) or in the gradient (Equation (11)), the output functions only appear

through scalar products with elements of the dictionary. We can then estimate those from $((\theta_i, \tilde{y}_i))_{i=1}^n$ and use a plug-in strategy. Interestingly, computing the gradient for the data attach term in Problem (5) shows that this is a feature of projection learning which is not specific to the vv-RKHS instantiation (see Section F.1 of the Supplement for details).

3 DICTIONARIES

In solving Problem (4) instead of Problem (2), we restrict the predictions of our model to $\text{Span}(\phi)$, the space of linear combinations of functions of ϕ . As a result ϕ must be chosen so that the functions $(y_i)_{i=1}^n$ can be approximated accurately by elements from $\text{Span}(\phi)$. To achieve this, several strategies are possible.

3.1 General dictionaries

Orthonormal and Riesz bases. We can consider families of functions known to provide sharp approximations of functions belonging to $L^2(\Theta)$. Orthogonal bases such as Fourier bases or wavelets bases (DeVore et al., 1992), as well as Riesz bases (see Definition 5.1) such as splines (Oswald, 1990), have proved their efficiency in signal compression. In practice, a choice among those families can be made from observed properties of the output functions or prior information on the generating process. Then within a family, dictionaries with different parameters (number of functions and/or other parameters) can be considered. A cross-validation can be performed to select one.

Families of random functions, such as random Fourier features (RFFs, Rahimi and Recht, 2008a) can enjoy good approximation properties as well. Through the choice of such family, we approximate the output functions in a space that is dense in a RKHS (Rahimi and Recht, 2008b). The link with this RKHS can moreover be made explicit as a family is associated to a given kernel. The kernel can then be chosen by cross-validation and number of functions to include results from a precision/computation time trade-off.

3.2 Dictionary learning

When the output functions are too complex, selecting a dictionary can however be difficult. The choice of a family may not be evident and it may take too many atoms (functions) to reach a satisfying approximation precision. While functional principal component analysis (FPCA; Ramsay and Silverman, 2005) addresses the first issue by ensuring that $\text{Span}(\phi)$ is close to $\text{Span}((y_i)_{i=1}^n)$, it does not address the second one. If the functions at hand are too complex, a very large number of eigenfunctions will be necessary to reach

an acceptable approximation quality. By opposition, dictionary learning (DL) solves both problems; it can generally synthesize faithfully the properties of a complicated set of functions while using very few atoms (Mairal et al., 2009). The DL problem is of the form

$$\min_{\phi \in \mathcal{C}, \beta \in \mathbb{R}^{d \times n}} \frac{1}{n} \sum_{i=1}^n \left(\|y_i - \Phi \beta_i\|_{L^2(\Theta)}^2 + \tau \Omega_{\mathbb{R}^d}(\beta_i) \right), \quad (6)$$

where \mathcal{C} is a set of constraint for the dictionary, $\Omega_{\mathbb{R}^d} : \mathbb{R}^d \rightarrow \mathbb{R}$ is a penalty on the learned representation coefficients and $\tau > 0$ is a trade-off parameter. $\mathcal{C} := \{\phi \in L^2(\Theta)^d, \|\phi_l\|_{L^2(\Theta)}^2 \leq 1, l \in [d]\}$ and $\Omega_{\mathbb{R}^d} := \|\cdot\|_1$ are the most common choices (Lee et al., 2007; Mairal et al., 2009), and most existing algorithms are based on alternating optimization schemes (Dumitrescu and Irofti, 2018, and references therein).

As opposed to other dictionary based methods (Oliva et al., 2015), KPL can handle the resulting non orthonormal dictionary and can thus benefit from the compression power of DL. Then combining the two, we obtain a FOR method that can deal directly with complex functional-output datasets at a low computational cost. Admittedly, solving Problem (6) has a cost, which must however be mitigated. Many efficient algorithms exist (Dumitrescu and Irofti, 2018) and the dictionary moreover needs to be learnt only once (when selecting other parameters through cross-validation, it needs only be learnt once per fold).

4 VV-RKHS INSTANTIATION

We now focus on projection learning using vv-RKHSs.

4.1 Vv-RKHSs and representer theorem

Let $\mathbf{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{L}(\mathbb{R}^d)$ be an OVK and $\mathcal{H}_{\mathbf{K}} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ its associated vv-RKHS. For $x \in \mathcal{X}$, we define $\mathbf{K}_x \in \mathcal{L}(\mathbb{R}^d, \mathcal{H}_{\mathbf{K}})$ as $\mathbf{K}_x : u \mapsto \mathbf{K}_x u$, with $\mathbf{K}_x u : x' \mapsto \mathbf{K}(x', x)u$. We consider Problem (5) taking $\mathcal{H} = \mathcal{H}_{\mathbf{K}}$ as vector-valued hypothesis class. Setting the regularization as $\Omega_{\mathcal{H}_{\mathbf{K}}}(h) := \|h\|_{\mathcal{H}_{\mathbf{K}}}^2$ yields the following instantiation of PL with vv-RKHS:

$$\min_{h \in \mathcal{H}_{\mathbf{K}}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i)) + \lambda \|h\|_{\mathcal{H}_{\mathbf{K}}}^2. \quad (7)$$

To solve Problem (7), we show in Proposition 4.1 that it benefits from a representer theorem, which proof is given in Section B.1 of the Supplement. It can then be restated as a problem with dn variables.

Proposition 4.1. (Representer theorem) For ℓ continuous and convex with respect to its second argument, Problem (7) admits a unique minimizer $h_{\mathbf{z}}^{\lambda}$.

Moreover there exists $\alpha \in \mathbb{R}^{d \times n}$ such that

$$h_{\mathbf{z}}^{\lambda} = \sum_{j=1}^n \mathbf{K}_{x_j} \alpha_j.$$

Choice of kernels. In vv-RKHSs, the choice of the kernel determines the regularization conveyed by the RKHS norm. In practice, the separable kernel is often used: $\mathbf{K} = k\mathbf{B} : (x_0, x_1) \mapsto k(x_0, x_1)\mathbf{B}$ (Alvarez et al., 2012), with k a scalar kernel on \mathcal{X} and $\mathbf{B} \in \mathbb{R}^{d \times d}$ a positive definite symmetric matrix encoding relations between the output variables. In KPL, \mathbf{B} can encode prior information on the dictionary. A diagonal matrix can for instance penalize higher frequencies/scales more. We exploit this with wavelets in the experiments related to biomedical imaging in Section 6.4.

4.2 Ridge solution

In this section, we focus on the functional square loss.

Fully observed setting. By Proposition 4.1, Problem (7) can be rewritten as

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^{d \times n}} \frac{1}{n} \|\mathbf{y} - \Phi_{(n)} \mathbf{K} \text{vec}(\alpha)\|_{L^2(\Theta)^n}^2 \\ + \lambda \langle \text{vec}(\alpha), \mathbf{K} \text{vec}(\alpha) \rangle_{\mathbb{R}^{dn}}, \end{aligned} \quad (8)$$

where $\mathbf{y} := (y_i)_{i=1}^n \in L^2(\Theta)^n$, the kernel matrix is defined block-wise as $\mathbf{K} := [\mathbf{K}(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{dn \times dn}$; and vec and $\Phi_{(n)}$ are introduced in Section 1. We then derive a closed-form for fully observed output functions.

Proposition 4.2. (Ridge solution) The minimum in Problem (8) is achieved by any $\alpha^* \in \mathbb{R}^{d \times n}$ verifying

$$(\mathbf{K}(\Phi^{\#}\Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I}) \text{vec}(\alpha^*) := \mathbf{K}\Phi_{(n)}^{\#} \mathbf{y}. \quad (9)$$

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

$$\text{vec}(\alpha^*) = ((\Phi^{\#}\Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I})^{-1} \Phi_{(n)}^{\#} \mathbf{y}. \quad (10)$$

We define the ridge estimator as $h_{\mathbf{z}}^{\lambda} := \sum_{j=1}^n \mathbf{K}_{x_j} \alpha_j^*$.

The proof is detailed in Section B.2 of the Supplement. $(\Phi^{\#}\Phi)_{(n)}$ is a block diagonal matrix with the Gram matrix $\Phi^{\#}\Phi$ of the dictionary repeated on its diagonal. Then if ϕ is orthonormal, Equation (10) simplifies to $\text{vec}(\alpha^*) = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \Phi_{(n)}^{\#} \mathbf{y}$.

Partially observed setting We can derive a solution for partially observed functions from Proposition 4.2. To that end, we remark that in Equation (10), the output functions only appear through the quantity $(\Phi_{(n)})^{\#} \mathbf{y} = \text{vec}((\Phi^{\#} y_i)_{i=1}^n) \in \mathbb{R}^{dn}$ with for $i \in [n]$, $\Phi^{\#} y_i = (\langle y_i, \phi_l \rangle_{L^2(\Theta)})_{l=1}^d$. As a consequence, we propose to estimate those scalar products from the available observations and then to plug the obtained estimators into Equation (10).

Definition 4.1. (Plug-in ridge estimator.) For all $l \in [d]$ and $i \in [n]$, let $\tilde{\nu}_{il} := \frac{1}{m_i} \sum_{p=1}^{m_i} \tilde{y}_{ip} \phi_l(\theta_{ip})$ be the entries of $\tilde{\nu} \in \mathbb{R}^{d \times n}$. Let $\tilde{\alpha}^* \in \mathbb{R}^{d \times n}$ be such that $\text{vec}(\tilde{\alpha}^*) = ((\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I})^{-1} \text{vec}(\tilde{\nu})$. We then define the plug-in ridge estimator as $\tilde{h}_{\tilde{\mathbf{z}}}^\lambda := \sum_{j=1}^n \mathbf{K}_{x_j} \tilde{\alpha}_j^*$.

We propose the following strategy to compute this estimator for a separable kernel $\mathbf{K} = \mathbf{k}\mathbf{B}$.

Fast algorithm for plug-in ridge estimator. The matrix \mathbf{K} can be rewritten as $\mathbf{K} = \mathbf{K}_\mathcal{X} \otimes \mathbf{B}$ with $\mathbf{K}_\mathcal{X} := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$. Solving the linear system in Equation (10) has time complexity $\mathcal{O}(n^3 d^3)$. However, $(\Phi_{(n)})^\# \Phi_{(n)} = \mathbf{I} \otimes (\Phi^\# \Phi)$, thus $(\Phi_{(n)})^\# \Phi_{(n)} \mathbf{K} = (\mathbf{I} \otimes (\Phi^\# \Phi))(\mathbf{K}_\mathcal{X} \otimes \mathbf{B})$. Using the mixed product property (Horn and Johnson, 1991, Lemma 4.2.10), we must solve $(\mathbf{K}_\mathcal{X} \otimes ((\Phi^\# \Phi) \mathbf{B}) + n\lambda \mathbf{I}) \text{vec}(\alpha) = \text{vec}(\tilde{\nu})$. Two classic resolution strategies can separate the contribution of n and d in the cubic term of the complexity. We can notice that the above linear system is equivalent to a discrete time Sylvester equation (Sima, 1996; Dinuzzo et al., 2011), which can be solved in $\mathcal{O}(n^3 + d^3 + n^2 d + nd^2)$ time. Or if we wish to test many values of λ , using the Kronecker structure, we can deduce an eigendecomposition of $\mathbf{K}_\mathcal{X} \otimes ((\Phi^\# \Phi) \mathbf{B})$ from one of $\mathbf{K}_\mathcal{X}$ and one of $(\Phi^\# \Phi) \mathbf{B}$ (Horn and Johnson, 1991, Theorem 4.2.12) in $\mathcal{O}(n^3 + d^3)$ time. For a given $\alpha \in \mathbb{R}^{d \times n}$, the predicted

Algorithm 1: Plug-in ridge estimator

Input: Sample $\tilde{\mathbf{z}}$, matrices \mathbf{B} , $\Phi^\# \Phi$
Compute: kernel matrix $\mathbf{K}_\mathcal{X} = (k(x_i, x_j))_{i,j=1}^n$
Compute: estimates $\tilde{\nu}$ of $(\langle y_i, \phi_d \rangle_{L^2(\Theta)})_{i=1, l=1}^{n, d}$
Solve: $(\mathbf{K}_\mathcal{X} \otimes ((\Phi^\# \Phi) \mathbf{B}) + n\lambda \mathbf{I}) \text{vec}(\alpha) = \text{vec}(\tilde{\nu})$
Output: Representer coefficients $\alpha \in \mathbb{R}^{d \times n}$.

function at a new input point $x \in \mathcal{X}$ is then given by $\Phi \mathbf{B} \alpha \mathbf{k}_\mathcal{X}(x)$ with $\mathbf{k}_\mathcal{X}(x) := (k(x, x_i))_{i=1}^n$.

4.3 Iterative optimization

For other losses, since it is no longer possible to find a closed-form, we resort to iterative optimization.

Fully observed setting For \mathbf{K} separable, using Proposition 4.1 and defining $\ell_{y_i}(y) := \ell(y_i, y)$; Problem (7) is rewritten as

$$\min_{\alpha \in \mathbb{R}^{d \times n}} \frac{1}{n} \sum_{i=1}^n \ell_{y_i}(\Phi \mathbf{B} \alpha \mathbf{k}_\mathcal{X}(x_i)) + \lambda \langle \mathbf{K}_\mathcal{X}, \alpha^T \mathbf{B} \alpha \rangle_{\mathbb{R}^{n \times n}}.$$

The gradient of the objective is given by

$$\frac{1}{n} \mathbf{B} \Phi_{\text{mat},(n)}^\# \mathbf{G}(\alpha) \mathbf{K}_\mathcal{X} + \lambda \mathbf{B} \alpha \mathbf{K}_\mathcal{X}, \quad (11)$$

with $\mathbf{G}(\alpha) := (\nabla \ell_{y_i}(\Phi \mathbf{B} \alpha \mathbf{k}_\mathcal{X}(x_i)))_{i=1}^n \in L^2(\Theta)^n$ and $\nabla \ell_{y_i} : L^2(\Theta) \mapsto L^2(\Theta)$ the gradient of ℓ_{y_i} .

Partially observed setting. We notice that the entries of $\Phi_{\text{mat},(n)}^\# \mathbf{G}(\alpha) \in \mathbb{R}^{d \times n}$ are the scalar products $(\langle \nabla \ell_{y_i}(\Phi \mathbf{B} \alpha \mathbf{k}_\mathcal{X}(x_i)), \phi_l \rangle_{L^2(\Theta)})_{l,i=1}^{d,n}$. For ℓ an integral loss (Equation (1)) based on a differentiable ground loss l , $\nabla \ell_{y_i} : y \mapsto (\theta \mapsto l(y_i(\theta), y(\theta)))$. We can thus estimate the columns $\Phi^\# \nabla \ell_{y_i}(\Phi \mathbf{B} \alpha \mathbf{k}_\mathcal{X}(x_i))$ as

$$\frac{1}{m_i} \sum_{p=1}^{m_i} l(y_i(\theta_{ip}), \phi(\theta_{ip})^T \mathbf{B} \mathbf{k}_\mathcal{X}(x_i)) \phi(\theta_{ip}), \quad (12)$$

where we have used the convention that for $\theta \in \Theta$, $\phi(\theta) := (\phi_l(\theta))_{l=1}^d \in \mathbb{R}^d$. The corresponding estimation of $\Phi_{\text{mat},(n)}^\# \mathbf{G}(\alpha)$ can be plugged into Equation (11) to yield an estimated gradient.

Link with ridge estimator. In the partially observed setting, for the square loss, iterative optimization and the plug-in ridge estimator do not yield the same result. In fact, they correspond to two different ridge closed-forms (see Section F.2 of the Supplement). While the former is slower to compute than the latter it can be more robust (see Section 6.3).

5 THEORETICAL ANALYSIS

In this section we give two finite sample excess risk bounds. One for the ridge estimator in the fully observed setting and one for the plug-in ridge estimator in the partially observed setting. In the first case, we study the effect of the number of samples n , and in the second case that of both n and the number of observations per function m . We suppose that for all $i \in [n]$, $m_i = m$. We leave however a detailed analysis with respect to the size of the dictionary d (including approximation aspects) for future work. Our analysis is based on the framework of integral operators (Caponnetto and De Vito, 2007; Smale and Zhou, 2007) to which we give an introduction in the context of our problem in Section C of the Supplement.

5.1 Fully observed setting

In this section, we suppose that \mathcal{X} is a separable metric space. We also need to relate the $L^2(\Theta)$ norm of any $g \in \text{Span}(\phi)$ to the square norm of its coefficients in the dictionary ϕ . To that end, a usual assumption is that it is a *Riesz family* (Casazza, 2000).

Definition 5.1. (Riesz family) $\phi \in L^2(\Theta)^d$ is a Riesz family of $L^2(\Theta)$ with constants (c_ϕ, C_ϕ) if it is linearly independent and for any $u \in \mathbb{R}^d$,

$$c_\phi \|u\|_{\mathbb{R}^d} \leq \left\| \sum_{l=1}^d u_l \phi_l \right\|_{L^2(\Theta)} \leq C_\phi \|u\|_{\mathbb{R}^d}.$$

If in addition for all $l \in [d]$, $\|\phi_l\|_{L^2(\Theta)} = 1$, it is a normed Riesz family.

Remark. Riesz families provide a natural generalization of orthonormal families as a normed Riesz family with $c_\phi = C_\phi = 1$ is orthonormal.

We make the following assumptions.

Assumption 5.1. 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 independent 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 independent of d .

Assumption 5.2. The dictionary ϕ is a normed Riesz family in $L^2(\Theta)$ with upper constant C_ϕ .

Remark. We do not use the lower constant c_ϕ .

Assumption 5.3. There exist $h_{\mathcal{H}_K} \in \mathcal{H}_K$ such that $h_{\mathcal{H}_K} = \inf_{h \in \mathcal{H}_K} \mathcal{R}(\Phi \circ h)$.

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}_K containing $h_{\mathcal{H}_K}$, as a consequence $\|h_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq R$.

Assumption 5.4. There exists $L \geq 0$ such that for all $\theta \in \Theta$, almost surely $|\mathbf{Y}(\theta)| \leq L$.

We then have the following excess risk bound for the ridge estimator defined in Proposition 4.2. We prove it in Section E.1 of the Supplement.

Proposition 5.1. 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 $1 - \eta$

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

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

This bound implies the consistency of the ridge estimator in the number of samples n .

5.2 Partially observed setting

To treat the partially observed setting, we need to make the following additional assumption.

Assumption 5.5. 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 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 included in ϕ , whereas for a Fourier dictionary we have $M(d) = 1$.

We then have the following excess risk bound for the plug-in ridge estimator from Definition 4.1 which we prove in Section E.2 of the Supplement.

Proposition 5.2. 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 $1 - \eta$,

$$\begin{aligned} & \mathcal{R}(\Phi \circ \tilde{h}_{\bar{z}}^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) \\ & \leq \left(\frac{B_2(d)\sqrt{n}}{m^2} + \frac{B_3(d)}{m^{3/2}} + \frac{9C(d)^2}{2\sqrt{nm}} + \frac{B_4(d)}{\sqrt{n}} \right) \log(6/\eta), \end{aligned}$$

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 5.1.

We highlight that if $m \asymp \sqrt{n}$, then this bounds yields consistency for the plug-in ridge estimator.

6 NUMERICAL EXPERIMENTS

Section 6.3 is dedicated to the study of several aspects of robustness of KPL algorithms. Then we compare KPL with the nonlinear FOR methods presented in Section 6.1 on two datasets. In Section 6.4 we explore a biomedical imaging dataset with relatively small number of samples ($n = 100$) and partially observed functions, whereas in Section 6.5 we study a speech inversion dataset with relatively large number of samples ($n = 413$) and fully observed output functions.

We use the mean squared error (MSE) as metric. Given observed functions $((\theta_i, \tilde{y}_i))_{i=1}^n$ and predicted ones $(\hat{y}_i)_{i=1}^n \in L^2(\Theta)$, we define it as $\text{MSE} := \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i} \sum_{p=1}^{m_i} (\hat{y}_i(\theta_{ip}) - \tilde{y}_{ip})^2$. The presented results are averaged either over 10 or 20 runs with different train/test splits. Full details of the experimental procedures are postponed to Section H of the Supplement.

6.1 Related works

We compare KPL to four existing nonlinear FOR methods that we present in this section. More detailed descriptions are given in Section G of the Supplement.

Functional kernel ridge regression (FKRR). Kadri et al. (2010, 2016) solve a functional KRR using function-valued-RKHSs. A representer theorem yields a closed-form solution computed by inverting an operator in $\mathcal{L}(L^2(\Theta))^{n \times n}$. For a separable kernel $\mathsf{K}^{\text{fun}} = kL$ with $L \in \mathcal{L}(L^2(\Theta))$, if an eigendecomposition of L is known in closed-form, an approximate solution is computed in $\mathcal{O}(n^3 + n^2 J m)$ time, with J the number of eigenfunctions considered and m the size of the discretization grid. If not, a discretized problem is solved in $\mathcal{O}(n^3 + m^3 + n^2 m + nm^2)$ time.

Triple basis estimator (3BE). In (Oliva et al., 2015), the input and the output functions are represented by decomposition coefficients on two orthonormal families. The output coefficients are then regressed

on the input ones using KRRs approximated with J RFFs in $\mathcal{O}(J^3 + J^2d)$ time, with d the size of the output family. As 3BE is specific to function-to-function regression with scalar-valued inputs, we deal with vector-valued input functions (as in Section 6.5), directly through a kernel. We call this extension **one basis estimator (1BE)**; it is solved in $\mathcal{O}(n^3 + n^2d)$ time. 1BE is in fact a particular case of the KPL plug-in ridge estimator with ϕ orthonormal and $K = kl$. However, our estimator offers the additional possibility to use non orthonormal dictionaries and to impose richer regularizations through kernels $K = kB$ with $B \neq I$. KPL can moreover can be used with a wide range of functional losses.

Kernel additive model (KAM). Reimherr et al. (2018) propose an additive function-to-function regression model using RKHSs. A representer theorem leads to a closed-form solution. Computations are performed in a truncated FPCA basis of size $J < n$. For a product of kernels, if the Kronecker structure is exploited (a possibility which is however not highlighted by the authors), the complexity is $\mathcal{O}(n^3 + J^3 + n^2J + nJ^2)$ time using a Sylvester solver. However, computing the matrix to form the linear system—matrix A in page 6 of (Reimherr et al., 2018)—is generally much more expensive; exploiting the product of kernels, $n^2 + J^2$ double integrals must be computed which has time complexity $\mathcal{O}(n^2t^2 + J^2m^2)$, with t the size of the input discretization grid. Those computations must moreover generally be repeated many times so as to tune the multiple kernel parameters.

Kernel Estimator (KE). Finally, an extension of the Nadaraya-Watson kernel estimator to Banach spaces is introduced and studied in (Ferraty et al., 2011).

6.2 Preliminary elements

Note on optimization. We compute the KPL plug-in ridge estimator as in Algorithm 1 with Sylvester solver. For iterative optimization, we use L-BFGS-B (Zhu et al., 1997a); the estimates of partial second order informations improve convergence speed. For FKRR, of the two possible approaches from Section G of the Supplement, we use the faster Sylvester approach. For KAM we exploit the separability as well using a Sylvester solver.

Logcosh functional loss. As an example of a robust integral loss, for $\gamma > 0$, we introduce $\ell_{\text{ich}}^{(\gamma)}$. It is obtained by taking $\ell_{\text{ich}}^{(\gamma)} : (a, b) \mapsto 1/\gamma \log(\cosh(\gamma(a - b)))$ as ground loss in Equation (1). This ground loss behaves similarly to the Huber loss (Huber, 1964)—almost quadratically around 0 and almost linearly elsewhere. The parameter γ gives us control on its behaviour around 0, as it grows bigger, $\ell_{\text{ich}}^{(\gamma)}$ tends to the absolute

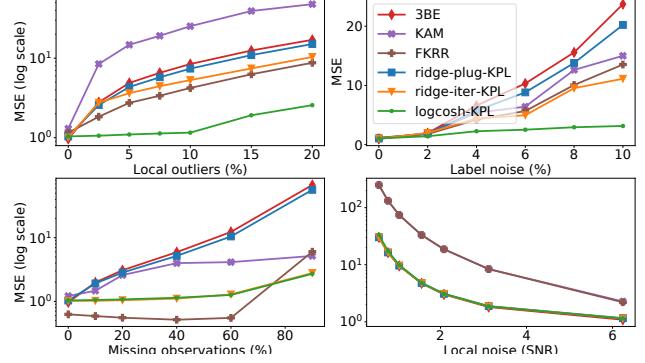


Figure 1: Several aspects of robustness.

loss (see Section H for examples). As opposed to our proposed integral loss $\ell_{\text{ich}}^{(\gamma)}$, the extension of the Huber loss to $L^2(\Theta) \times L^2(\Theta)$ (e. g. Bauschke and Combettes, 2017, Example 13.7) is not differentiable everywhere.

6.3 Toy data

In this section, we take $K = kl$ with k a scalar-valued Gaussian kernel. We use a generated toy dataset: inputs are random mixtures of cubic B-splines (de Boor, 2001) centered at different locations and outputs are associated mixtures of Gaussian processes (drawn once and then fixed). The full generation procedure is described in Section H of the Supplement. We use $n_{\text{train}} = 100$ samples for training and $n_{\text{test}} = 100$ samples and use Fourier dictionaries for KPL and 3BE.

Corruption modalities. We study the effect of four types of corruptions of the training data: local outliers, label noise, missing observations and local noise. In the first case, observations from the output functions are replaced with random draws in their range. In the second case, some output functions are replaced with erroneous ones. In the third case we remove observations from the output functions uniformly at random. Finally, in the last one we add Gaussian noise to those observations. We then use the signal to noise ratio as x-axis; for a noise level σ and a sample \tilde{z} , we define it as $\text{SNR} := \frac{1}{\sigma n} \sum_{i=1}^n \frac{1}{m_i} \sum_{p=1}^{m_i} |\tilde{y}_{ip}|$.

Comments on the results. The evolution of the MSEs for several levels of corruption are displayed in Figure 1. For each type, at least one KPL algorithm is particularly robust which demonstrates the versatility of our framework. KPL can be combined with the functional logcosh loss to obtain a FOR algorithm that is robust to outliers (*logcosh-KPL*). Dealing with partially observed functions, KPL solved iteratively using estimated gradients works especially well (*ridge-iter-KPL*, *logcosh-KPL*). Finally all proposed KPL algorithms are robust to local noise.

Table 1: MSEs on the DTI dataset.

KE	0.231 ± 0.025
3BE	0.227 ± 0.017
KAM	0.222 ± 0.021
FKRR	0.215 ± 0.020
RIDGE-KPL	0.211 ± 0.022
LOGCOSH-KPL	0.209 ± 0.020

6.4 Diffusion tensor imaging dataset (DTI)

Dataset. We now consider the DTI dataset.¹ It 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, however the structure of this process is not well understood. Using the proxy of FA profiles, we propose to predict one tract (RCS) from the other (CCA). We consider only the first $n = 100$ scans of MS patients. Finally, we highlight that the functions are partially observed: significant parts of the FA profiles along the RCS tract are missing.

Experimental setting. We perform linear smoothing if necessary—for FKRR and KAM. We split the data as $n_{\text{train}} = 70$ and $n_{\text{test}} = 30$ and use wavelets dictionaries for 3BE and KPL. For KPL, we take a kernel of the form $\mathbf{K} = k\mathbf{D}$ with k a Gaussian kernel and \mathbf{D} a diagonal matrix with diagonal decreasing with the corresponding wavelet scale. Finally, when using wavelets, we extend the signal symmetrically to avoid boundary effects. The MSEs are shown in Table 1.

Comments on the results. The studied methods perform almost equally well, with a slight advantage for ours. The combination of an efficient use of wavelets (well suited to non-smooth data) with the scale-dependant regularization induced by the kernel $\mathbf{K} = k\mathbf{D}$ may explain this.

6.5 Synthetic speech inversion dataset

Dataset. We consider a speech inversion problem: from an acoustic speech signal, we estimate the underlying vocal tract (VT) configuration that produced it (Richmond, 2002). Such information can improve performance in speech recognition systems or in speech synthesis. The dataset was introduced by Mitra et al. (2009); it is generated by a software synthesizing words

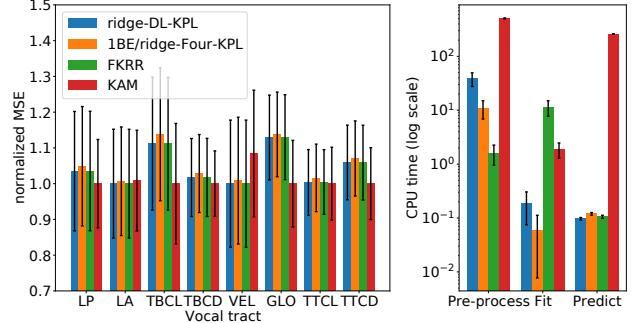


Figure 2: MSEs and CPU times on the speech dataset.

from an articulatory model. It consists of a corpus of $n = 413$ pronounced words with 8 distinct VT functions: lip aperture (LA), lip protrusion (LP), tongue tip constriction degree (TTCD), tongue tip constriction location (TTCL), tongue body constriction degree (TBCD), tongue body constriction location (TBCL), Velum (VEL) and Glottis (GLO).

Experimental setting. To match words of varying lengths, we extend symmetrically both the input sounds and the VT functions matching the longest word. We represent the sounds using 13 mel-frequency cepstral coefficients (MFCC), the input data thus consist of vector-valued functions. We split the data as $n_{\text{train}} = 300$ and $n_{\text{test}} = 113$. We normalize the output functions so that they take their values in $[-1, 1]$. To deal with the vector-valued functional inputs, we use an integral of Gaussian kernels on the standardized MFCCs (KPL, FKRR, 1BE/KPL). For KAM we take Laplace kernels for both input and output locations, and use a Gaussian kernel defined on \mathbb{R}^{13} to compare the evaluations of the standardized MFCCs (see Section H of the Supplement for details on the employed kernels).

The MSEs for the 8 VTs (left panel) as well as an analysis of the computation times (right panel) are displayed in Figure 2. *Pre-process* entails all pre-processing operations (e. g. computing the kernel matrices, learning the dictionary, computing the gram matrix of ϕ), *fit* measures the fitting time per se (solving the relevant linear system) and *predict* measures the prediction time on the test set (for all methods, it entails computing new kernel matrices). *ridge-DL-KPL* is the KPL ridge estimator with ϕ learnt by solving Problem (6) with \mathcal{C} and $\Omega_{\mathbb{R}^d}$ as introduced in Section 3.2. *1BE/ridge-Four-KPL* corresponds to 1BE (or equivalently KPL with $\mathbf{K} = k\mathbf{I}$) with ϕ a Fourier family. To give an order of idea, we use 30 atoms for the learnt dictionaries while the numbers of atoms selected by cross-validation for the Fourier ones are around 100. We do not include KE in the figure as it performed poorly on this dataset.

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

Comments on the results. For 4 out of 8 VTs (LP, LA, TBCD, TTCL), the performances of the methods are comparable, with KAM being slightly more precise. On the remaining 4 VTs, ridge-DL-KPL, 1BE/ridge-Four-KPL and FKRR beat KAM on one (VEL) and are beaten by KAM on the 3 other (TBCL, GLO, TTCD). This could be explained by the fact that KAM predicts locally the functions while the other three methods have more of a global approach. Depending on the properties of the functions and the nature of the dependency between input and output functions, one or the other could be more favorable. However KAM's main weakness is its computational cost for pre-processing and prediction, which makes it unpractical to use on medium-sized datasets and impossible to use on larger ones. The particularly time-consuming operation in question is the computation of an analogous to the kernel matrix (see Section 6.1). The three other methods display very close MSEs, with 1BE/ridge-Four-KPL being a bit less precise than the two others. Ridge-DL-KPL and FKRR perform equally well. However for the former the main computational burden comes from a pre-processing operation (learning the dictionary), which is performed only once per dataset (or once per fold in a cross-validation); whereas for the latter it comes from fitting the method, which must be done many times so as to tune its parameters. Moreover for Ridge-DL-KPL, once a number of atoms yielding a good approximation has been found and the dictionary has been learnt, no further tuning must be performed for the outputs, whereas for FKRR an output kernel must be chosen.

7 CONCLUSION

We introduced PL, a general dictionary-based framework to address FOR. It can be used with a wide class of functional losses and non orthonormal dictionaries. Through an extensive study in the context of vv-RKHSs, we illustrated some aspects of its versatility and demonstrated that the approach is efficient and can be backed theoretically in some cases. For future research, PL could be instantiated using other hypothesis classes than vv-RKHS and the possibilities offered by dictionary learning could be investigated further.

Acknowledgements

The authors thank Zoltán Szabó for his insightful feedbacks. This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

References

- A. M. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.
- 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.
- T. T. Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39: 2330–2355, 2011.
- 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.
- P. G. Casazza. The art of frame theory. *Taiwanese journal of mathematics*, 4:129–201, 2000.
- 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.
- R. A. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *American Journal of Mathematics*, 114(4):737–785, 1992.
- F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 49–56, 2011.
- B. Dumitrescu and P. Irofti. *Dictionary Learning, Algorithms and Applications*. Springer, 2018.
- F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer, 2006.
- 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.

- R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- P. J. Huber. Robust estimation of a location parameter. *Annals of Statistics*, 53:73–101, 1964.
- 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.
- P. Kokoska and M. Reimherr. *Introduction to Functional Data Analysis*. CRC Press, 2017.
- H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS) 19*, pages 801–808, 2007.
- Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38:3321–3351, 2010.
- Z. Li, J.-F. Ton, D. Ogle, 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.
- H. Lian. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics*, pages 597–606, 2007.
- J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 689–696, 2009.
- S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2008.
- C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.
- V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C. Y. Espy-Wilson. From acoustics to vocal tract time functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4500, 2009.
- J. S. Morris. Functional regression. *The Annual Review of Statistics and Its Application*, 2:321–359, 2015.
- 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.
- P. Oswald. On the degree of nonlinear spline approximation in Besov-Sobolev spaces. *Journal of approximation theory*, 61(2):131–157, 1990.
- 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., 2008a.
- A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561, 2008b.
- 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.
- K. Richmond. *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Center for Speech Technology Research, Edinburgh University, 2002.
- E. Senkene and A. Templeman. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 1973.
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- V. Sima. *Algorithms for Linear-Quadratic Optimization*. Chapman and Hall/CRC, 1996.
- 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.
- C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transations on Mathematical Software*, 1997a.
- 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, 1997b.

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 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. The Python code is provided in a separate zip file.

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 $\mathsf{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{U})$ 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 $(x_1, \dots, x_n) \in \mathcal{X}^n$, for all $(u_1, \dots, u_n) \in \mathcal{U}^n$,

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

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 $\mathsf{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{U})$. For any $x \in \mathcal{X}$, we define K_x as*

$$\mathsf{K}_x : u \mapsto \mathsf{K}_x u, \quad \text{with} \quad \mathsf{K}_x u : x' \mapsto \mathsf{K}(x', x) u. \quad (13)$$

There exists a unique Hilbert space \mathcal{H}_{K} of functions $h : \mathcal{X} \rightarrow \mathcal{U}$ 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 $h \in \mathcal{H}_{\mathsf{K}}$,

$$\langle \mathsf{K}_x u, h \rangle_{\mathcal{H}_{\mathsf{K}}} = \langle u, h(x) \rangle_{\mathcal{U}}. \quad (14)$$

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

The scalar product on \mathcal{H}_{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 $\|h\|_{\mathcal{H}_{\mathsf{K}}}^2 = \langle h, h \rangle_{\mathcal{H}_{\mathsf{K}}}$.

This RKHS \mathcal{H}_{K} can be built by taking the closure of the set $\{\mathsf{K}_x u \mid x \in \mathcal{X}, u \in \mathcal{U}\}$ 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 $\mathcal{H}_K \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ a vv-RKHS associated to a positive matrix-valued kernel K . Then we have for all $x \in \mathcal{X}$:

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

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

$$\|K_x\|_{\mathcal{L}(\mathbb{R}^d, \mathcal{H}_K)} = \|K_x^\#\|_{\mathcal{L}(\mathcal{H}_K, \mathbb{R}^d)} \leq \|K(x, x)\|_{\mathcal{L}(\mathbb{R}^d)}^{1/2}. \quad (15)$$

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 $K : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{L}(\mathbb{R}^d)$ an OVK with $\mathcal{H}_K \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ its associated vv-RKHS, we want to solve the following optimization problem

$$\min_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i)) + \lambda \|h\|_{\mathcal{H}_K}^2. \quad (16)$$

Proposition B.1. (*Representer theorem.*) For ℓ continuous and convex with respect to its second argument, Problem (16) admits a unique minimizer $h_\mathbf{z}^\lambda$. Moreover there exists $\alpha \in \mathbb{R}^{d \times n}$ such that $h_\mathbf{z}^\lambda = \sum_{j=1}^n K_{x_j} \alpha_j$.

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

Let $\mathcal{U} := \left\{ h \mid h = \sum_{j=1}^n K_{x_j} \alpha_j, \alpha \in \mathbb{R}^{d \times n} \right\}$. Since it is a closed subspace of \mathcal{H}_K , $\mathcal{H}_K = \mathcal{U} \oplus \mathcal{U}^\perp$ and we can decompose $h_\mathbf{z}^\lambda$ as $h_\mathbf{z}^\lambda = h_{\mathbf{z}, \mathcal{U}}^\lambda + h_{\mathbf{z}, \mathcal{U}^\perp}^\lambda$ with $(h_{\mathbf{z}, \mathcal{U}}^\lambda, h_{\mathbf{z}, \mathcal{U}^\perp}^\lambda) \in \mathcal{U} \times \mathcal{U}^\perp$. We recall that $\phi \in L^2(\Theta)^d = (\phi_l)_{l=1}^d$ 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:

$$(\Phi h_\mathbf{z}^\lambda(x_i))(\theta) = \langle \phi(\theta), h_\mathbf{z}^\lambda(x_i) \rangle_{\mathbb{R}^d} = \langle K_{x_i} \phi(\theta), h_\mathbf{z}^\lambda \rangle_{\mathcal{H}_K}.$$

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

$$(\Phi h_\mathbf{z}^\lambda(x_i))(\theta) = \langle K_{x_i} \phi(\theta), h_{\mathbf{z}, \mathcal{U}}^\lambda \rangle_{\mathcal{H}_K} = \langle \phi(\theta), h_{\mathbf{z}, \mathcal{U}}^\lambda(x_i) \rangle_{\mathbb{R}^d} = (\Phi h_{\mathbf{z}, \mathcal{U}}^\lambda(x_i))(\theta).$$

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

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):

$$\min_{\alpha \in \mathbb{R}^{d \times n}} \frac{1}{n} \|\mathbf{y} - \Phi_{(n)} \mathbf{K} \text{vec}(\alpha)\|_{L^2(\Theta)^n}^2 + \lambda \langle \text{vec}(\alpha), \mathbf{K} \text{vec}(\alpha) \rangle_{\mathbb{R}^{dn}}. \quad (17)$$

Proposition B.2. (*Ridge solution*) The minimum in Problem (17) is achieved by any $\alpha^* \in \mathbb{R}^{d \times n}$ verifying

$$(\mathbf{K}(\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{K}) \text{vec}(\alpha^*) := \mathbf{K} \Phi_{(n)}^\# \mathbf{y}. \quad (18)$$

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

$$\text{vec}(\alpha^*) = ((\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I})^{-1} \Phi_{(n)}^\# \mathbf{y}. \quad (19)$$

We then define the ridge estimator as $h_\mathbf{z}^\lambda := \sum_{j=1}^n K_{x_j} \alpha_j^*$.

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

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

Up to an additional term not dependant on $\boldsymbol{\alpha}$, this corresponds to the objective function in Problem (17) where we have set $\boldsymbol{\alpha} = \text{vec}(\alpha)$ to simplify the exposition.

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

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

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

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

We want to prove that $\boldsymbol{\alpha}^*$ is then a solution to Problem (17). Observe now that

$$\begin{aligned} \langle \boldsymbol{\alpha}^*, \mathbf{K} ((\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I}) \boldsymbol{\alpha} \rangle_{\mathbb{R}^{dn}} &= \langle \boldsymbol{\alpha}, \mathbf{K} ((\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I}) \boldsymbol{\alpha}^* \rangle_{\mathbb{R}^{dn}} \\ &= \langle \boldsymbol{\alpha}, \mathbf{K} \Phi_{(n)}^\# \mathbf{y} \rangle_{\mathbb{R}^{dn}} \\ &= \langle \Phi_{(n)}^\# \mathbf{y}, \mathbf{K} \boldsymbol{\alpha} \rangle_{\mathbb{R}^{dn}}. \end{aligned} \quad (20)$$

Using Equation (20), we deduce that

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

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

We now show that Equation (18) always has a solution $\boldsymbol{\alpha}^*$ in \mathbb{R}^{dn} and conclude with the special case where \mathbf{K} is full rank. Note that $(\mathbf{K} (\Phi^\# \Phi)_{(n)} \mathbf{K} + n\lambda \mathbf{I})$ is a positive symmetric matrix and its null space is exactly that of \mathbf{K} . Hence it is bijective on the image of \mathbf{K} , which shows that Equation (18) always has a solution. If \mathbf{K} is moreover full rank then

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

is also invertible and we can simplify by \mathbf{K} on both sides of Equation (18) and obtain the claimed formula for $\boldsymbol{\alpha}^*$. Taking $\alpha^* \in \mathbb{R}^{d \times n}$ such that $\text{vec}(\alpha^*) = \boldsymbol{\alpha}^*$ yields the desired results. \square

C LEARNING THEORY AND INTEGRAL OPERATORS

This section is devoted to the study of Problem (16) 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 $f \in \mathcal{F}(\mathcal{X}, \mathbb{L}^2(\Theta))$

$$\mathcal{R}(f) := \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \rho} \left[\|\mathbf{Y} - f(\mathbf{X})\|_{\mathbb{L}^2(\Theta)}^2 \right], \quad (21)$$

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

$$\widehat{\mathcal{R}}(f, \mathbf{z}) := \frac{1}{n} \sum_{i=1}^n \|y_i - f(x_i)\|_{\mathbb{L}^2(\Theta)}^2. \quad (22)$$

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

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

and its associated norm $\|\cdot\|_\rho$. Then, the expected risk in Equation (21) of a regressor f can then be equivalently formulated as

$$\mathcal{R}(f) = \|f \circ X - Y\|_\rho^2, \quad (23)$$

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

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

$$\mathbf{A}_\Phi : h \mapsto \mathbf{A}_\Phi h \text{ with } (\mathbf{A}_\Phi h) : (x, y) \in \mathcal{Z} \mapsto \Phi K_x^\# h. \quad (24)$$

We can reformulate the expected risk in terms of \mathbf{A}_Φ for any $h \in \mathcal{H}_K$,

$$\|\mathbf{A}_\Phi h - Y\|_\rho^2 = \int_{\mathcal{Z}} \|\Phi K_x^\# h - y\|_{\mathbb{L}^2(\Theta)}^2 d\rho(x, y) = \int_{\mathcal{Z}} \|\Phi h(x) - y\|_{\mathbb{L}^2(\Theta)}^2 d\rho(x, y) = \mathcal{R}(\Phi \circ h). \quad (25)$$

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

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

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

Then, for all $h \in \mathcal{H}_K$,

$$\langle h, \mathbf{T}_\Phi h_{\mathcal{H}_K} - \mathbf{A}_\Phi^\# Y \rangle_{\mathcal{H}_K} = 0; \quad (26)$$

or equivalently:

$$\mathbf{T}_\Phi h_{\mathcal{H}_K} = \mathbf{A}_\Phi^\# Y, \quad (27)$$

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

Proof. We use the formulation of the expected risk from Equation (25). The function $h \mapsto \mathcal{R}(\Phi \circ h) = \|\mathbf{A}_\Phi h - Y\|_\rho^2$ is convex as a convex function composed with an affine mapping. Its differential is given by

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

We then must have for all $h \in \mathcal{H}_K$,

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

□

Using the formulation of the expected risk from Equation (25) as well as the characterization of $h_{\mathcal{H}_K}$ in Equation (26), for any $h \in \mathcal{H}_K$, we can then reformulate the excess risk of h as a distance in \mathcal{H}_K between h and $h_{\mathcal{H}_K}$ taken through the operator T_Φ .

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

$$\mathcal{R}(\Phi \circ h) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) = \|\sqrt{T_\Phi}(h - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}^2. \quad (28)$$

Proof.

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

where we have used Equation (26). Since we have the following polar decomposition $A_\Phi = U\sqrt{A_\Phi^\# A_\Phi} = U\sqrt{T_\Phi}$ with U a partial isometry from the closure of $\text{Im}(\sqrt{T_\Phi})$ onto the closure of $\text{Im}(A_\Phi)$,

$$\|A_\Phi(h - h_{\mathcal{H}_K})\|_\rho = \|U\sqrt{T_\Phi}(h - h_{\mathcal{H}_K})\|_\rho = \|\sqrt{T_\Phi}(h - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}.$$

□

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 $A_\Phi^\#$ 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 $K_{x,\Phi} := K_x \Phi^\#$ and $T_{x,\Phi} := K_{x,\Phi} K_{x,\Phi}^\#$.

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

$$A_\Phi^\# \psi = \int_{\mathcal{Z}} K_{x,\Phi} \psi(x, y) d\rho(x, y), \quad (29)$$

with the integral converging in \mathcal{H}_K . And $A_\Phi^\# A_\Phi$ is the Hilbert Schmidt operator on \mathcal{H}_K given by

$$A_\Phi^\# A_\Phi = T_\Phi = \int_{\mathcal{X}} T_{x,\Phi} d\rho_X(x), \quad (30)$$

with the integral converging in $L_2(\mathcal{H}_K)$.

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

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

$$(A_{\mathbf{x},\Phi} h)_i = K_{x_i,\Phi}^\# h = \Phi h(x_i), \quad h \in \mathcal{H}_K, \quad \forall i \in [n].$$

$$T_{\mathbf{x},\Phi} = A_{\mathbf{x},\Phi}^\# A_{\mathbf{x},\Phi} = \frac{1}{n} \sum_{i=1}^n T_{x_i,\Phi}.$$

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

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

the following closed form for its minimizer can be derived.

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

$$h_\mathbf{z}^\lambda := (\mathsf{T}_{\mathbf{x}, \Phi} + \lambda I)^{-1} \mathsf{A}_{\mathbf{x}, \Phi}^\# \mathbf{y}. \quad (31)$$

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

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

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

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

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

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

□

Importantly, $h_\mathbf{z}^\lambda$ 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 $\phi := (\phi_1, \dots, \phi_d)$ be a Riesz family, let Φ be its associated projection operator (see Definition 2.1 from the main paper). Then*

$$\|\Phi\|_{\mathcal{L}(\mathbb{R}^d, L^2(\Theta))} \leq C_\phi \quad (32)$$

$$\|\Phi^\#\|_{\mathcal{L}(L^2(\Theta), \mathbb{R}^d)} \leq C_\phi \quad (33)$$

$$\|\Phi^\# \Phi\|_{\mathcal{L}(\mathbb{R}^d)} \leq C_\phi^2. \quad (34)$$

Proof. Equation (32) is a direct consequence of the definition of a Riesz family (Definition 5.1 from the main paper). Since the operator Φ is bounded, $\|\Phi^\#\|_{\mathcal{L}(L^2(\Theta), \mathbb{R}^d)} = \|\Phi\|_{\mathcal{L}(\mathbb{R}^d, L^2(\Theta))}$ implying Equation (33). Finally combining the two inequalities yields Equation (34). □

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

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

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

- $\mathbf{K}_{x,\Phi} : L^2(\Theta) \longrightarrow \mathcal{H}_K$ is defined by $\mathbf{K}_{x,\Phi} := \mathbf{K}_x \Phi^\#$ with \mathbf{K}_x as defined in Equation (13).
- $\mathbf{T}_{x,\Phi} : \mathcal{H}_K \longrightarrow \mathcal{H}_K$ is defined as $\mathbf{T}_{x,\Phi} := \mathbf{K}_{x,\Phi} \mathbf{K}_{x,\Phi}^\#$.

Observe that $\mathbf{T}_{x,\Phi}$ 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 $x \in \mathcal{X}$,*

$$\|\mathbf{K}(x, x)\|_{\mathcal{L}(\mathbb{R}^d)} \leq \kappa, \quad (35)$$

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

$$\|\mathbf{T}_{x,\Phi}\|_{\mathcal{L}_2(\mathcal{H}_K)} \leq \sqrt{d} \kappa C_\phi^2. \quad (36)$$

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

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

Using Cauchy-Schwartz in the previous expression along with Equation (34), Equation (35) and Equation (15) we have that

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

which achieves the proof. \square

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 $\sqrt{\mathbf{T}_{x,\Phi}}$ from concentration properties of $\mathbf{T}_{x,\Phi}$.

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*

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

Let $n \in \mathbb{N}$ and (ξ_1, \dots, ξ_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 $(U_i)_{i=1}^n$ be independent random variables taking their values in a real separable Hilbert space \mathcal{K} such that for all $i \in [n]$*

$$\mathbb{E}[U_i] = 0,$$

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

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

Let $0 < \eta < 1$, then

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

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 $A, B \in \mathcal{L}(\mathcal{K})$ be two positive operators. Then, we have*

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

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 $x \in \mathcal{X}$, $\|K(x, x)\|_{\mathcal{L}(\mathbb{R}^d)} \leq \kappa$.*

Remark. We suppose that κ is independant from d . This is for instance the case if for $x \in \mathcal{X}$, $K(x, x)$ 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 $K(x, x)$, which can easily be imposed to be independent of d .

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

Remark. We do not use the lower constant c_ϕ .

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

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}_K containing $h_{\mathcal{H}_K}$, as a consequence

$$\|h_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq R. \quad (37)$$

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

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

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 $1 - \eta$

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

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 $1 - \eta$*

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

with δ_1 defined as

$$\delta_1(n, \eta) := 6(\sqrt{\kappa}C_\phi L + \kappa C_\phi^2 R) \frac{\log(2/\eta)}{\sqrt{n}}. \quad (38)$$

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

Observe that

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

and using Equation (27), that

$$\mathbb{E}_{X, Y \sim \rho} [\xi_1(X, Y)] = \int_{\mathcal{Z}} K_{x, \Phi} y \, d\rho(x, y) - \left(\int_{\mathcal{Z}} K_{x, \Phi} K_{x, \Phi}^\# \, d\rho(x, y) \right) h_{\mathcal{H}_K} = A_\Phi^\# Y - T_\Phi h_{\mathcal{H}_K} = 0.$$

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

$$\begin{aligned} \|\xi_1(X, Y)\|_{\mathcal{H}_K} &= \|K_{X, \Phi}(Y - \Phi h_{\mathcal{H}_K}(X))\|_{\mathcal{H}_K} \leq \|K_{X, \Phi}\|_{\mathcal{L}(\mathcal{L}^2(\Theta), \mathcal{H}_K)} \|Y - \Phi h_{\mathcal{H}_K}(X)\|_{\mathcal{L}^2(\Theta)} \\ &\leq \sqrt{\kappa}C_\phi (\|Y\|_{\mathcal{L}^2(\Theta)} + \|K_{X, \Phi}^\# h\|_{\mathcal{L}^2(\Theta)}) \\ &\leq \sqrt{\kappa}C_\phi (L + \sqrt{\kappa}C_\phi R), \end{aligned} \quad (39)$$

where we have used the inequality $\|K_{X, \Phi}\|_{\mathcal{L}(\mathcal{L}^2(\Theta), \mathcal{H}_K)} = \|K_{X, \Phi}^\#\|_{\mathcal{L}(\mathcal{L}^2(\Theta), \mathcal{H}_K)} \leq \sqrt{\kappa}C_\phi$ (immediate consequence of Equations (32) and (15)), as well as Assumptions E.4 and E.3.

Equation (39) also implies

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

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

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

□

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

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

with δ_2 defined as

$$\delta_2(n, d, \eta) := 6\kappa C_\phi^2 \frac{\log(2/\eta) \sqrt{d}}{\sqrt{n}}. \quad (40)$$

Proof. We introduce the $\xi_2 : \mathcal{Z} \rightarrow \mathcal{L}_2(\mathcal{H}_K)$ as $\xi_2 : x, y \mapsto T_{x,\Phi}$.

We have that

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

And from Equation (36), we have almost surely

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

which implies as well

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

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

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

□

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

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

with δ_1 and δ_2 defined respectively in Equations (38) and (40).

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

$$\lambda \geq 6\kappa C_\phi^2 \frac{\log(4/\eta) \sqrt{d}}{\sqrt{n}} = \delta_2(n, d, \eta/2), \quad (41)$$

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

$$\mathcal{R}(\Phi \circ h_z^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) \leq \frac{9}{2} \left(\frac{36(\sqrt{\kappa} C_\phi L + \kappa C_\phi^2 R)^2 \log(4/\eta)^2}{\lambda n} + \lambda R^2 \right). \quad (42)$$

Proof. We introduce h^λ as

$$h^\lambda := (T_{x,\Phi} + \lambda I)^{-1} T_{x,\Phi} h_{\mathcal{H}_K}. \quad (43)$$

We consider the following decomposition of the risk using Equation (28),

$$\begin{aligned} \mathcal{R}(\Phi \circ h_z^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) &= \|\sqrt{T_\Phi}(h_z^\lambda - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}^2 \\ &\leq 2\|\sqrt{T_\Phi}(h_z^\lambda - h^\lambda)\|_{\mathcal{H}_K}^2 + 2\|\sqrt{T_\Phi}(h^\lambda - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}^2. \end{aligned} \quad (44)$$

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

$$\begin{aligned}\sqrt{\mathbf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda) &= \sqrt{\mathbf{T}_{x,\Phi}}(\mathbf{T}_{x,\Phi} + \lambda I)^{-1}(\mathbf{A}_{x,\Phi}^\# \mathbf{y} - \mathbf{T}_{x,\Phi} h_{\mathcal{H}_K}) \\ &\quad + (\sqrt{\mathbf{T}_\Phi} - \sqrt{\mathbf{T}_{x,\Phi}})(\mathbf{T}_{x,\Phi} + \lambda I)^{-1}(\mathbf{A}_{x,\Phi}^\# \mathbf{y} - \mathbf{T}_{x,\Phi} h_{\mathcal{H}_K}).\end{aligned}\tag{45}$$

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

$$\|\sqrt{\mathbf{T}_{x,\Phi}}(\mathbf{T}_{x,\Phi} + \lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H}_K)} \leq \max_{a \in \text{Sp}(\mathbf{T}_{x,\Phi})} \frac{\sqrt{a}}{a+\lambda} \leq \max_{a \in \mathbb{R}_+} \frac{\sqrt{a}}{a+\lambda} \leq \frac{1}{2\sqrt{\lambda}},\tag{46}$$

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

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

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

Taking the norm in Equation (45), applying Minkowski's inequality and using Lemma D.5 as well as the last two displays yields

$$\|\sqrt{\mathbf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_K} \leq \|\mathbf{A}_{x,\Phi}^\# \mathbf{y} - \mathbf{T}_{x,\Phi} h_{\mathcal{H}_K}\|_{\mathcal{H}_K} \left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\|\mathbf{T}_\Phi - \mathbf{T}_{x,\Phi}\|_{\mathcal{L}(\mathcal{H}_K)}}}{\lambda} \right).\tag{47}$$

Now dealing with the term on the right-hand side in Equation (44), using the definition of h^λ in Equation (43), we have that

$$\begin{aligned}\sqrt{\mathbf{T}_\Phi}(h_{\mathcal{H}_K} - h^\lambda) &= \sqrt{\mathbf{T}_\Phi}(I - (\mathbf{T}_{x,\Phi} + \lambda I)^{-1}\mathbf{T}_{x,\Phi})h_{\mathcal{H}_K} \\ &= (\sqrt{\mathbf{T}_\Phi} - \sqrt{\mathbf{T}_{x,\Phi}})(I - (\mathbf{T}_{x,\Phi} + \lambda I)^{-1}\mathbf{T}_{x,\Phi})h_{\mathcal{H}_K} \\ &\quad + \sqrt{\mathbf{T}_{x,\Phi}}(I - (\mathbf{T}_{x,\Phi} + \lambda I)^{-1}\mathbf{T}_{x,\Phi})h_{\mathcal{H}_K}.\end{aligned}\tag{48}$$

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 (46) yields

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

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

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

Thus, taking the norm in Equation (48), using Minkowski's inequality, Lemma D.5 and Equation (37) yields

$$\|\sqrt{\mathbf{T}_\Phi}(h_{\mathcal{H}_K} - h^\lambda)\|_{\mathcal{H}_K} \leq R\sqrt{\|\mathbf{T}_\Phi - \mathbf{T}_{x,\Phi}\|_{\mathcal{L}(\mathcal{H}_K)}} + \frac{R}{2}\sqrt{\lambda}.\tag{49}$$

Combining Equations (47) and (49) with Lemma E.3, for $0 < \eta < 1$, we have with probability at least $1 - \eta$

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

Using the condition on λ given by Equation (41), still with probability at least $1 - \eta$, we have

$$\|\sqrt{\mathbf{T}_\Phi}(h_{\mathbf{z}}^\lambda - h^\lambda)\|_{\mathcal{H}_K} \leq \frac{3}{2\sqrt{\lambda}}\delta_1(n, \eta/2), \quad (50)$$

$$\|\sqrt{\mathbf{T}_\Phi}(h_{\mathcal{H}_K} - h^\lambda)\|_{\mathcal{H}_K} \leq \frac{3R}{2}\sqrt{\lambda}. \quad (51)$$

Combining Equations (50) and (51) into Equation (44) yields that with probability at least $1 - \eta$,

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

□

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 (41), 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 $1 - \eta$,

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

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:

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

where for all $i \in [n]$, $\theta_i \in \Theta^{m_i}$, $\tilde{y}_i \in \mathbb{R}^{m_i}$ 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 $\tilde{y}_{ip} \in \mathbb{R}$. 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 $\tilde{\mathbf{y}} := (\tilde{y}_i)_{i=1}^n$ and highlight that since there is no added noise, we have for all $i \in [n]$

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

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 $\tilde{\Phi}_i \in \mathbb{R}^{m \times d}$ the approximation of Φ using the locations θ_i as

$$\tilde{\Phi}_i := (\phi_1(\theta_i), \dots, \phi_d(\theta_i)),$$

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 (31)) reads:

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

with

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

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

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

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

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

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 $1 - \eta$*

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

with δ_3 defined as

$$\delta_3(n, m, d, \eta) := \left(\frac{4(L\sqrt{\kappa}\sqrt{d}M(d) + \sqrt{\kappa}C_\phi R)}{m} + \frac{2L\sqrt{\kappa}\sqrt{d}M(d)}{\sqrt{n}\sqrt{m}} \right) \log(2/\eta). \quad (52)$$

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

The proof relies on the fact that

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

Let $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ 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

$$\begin{aligned} W_{ip} &:= \xi_3(\mathbf{X}_i, \mathbf{Y}_i, \vartheta_{ip}) \\ &= \mathbf{Y}_i(\vartheta_{ip})\mathbf{K}_{\mathbf{X}_i} \phi(\vartheta_{ip}) - \mathbf{K}_{\mathbf{X}_i} \Phi^\# \mathbf{Y}_i \\ &= \mathbf{Y}_i(\vartheta_{ip})\mathbf{K}_{\mathbf{X}_i} \phi(\vartheta_{ip}) - \mathbb{E}[\mathbf{Y}_i(\vartheta)\mathbf{K}_{\mathbf{X}_i} \phi(\vartheta)|\mathbf{X}_i, \mathbf{Y}_i], \end{aligned} \quad (53)$$

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

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

$$\mathbb{P}[\cdot | \mathbf{z}] = \mathbb{P}[\cdot | X_i = x_i, Y_i = y_i, i \in [n]]$$

Then, Equation (53) implies that $\mathbb{E}[W_{ip} | \mathbf{z}] = 0$.

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

We have almost surely that

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

We have used Assumptions E.4 and E.5 as well as Equation (33).

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

$$\mathbb{E}[\|\bar{W}_p\|_{\mathcal{H}_K}^2 | \mathbf{z}] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[\|W_{ip}\|_{\mathcal{H}_K}^2 | \mathbf{z}]. \quad (54)$$

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

$$\mathbb{E}[\|W_{ip}\|_{\mathcal{H}_K}^2 | \mathbf{z}] = \mathbb{E}[\|\mathbb{Y}_i(\vartheta_{ip}) \mathsf{K}_{X_i} \phi(\vartheta_{ip})\|_{\mathcal{H}_K}^2 | \mathbf{z}]. \quad (55)$$

Then using Equation (55) into Equation (54) along with Assumptions E.4 and E.5 yields

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

We can then apply Lemma D.4 to obtain that

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

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

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

□

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

$$\|\mathsf{A}_{\mathbf{x}, \Phi}^\# \mathbf{y} - \mathsf{T}_{\mathbf{x}, \Phi} h_{\mathcal{H}_K}\|_{\mathcal{H}_K} \leq \delta_1(n, \eta/3) \quad (56)$$

$$\|\mathsf{T}_{\mathbf{x}, \Phi} - \mathsf{T}_\Phi\|_{\mathcal{L}_2(\mathcal{H}_K)} \leq \delta_2(n, d, \eta/3) \quad (57)$$

$$\|\mathsf{A}_{\mathbf{x}, \tilde{\Phi}}^\# \tilde{\mathbf{y}} - \mathsf{A}_{\mathbf{x}, \Phi}^\# \mathbf{y}\|_{\mathcal{H}_K} \leq \delta_3(n, m, d, \eta/3), \quad (58)$$

with δ_1 , δ_2 and δ_3 respectively defined as in Equations (38), (40) and (52).

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*

$$\lambda \geq 6\kappa C_\phi^2 \frac{\log(6/\eta) \sqrt{d}}{\sqrt{n}} = \delta_2(n, d, \eta/3), \quad (59)$$

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

$$\mathcal{R}(\Phi \circ \tilde{h}_z^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) \leq \frac{27}{4} \left(\left(\frac{A_0(d)^2}{\lambda m^2} + \frac{2A_0(d)A_1(d)}{\lambda \sqrt{nm}^{3/2}} + \frac{A_1(d)^2}{\lambda nm} + \frac{A_2^2}{\lambda n} \right) \log(6/\eta)^2 + \lambda R^2 \right), \quad (60)$$

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 (43), we consider the following decomposition of the risk using Equation (28)

$$\begin{aligned} \mathcal{R}(\Phi \circ \tilde{h}_z^\lambda) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) &= \|\sqrt{T_\Phi}(\tilde{h}_z^\lambda - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}^2 \\ &\leq 3\|\sqrt{T_\Phi}(\tilde{h}_z^\lambda - h_z^\lambda)\|_{\mathcal{H}_K}^2 + 3\|\sqrt{T_\Phi}(h_z^\lambda - h^\lambda)\|_{\mathcal{H}_K}^2 + 3\|\sqrt{T_\Phi}(h^\lambda - h_{\mathcal{H}_K})\|_{\mathcal{H}_K}^2. \end{aligned} \quad (61)$$

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 (47), we get that

$$\|\sqrt{T_\Phi}(\tilde{h}_z^\lambda - h_z^\lambda)\|_{\mathcal{H}_K} \leq \|A_{x, \tilde{\Phi}}^\# \tilde{y} - A_{x, \Phi}^\# y\|_{\mathcal{H}_K} \left(\frac{1}{2\sqrt{\lambda}} + \frac{\sqrt{\|\mathbf{T}_\Phi - \mathbf{T}_{x, \Phi}\|_{\mathcal{L}(\mathcal{H}_K)}}}{\lambda} \right). \quad (62)$$

Combining Equations (47) , (49) and (62) with Lemma E.5, for $0 < \eta < 1$, the three following inequalities are verified with probability at least $1 - \eta$

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

Using the condition on λ given by Equation (59), still with probability at least $1 - \eta$, we have

$$\|\sqrt{T_\Phi}(\tilde{h}_z^\lambda - h_z^\lambda)\|_{\mathcal{H}_K} \leq \frac{3}{2\sqrt{\lambda}} \delta_3(n, m, d, \eta/3) \quad (63)$$

$$\|\sqrt{\mathbf{T}_\Phi}(h_z^\lambda - h^\lambda)\|_{\mathcal{H}_K} \leq \frac{3}{2\sqrt{\lambda}} \delta_1(n, \eta/3) \quad (64)$$

$$\|\sqrt{\mathbf{T}_\Phi}(h_{\mathcal{H}_K} - h^\lambda)\|_{\mathcal{H}_K} \leq \frac{3R}{2}\sqrt{\lambda}. \quad (65)$$

Combining Equation (63), (64) and (65) into Equation (61) yields that with probability at least $1 - \eta$,

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

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 (59)), 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. \square

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):

$$\min_{h \in \mathcal{H}} \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda \Omega_{\mathcal{H}}(h), \quad (66)$$

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

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

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 $x \in \mathcal{X}$, $\mathbf{w} \mapsto h_{\mathbf{w}}$ is differentiable. The gradient is given by

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

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

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

Then replacing $h_{\mathbf{w}}$ by the regressor corresponding to the vv-RKHS hypothesis class with separable kernel: $x \mapsto \mathbf{B}k(x)$, 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 $\tilde{\mathbf{z}}$:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i} \sum_{p=1}^{m_i} l(y_i(\theta_{ip}), \phi(\theta_{ip})^T h_{\mathbf{w}}(x_i)). \quad (67)$$

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

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

$$\tilde{\Phi}_i := (\phi_1(\theta_i), \dots, \phi_d(\theta_i)),$$

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

$$\min_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n \left\| \frac{\tilde{y}_i}{\sqrt{m_i}} - \frac{\tilde{\Phi}_i}{\sqrt{m_i}} h(x_i) \right\|_{\mathbb{R}^{m_i}}^2 + \lambda \|h\|_{\mathcal{H}_K}^2 \quad (68)$$

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

Then using Proposition 4.1 from the main paper, we can rewrite Problem (67) as

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

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

$$\text{vec}(\alpha^*) = ((\tilde{\Phi}^\# \tilde{\Phi}) \mathbf{K} + n\lambda \mathbf{I})^{-1} \tilde{\Phi}^\# \text{vec}(\tilde{\mathbf{y}}). \quad (69)$$

We remark that $\tilde{\Phi}^\# \text{vec}(\tilde{\mathbf{y}}) \in \mathbb{R}^{dn}$ 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 $\tilde{\Phi}^\# \text{vec}(\tilde{\mathbf{y}}) = \text{vec}(\tilde{\nu})$. Then the only difference with the plug-in ridge estimator is that the matrix $(\Phi^\# \Phi)_{(n)}$ is replaced by the matrix $(\tilde{\Phi}^\# \tilde{\Phi})$ which is block-diagonal with the matrices $\left(\frac{1}{m_i} \tilde{\Phi}_i^\# \tilde{\Phi}_i\right)_{i=1}^n$ 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:

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

with $\mathcal{H}_{K^{\text{fun}}}$ the fv-RKHS associated to some OVK $K^{\text{fun}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Y})$, 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:

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

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

We now focus on the case of the separable kernel $K^{\text{fun}}(x, x') = k^{\text{in}}(x, x') \mathbf{L}$. k^{in} is a scalar-valued kernel and $\mathbf{L} \in \mathcal{L}(\mathcal{Y})$ 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 \mathbf{L} :

$$\mathbf{L} \mathbf{y} : \theta' \mapsto \int_{\theta \in \Theta} \exp\left(-\frac{|\theta' - \theta|}{\sigma_{k^{\text{out}}}}\right) d\theta. \quad (70)$$

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

1. An eigendecomposition can be performed. If such decomposition of \mathbf{L} is known in closed-form, the Kronecker product can be exploited to solve the system in $\mathcal{O}(n^3 + n^2 Jm)$ 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., 1997b). 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 $\mathcal{O}(n^3 + m^3 + n^2 m + nm^2)$ 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: $(\beta^{\text{in}}, \beta^{\text{out}})$ with $\beta^{\text{in}} \in \mathbb{R}^{n \times c}$ and $\beta^{\text{out}} \in \mathbb{R}^{n \times d}$; $c \in \mathbb{N}^*$ being the cardinality of the input basis and $d \in \mathbb{N}^*$ that of the output basis. Then, each set of output coefficient (β_l^{out} for $l \in [d]$) is regressed on the input coefficients β^{in} using KRRs approximated with RFFs (Rahimi and Recht, 2008a). Denoting by $\mathbf{R}(\beta^{\text{in}}) \in \mathbb{R}^{n \times J}$ the matrix of RFFs evaluated on the input coefficients β^{in} , for all $l \in [d]$, the following (scalar-valued) sub-problem is solved:

$$\min_{c_l \in \mathbb{R}^J} \|\beta_l^{\text{out}} - \mathbf{R}(\beta^{\text{in}})c_l\|_{\mathbb{R}^n}^2 + \lambda \|c_l\|_{\mathbb{R}^J}^2.$$

All those sub-problems require the inversion of the same matrix $(\mathbf{R}(\beta^{\text{in}})^T \mathbf{R}(\beta^{\text{in}}) + \lambda \mathbf{I})$, 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 $\mathbf{K} = k\mathbf{I}$. In that case, the time complexity is $\mathcal{O}(n^3 + n^2 d)$ (we 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:

$$\sum_{i=1}^n \int_0^1 \left(y_i(\theta) - a(\theta) - \int_0^1 b(\zeta, \theta) x_i(\zeta) d\zeta \right)^2 d\theta. \quad (71)$$

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

$$\begin{aligned} a(\theta) &= \mathbf{A} e^{\text{out}}(\theta) \\ b(\zeta, \theta) &= (e^{\text{in}}(\zeta))^T \mathbf{B} e^{\text{out}}(\theta). \end{aligned}$$

Then, we use those expressions for a and b and minimize the objective from Equation (71) in the variables $\mathbf{A} \in \mathbb{R}^{1 \times d}$ and $\mathbf{B} \in \mathbb{R}^{c \times d}$. 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

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

where $\mathcal{H}_{k^{\text{add}}}$ is the RKHS of a scalar-valued kernel $k^{\text{add}} : ([0, 1] \times [0, 1] \times \mathbb{R})^2 \rightarrow \mathbb{R}$ and $\lambda > 0$. A representer theorem leads to a closed-form 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 $\mathcal{O}(n^3 J^3)$. However, if k^{add} is chosen as a product of three kernels, the separability property can be exploited to solve the problem in $\mathcal{O}(n^3 + J^3 + n^2 J + nJ^2)$ 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 $\mathcal{O}(n^2 t^2 + J^2 m^2)$ 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:

$$k^{\text{add}} : ((\zeta, \theta, s), (\zeta', \theta', s)) \mapsto \exp\left(\frac{-(\zeta - \zeta')^2}{\sigma_1^2}\right) \exp\left(\frac{-(\theta - \theta')^2}{\sigma_2^2}\right) \exp\left(\frac{-(s - s')^2}{\sigma_3^2}\right). \quad (72)$$

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 $k^{\text{add}} : ([0, 1] \times [0, 1] \times \mathbb{R}^o)^2 \rightarrow \mathbb{R}$ 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 $K : \mathbb{R} \mapsto \mathbb{R}$ 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:

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

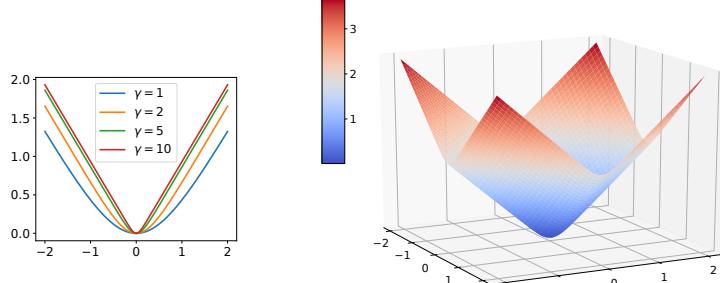

 Figure 3: Logcosh loss on \mathbb{R} .

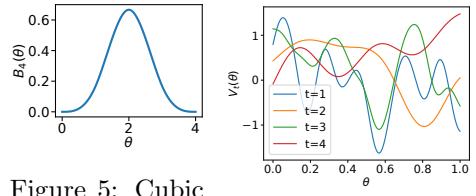
 Figure 4: Logcosh loss on \mathbb{R}^2 ($\gamma = 5$).


Figure 5: Cubic B-spline.

Figure 6: GP draws.

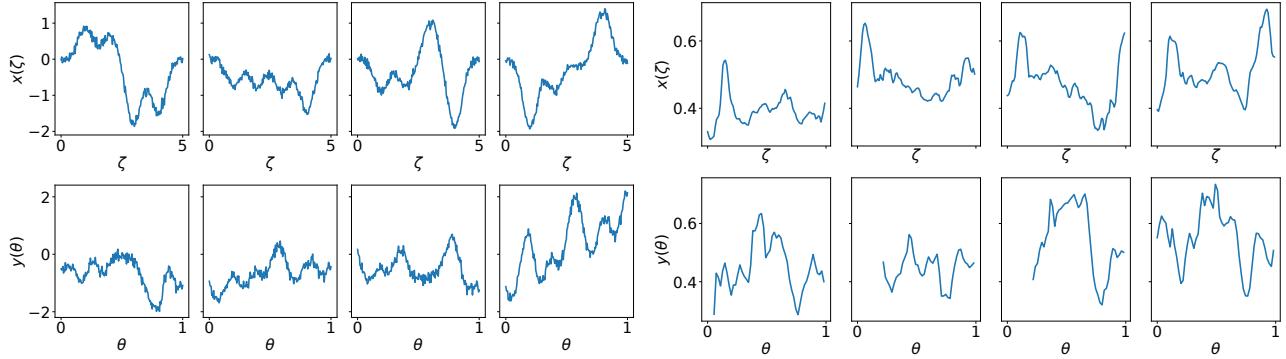


Figure 7: Examples of generated toy data.

Figure 8: Examples from the DTI dataset.

It corresponds to the loss $l_{\text{LCH}}^{(\gamma)}$ defined in Section 6.2 from the main paper. We illustrate the effect of the parameter γ in Figure 3.

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

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

We plot this loss for $\gamma = 5$ in Figure 4.

H.2 Toy dataset

H.2.1 Generating process

We consider a functional toy dataset. To generate it, we draw $r \in \mathbb{N}$ independent zero mean Gaussian processes (GP) with Gaussian covariance functions. More precisely, for $t \in [r]$ the Gaussian process V_t has covariance $(\theta_1, \theta_2) \mapsto \exp\left(-\frac{(\theta_2 - \theta_1)^2}{b_t^2}\right)$. 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 6. 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 5). Let then $\bar{B}_4 : \zeta \mapsto B_4(4\zeta + 2)$ (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 7.

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 $K(x, x') := k(x, x')I$ with k a scalar-valued Gaussian kernel with standard deviation $\sigma_k = 20$ and $I \in \mathbb{R}^{d \times d}$ the identity matrix. When using the logcosh loss, the parameter γ is set to $\gamma = 25$ for the 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 (72) 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 (70) 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) dataset¹ 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 8.

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 $K(x, x') = k(x, x')D$ 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 $\{7.5, 10, 12.5, 15, 17.5, 20\}$.

¹This 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 (72) 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 (70) 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:

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

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

$$(\tilde{x}_0, \tilde{x}_1) \mapsto \frac{1}{m} \sum_{p=1}^m \exp\left(\frac{-\|\tilde{x}_{1p} - \tilde{x}_{0p}\|_2^2}{\sigma^2}\right). \quad (73)$$

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

$$((\zeta, \theta, w), (\zeta', \theta', w')) \mapsto \exp\left(\frac{-|\zeta - \zeta'|}{\sigma_1}\right) \exp\left(\frac{-|\theta - \theta'|}{\sigma_2}\right) \exp\left(\frac{-\|w - w'\|_2^2}{\sigma_3^2}\right). \quad (74)$$

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 $\text{avg}^{(r)} := \frac{1}{n_{\text{train}} m} \sum_{i=1}^{n_{\text{train}}} \sum_{p=1}^m \tilde{x}_{ip}^{(r)}$ and $\text{std}^{(r)} := \sqrt{\frac{1}{n_{\text{train}} m - 1} \sum_{i=1}^{n_{\text{train}}} \sum_{p=1}^m (\tilde{x}_{ip}^{(r)} - \text{avg}^{(r)})^2}$, and use as input data $\left(\left(\frac{\tilde{x}_{ip}^{(r)}}{\text{std}^{(r)}}\right)_{r=1}^{13}\right)_{i=1}^{n_{\text{train}}}$.

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 (73) 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 (70) 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 (74) 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 (73) 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 (70) 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 (74) 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 9. 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 \mathbf{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 9, 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.

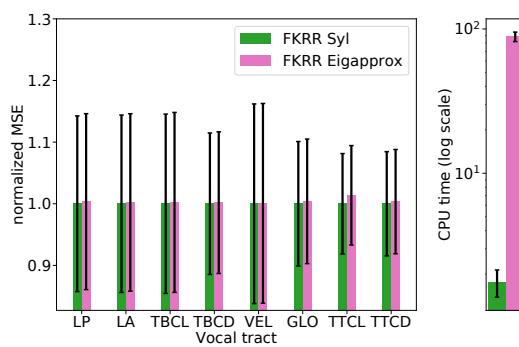


Figure 9: Comparison of two solvers for FKRR on speech dataset