Multiple Operator-valued Kernel Learning

Hachem Kadri, Alain Rakotomamonjy, Francis Bach, Philippe Preux

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

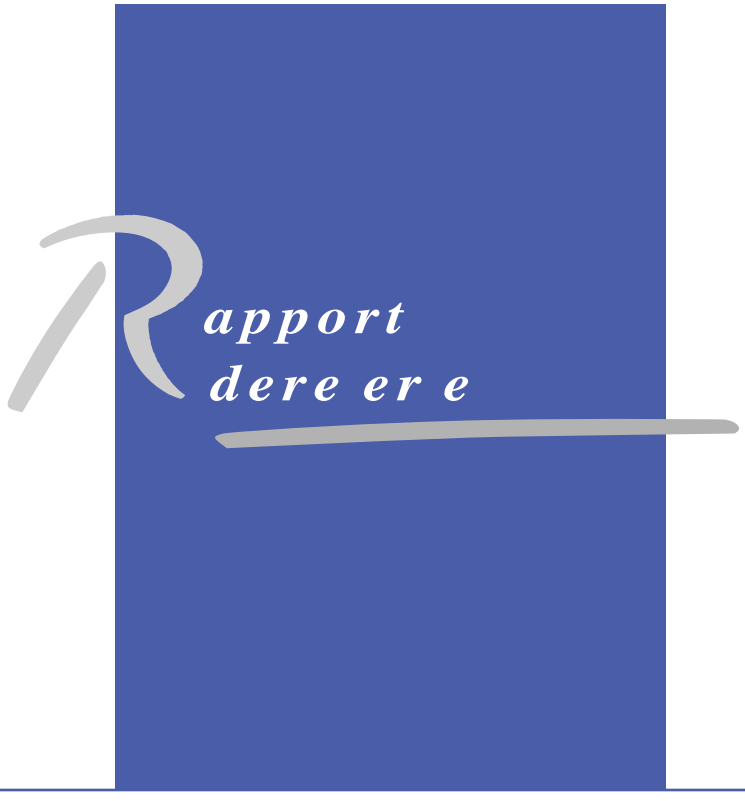
Multiple Operator-valued Kernel Learning

Hachem Kadri — Alain Rakotomamonjy — Francis Bach — Philippe Preux

N° 7900

February 2012

Thème COG



ISSN 0249-6399 ISRN INRIA/RR--7900--FR+ENG

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche LILLE - NORD EUROPE

Multiple Operator-valued Kernel Learning

Hachem Kadri*, Alain Rakotomamonjy†, Francis Bach‡, Philippe Preux§

Thème COG — Systèmes cognitifs
Équipes-Projets SequeL

Rapport de recherche n° 7900 — February 2012 — 20 pages

Abstract: Positive definite operator-valued kernels generalize the well-known notion of reproducing kernels, and are naturally adapted to multi-output learning situations. This paper addresses the problem of learning a finite linear combination of infinite-dimensional operator-valued kernels which are suitable for extending functional data analysis methods to nonlinear contexts. We study this problem in the case of kernel ridge regression for functional responses with an ℓ_r-norm constraint on the combination coefficients ($r \geq 1$). The resulting optimization problem is more involved than those of multiple scalar-valued kernel learning since operator-valued kernels pose more technical and theoretical issues. We propose a multiple operator-valued kernel learning algorithm based on solving a system of linear operator equations by using a block coordinate-descent procedure. We experimentally validate our approach on a functional regression task in the context of finger movement prediction in brain-computer interfaces.

Key-words: Operator-valued kernels, multiple kernel learning, nonparametric functional data analysis, function-valued reproducing kernel Hilbert spaces.

* Sequel Team, INRIA Lille. E-mail: hachem.kadri@inria.fr
† LITIS, Université de Rouen. E-mail: alain.rakotomamonjy@insa-rouen.fr
‡ Sierra Team/INRIA, Ecole Normale Supérieure. E-mail: francis.bach@inria.fr
§ Sequel/INRIA-Lille, LIFL/CNRS. E-mail: philippe.preux@inria.fr

Centre de recherche INRIA Lille – Nord Europe
Parc Scientifique de la Haute Borne
40, avenue Halley, 59650 Villeneuve d'Ascq
Téléphone : +33 3 59 57 78 00 — Télécopie : +33 3 59 57 78 50

Apprentissage de Noyaux à Valeurs Opérateurs Multiples

Résumé : Dans cet article, nous proposons une méthode d'apprentissage de noyaux multiples à valeurs opérateurs dans le cas d'une régression ridge à réponse fonctionnelle. Notre méthode est basée sur la résolution d'un système d'équations linéaires d'opérateurs en utilisant une procédure de type Iterative Coordinate Descent. Nous validons expérimentalement notre approche sur un problème de prédiction de mouvement de doigt dans un contexte d'Interface Cerveau-Machine.

Mots-clés : noyaux à valeurs opérateurs, apprentissage de noyaux multiples, analyse des données fonctionnelles, espace de Hilbert à noyau reproduisant

1 Introduction

During the past decades, a large number of algorithms have been proposed to deal with learning problems in the case of single-valued functions (*e.g.*, binary-output function for classification or real output for regression). Recently, there has been considerable interest in estimating vector-valued functions [20, 5, 7]. Much of this interest has arisen from the need to learn tasks where the target is a complex entity, not a scalar variable. Typical learning situations include multi-task learning [11], functional regression [12], and structured output prediction [4].

In this paper, we are interested in the problem of functional regression with functional responses in the context of brain-computer interface (BCI) design. More precisely, we are interested in finger movement prediction from electrocorticographic signals [27]. Indeed, from a set of signals measuring brain surface electrical activity on d channels during a given period of time, we want to predict, for any instant of that period whether a finger is moving or not and the amplitude of the finger flexion. Formally, the problem consists in learning a functional dependency between a set of d signals and a sequence of labels (a step function indicating whether a finger is moving or not) and between the same set of signals and vector of real values (the amplitude function). While, it is clear that this problem can be formalized as functional regression problem, from our point of view, such problem can benefit from the multiple operator-valued kernel learning framework. Indeed, for these problems, one of the difficulties arises from the unknown latency between the signal related to the finger movement and the actual movement [22]. Hence, instead of fixing in advance some value for this latency in the regression model, our framework allows to learn it from the data by means of several operator-valued kernels.

If we wish to address functional regression problem in the principled framework of reproducing kernel Hilbert spaces (RKHS), we have to consider RKHSs whose elements are operators that map a function to another function space, possibly source and target function spaces being different. Working in such RKHSs, we are able to draw on the important core of work that has been performed on scalar-valued and vector-valued RKHSs [29, 20]. Such a functional RKHS framework and associated operator-valued kernels have been introduced recently [12, 13]. A basic question with reproducing kernels is how to build these kernels and what is the optimal kernel choice for a given application. In order to overcome the need for choosing a kernel before the learning process, several works have tried to address the problem of learning the scalar-valued kernel jointly with the decision function [17, 30]. Since these seminal works, many efforts have been carried out in order to theoretically analyze the kernel learning framework [9, 3] or in order to provide efficient algorithms [23, 1, 14]. While many works have been devoted to multiple *scalar-valued* kernel learning, this problem of kernel learning have been barely investigated for *operator-valued* kernels. One motivation of this work is to bridge the gap between multiple kernel learning (MKL) and operator-valued kernels by proposing a framework and an algorithm for learning a finite linear combination of operator-valued kernels. While each step of the scalar-valued MKL framework can be extended without major difficulties to operator-valued kernels, technical challenges arise at all stages because we deal with infinite dimensional spaces. It should be pointed out that in a recent work [10], the problem of learning the output kernel was

formulated as an optimization problem over the cone of positive semidefinite matrices, and proposed a block-coordinate descent method to solve it. However, they did not focus on learning the input kernel. In contrast, our multiple operator-valued kernel learning formulation can be seen as a way of learning simultaneously input and output kernels, although we consider a linear combination of kernels which are fixed in advance.

In this paper, we make the following contributions:

- we introduce a novel approach to infinite-dimensional multiple operator-valued kernel learning (MovKL) suitable for learning the functional dependencies and interactions between continuous data,
- we extend the original formulation of ridge regression in dual variables to the functional data analysis domain, showing how to perform nonlinear functional regression with functional responses by constructing a linear regression operator in an operator-valued kernel feature space (Section 2),
- we derive a dual form of the MovKL problem with functional ridge regression, and show that a solution of the related optimization problem exists (Section 2),
- we propose a block-coordinate descent algorithm to solve the MovKL optimization problem which involves solving a challenging linear system with a sum of block operator matrices, and show its convergence in the case of compact operator-valued kernels (Section 3),
- we provide an empirical evaluation of MovKL performance which demonstrates its effectiveness on a BCI dataset (Section 4).

2 Problem Setting

Before describing the multiple operator-valued kernel learning algorithm that we will study and experiment with in this paper, we first review notions and properties of reproducing kernel Hilbert spaces with operator-valued kernels, show their connection to learning from multiple response data (multiple outputs; see [20] for discrete data and [12] for continuous data), and describe the optimization problem for learning kernels with functional response ridge regression.

2.1 Notations and Preliminaries

We start by some standard notations and definitions used all along the paper. Given a Hilbert space $\mathcal{H}$, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ refer to its inner product and norm, respectively. We denote by $\mathcal{G}_x$ and $\mathcal{G}_y$ the separable real Hilbert spaces of input and output functional data, respectively. In functional data analysis domain, continuous data are generally assumed to belong to the space of square integrable functions L^2. In this work, we consider that $\mathcal{G}_x$ and $\mathcal{G}_y$ are the Hilbert space $L^2(\Omega)$ which consists of all equivalence classes of square integrable functions on a finite set Ω. Ω being potentially different for $\mathcal{G}_x$ and $\mathcal{G}_y$. We denote by $\mathcal{F}(\mathcal{G}_x, \mathcal{G}_y)$ the vector space of functions $f : \mathcal{G}_x \longrightarrow \mathcal{G}_y$, and by $\mathcal{L}(\mathcal{G}_y)$ the set of bounded linear operators from $\mathcal{G}_y$ to $\mathcal{G}_y$.

We consider the problem of estimating a function f such that $f(x_i) = y_i$ when observed functional data $(x_i, y_i)_{i=1,\ldots,n} \in (\mathcal{G}_x, \mathcal{G}_y)$. Since $\mathcal{G}_x$ and $\mathcal{G}_y$ are

spaces of functions, the problem can be thought of as an operator estimation problem, where the desired operator maps a Hilbert space of factors to a Hilbert space of targets. We can define the regularized operator estimate of $f \in \mathcal{F}$ as:

$$f_\lambda \triangleq \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \|y_i - f(x_i)\|^2_{\mathcal{G}_y} + \lambda\|f\|^2_{\mathcal{F}}. \tag{1}$$

In this work, we are looking for a solution to this minimization problem in a function-valued reproducing kernel Hilbert space $\mathcal{F}$. More precisely, we mainly focus on the RKHS $\mathcal{F}$ whose elements are continuous linear operators on $\mathcal{G}_x$ with values in $\mathcal{G}_y$. The continuity property is obtained by considering a special class of reproducing kernels called Mercer kernels [7, Proposition 2.2]. Note that in this case, $\mathcal{F}$ is separable since $\mathcal{G}_x$ and $\mathcal{G}_y$ are separable [6, Corollary 5.2].

We now introduce (function) $\mathcal{G}_y$-valued reproducing kernel Hilbert spaces and show the correspondence between such spaces and positive definite (operator) $\mathcal{L}(\mathcal{G}_y)$-valued kernels. These extend the traditional properties of scalar-valued kernels.

Definition 1 *(function-valued RKHS)*
A Hilbert space $\mathcal{F}$ of functions from $\mathcal{G}_x$ to $\mathcal{G}_y$ is called a reproducing kernel Hilbert space if there is a positive definite $\mathcal{L}(\mathcal{G}_y)$-valued kernel $K_{\mathcal{F}}(w,z)$ on $\mathcal{G}_x \times \mathcal{G}_x$ such that:

i. the function $z \longmapsto K_{\mathcal{F}}(w,z)g$ belongs to $\mathcal{F}$, $\forall z \in \mathcal{G}_x,\ w \in \mathcal{G}_x,\ g \in \mathcal{G}_y$,

ii. $\forall f \in \mathcal{F}, w \in \mathcal{G}_x,\ g \in \mathcal{G}_y,\quad \langle f, K_{\mathcal{F}}(w,\cdot)g\rangle_{\mathcal{F}} = \langle f(w), g\rangle_{\mathcal{G}_y}$ (reproducing property).

Definition 2 *(operator-valued kernel)*
An $\mathcal{L}(\mathcal{G}_y)$-valued kernel $K_{\mathcal{F}}(w,z)$ on $\mathcal{G}_x$ is a function $K_{\mathcal{F}}(\cdot,\cdot) : \mathcal{G}_x \times \mathcal{G}_x \longrightarrow \mathcal{L}(\mathcal{G}_y)$; furthermore:

i. $K_{\mathcal{F}}$ is Hermitian if $K_{\mathcal{F}}(w,z) = K_{\mathcal{F}}(z,w)^*$, where $*$ denotes the adjoint operator,

ii. $K_{\mathcal{F}}$ is positive definite on $\mathcal{G}_x$ if it is Hermitian and for every natural number r and all $\{(w_i,u_i)_{i=1,\ldots,r}\} \in \mathcal{G}_x \times \mathcal{G}_y$, $\sum_{i,j}\langle K_{\mathcal{F}}(w_i,w_j)u_j, u_i\rangle_{\mathcal{G}_y} \geq 0$.

Theorem 1 *(bijection between function valued RKHS and operator-valued kernel)*
An $\mathcal{L}(\mathcal{G}_y)$-valued kernel $K_{\mathcal{F}}(w,z)$ on $\mathcal{G}_x$ is the reproducing kernel of some Hilbert space $\mathcal{F}$, if and only if it is positive definite.

The proof of Theorem 1 can be found in [20]. For further reading on operator-valued kernels and their associated RKHSs, see, e.g., [5, 6, 7].

2.2 Functional Response Ridge Regression in Dual Variables

We can write the ridge regression with functional responses optimization problem (1) as follows:

$$\min_{f\in\mathcal{F}} \frac{1}{2}\|f\|^2_{\mathcal{F}} + \frac{1}{2n\lambda}\sum_{i=1}^{n}\|\xi_i\|^2_{\mathcal{G}_y} \quad \text{with } \xi_i = y_i - f(x_i). \tag{2}$$

Now, we introduce the Lagrange multipliers $\alpha_i, i = 1, \ldots, n$ which are functional variables since the output space is the space of functions $\mathcal{G}_y$. For the optimization problem (2), the Lagrangian multipliers exist and the Lagrangian function is well defined. The method of Lagrange multipliers on Banach spaces, which is a generalization of the classical (finite-dimensional) Lagrange multipliers method suitable to solve certain infinite-dimensional constrained optimization problems, is applied here. For more details, see [15]. Let $\alpha = (\alpha_i)_{i=1,\ldots,n} \in \mathcal{G}_y^n$ the vector of functions containing the Lagrange multipliers, the Lagrangian function is defined as

$$L(f, \alpha, \xi) = \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{2n\lambda}\|\xi\|_{\mathcal{G}_y^n}^2 + \langle \alpha, y - f(x) - \xi \rangle_{\mathcal{G}_y^n}, \tag{3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathcal{G}_y^n$, $y = (y_1, \ldots, y_n) \in \mathcal{G}_y^n$, $\xi = (\xi_1, \ldots, \xi_n) \in \mathcal{G}_y^n$, $f(x) = (f(x_1), \ldots, f(x_n)) \in \mathcal{G}_y^n$, and $\forall a, b \in \mathcal{G}_y^n$, $\langle a, b \rangle_{\mathcal{G}_y^n} = \sum_{i=1}^{n} \langle a_i, b_i \rangle_{\mathcal{G}_y}$.

Differentiating (3) with respect to $f \in \mathcal{F}$ and setting to zero, we obtain

$$f(.) = \sum_{i=1}^{n} K(x_i, .)\alpha_i, \tag{4}$$

where $K : \mathcal{G}_x \times \mathcal{G}_x \longrightarrow \mathcal{L}(\mathcal{G}_y)$ is the operator-valued kernel of $\mathcal{F}$.

Substituting this into (3) and minimizing with respect to ξ, we obtain the dual of the functional response ridge regression problem

$$\max_{\alpha} -\frac{n\lambda}{2}\|\alpha\|_{\mathcal{G}_y^n}^2 - \frac{1}{2}\langle \mathbf{K}\alpha, \alpha \rangle_{\mathcal{G}_y^n} + \langle \alpha, y \rangle_{\mathcal{G}_y^n}, \tag{5}$$

where $\mathbf{K} = [K(x_i, x_j)]_{i,j=1}^n$ is the block operator kernel matrix.

2.3 MovKL in Dual Variables

Let us now consider that the function $f(\cdot)$ is sum of M functions $\{f_k(\cdot)\}_{k=1}^M$ where each f_k belongs to a $\mathcal{G}_y$-valued RKHS with kernel $K_k(\cdot, \cdot)$. Similarly to scalar-valued multiple kernel learning, we adopt the convention that $\frac{x}{0} = 0$ if $x = 0$ and ∞ otherwise, and we can cast the problem of learning these functions f_k as

$$\begin{aligned} \min_{d \in \mathcal{D}} \min_{f_k \in \mathcal{F}_k} & \sum_{k=1}^{M} \frac{\|f_k\|_{\mathcal{F}_k}^2}{2d_k} + \frac{1}{2\lambda}\sum_{i=1}^{n} \|\xi_i\|_{\mathcal{G}_y}^2 \\ & \text{with } \xi_i = y_i - \textstyle\sum_{k=1}^{M} f_k(x_i), \end{aligned} \tag{6}$$

where $d = [d_1, \cdots, d_M]$, $\mathcal{D} = \{d : \forall k, d_k \geq 0 \text{ and } \sum_k d_k^r \leq 1\}$ and $1 \leq r \leq \infty$. Note that this problem can equivalently be rewritten as an unconstrained optimization problem. Before deriving the dual of this problem, it can be shown by means of the generalized Weierstrass theorem [16] that this problem admits a solution (a detailed proof is provided in the supplementary material).

Now, following the lines of [23], a dualization of this problem leads to the following equivalent one

$$\min_{\mathbf{d} \in \mathcal{D}} \max_{\alpha \in \mathcal{G}_y^n} -\frac{n\lambda}{2}\|\alpha\|_{\mathcal{G}_y^n}^2 - \frac{1}{2}\langle \mathbf{K}\alpha, \alpha \rangle_{\mathcal{G}_y^n} + \langle \alpha, y \rangle_{\mathcal{G}_y^n}, \tag{7}$$

where $\mathbf{K} = \sum_{k=1}^{M} d_k \mathbf{K}_k$ and $\mathbf{K}_k$ is the block operator kernel matrix associated to the operator-valued kernel K_k. The KKT conditions also state that at optimality we have $f_k(\cdot) = \sum_{i=1}^{n} d_k K_k(x_i, \cdot)\alpha_i$.

3 Solving the MovKL Problem

After having presented the framework, we now devise an algorithm for solving this MovKL problem.

3.1 Block-coordinate descent algorithm

Since the optimization problem (6) has the same structure as a multiple scalar-valued kernel learning problem, we can build our MovKL algorithm upon the MKL literature. Hence, we propose to borrow from [14], and consider a block-coordinate descent method. The convergence of a block coordinate descent algorithm which is related closely to the Gauss-Seidel method was studied in works of [31] and others. The difference here is that we have operators and block operator matrices rather than matrices and block matrices, but this doesn't increase the complexity if the inverse of the operators are computable (typically analytically or by spectral decomposition). Our algorithm iteratively solves the problem with respect to α with d being fixed and then with respect to d with α being fixed (see Algorithm 1). After having initialized $\{d_k\}$ to non-zero values, this boils down to the following steps :

1. with $\{d_k\}$ fixed, the resulting optimization problem with respect to α has a simple closed-form solution:

$$(\mathbf{K} + \lambda I)\alpha = 2y, \tag{8}$$

   where $\mathbf{K} = \sum_{k=1}^{M} d_k \mathbf{K}_k$. While the form of solution is rather simple, solving this linear system is still challenging in the operator setting and we propose below an algorithm for its resolution.

2. with $\{f_k\}$ fixed, according to problem (6), we can rewrite the problem as

$$\min_{d \in \mathcal{D}} \sum_{k=1}^{M} \frac{\|f_k\|^2_{\mathcal{F}_k}}{d_k} \tag{9}$$

   which has a closed-form solution and for which optimality occurs at [19]:

$$d_k = \frac{\|f_k\|^{\frac{2}{r+1}}}{(\sum_k \|f_k\|^{\frac{2r}{r+1}})^{1/r}}. \tag{10}$$

This algorithm is similar to that of [8] and [14] both being based on alternating optimization. The difference here is that we have to solve a linear system involving a block-operator kernel matrix which is a combination of basic kernel matrices associated to M operator-valued kernels. This makes the system very challenging, and we present an algorithm for solving it in the next paragraph.

Algorithm 1 ℓ_r-norm MovKL

Input $\mathbf{K}_k$ for $k = 1, \ldots, M$

$d_k^1 \longleftarrow \frac{1}{M}$ for $k = 1, \ldots, M$

$\alpha \longleftarrow 0$

for $t = 1, 2, \ldots$ **do**

  $\alpha' \longleftarrow \alpha$

  $\mathbf{K} \longleftarrow \sum_k d_k^t \mathbf{K}_k$

  $\alpha \longleftarrow$ solution of $(\mathbf{K} + \lambda I)\alpha = 2y$

  if $\|\alpha - \alpha'\| < \epsilon$ **then**

    break

  end if

  $d_k^{t+1} \longleftarrow \frac{\|f_k\|^{\frac{2}{r+1}}}{(\sum_k \|f_k\|^{\frac{2r}{r+1}})^{1/r}}$ for $k = 1, \ldots, M$

end for

A detailed proof of convergence of the MovKL algorithm, in the case of compact operator-valued kernels, is given in the supplementary material. It proceeds by showing that the sequence of functions $\{f_k^{(n)}\}$ generated by the above alternating optimization produces a non-increasing sequence of objective values of Equation (6). Then, using continuity and boundedness arguments, we can also show that the sequence $\{f_k^{(n)}\}$ is bounded and thus converges to a minimizer of Equation (6). The proof is actually an extension of results obtained by [2] and [24] for scalar-valued kernels. However, the extension is not straighforward since infinite-dimensional Hilbert spaces with operator-valued reproducing kernels raise some technical issues that we have leveraged in the case of compact operators.

3.2 Solving a linear system with multiple block operator-valued kernels

One common way to construct operator-valued kernels is to build scalar-valued ones which are carried over to the vector-valued (resp. function-valued) setting by a positive definite matrix (resp. operator). In this setting an operator-valued kernel has the following form:

$$K(w, z) = G(w, z)T,$$

where G is a scalar-valued kernel and T is an operator in $\mathcal{L}(\mathcal{G}_y)$. In multi-task learning, T is a finite dimensional matrix that is expected to share information between tasks [11, 5]. More recently and for supervised functional output learning problems, T is chosen to be a multiplication or an integral operator [12, 13]. This choice is motivated by the fact that functional linear models for functional responses [25] are based on these operators and then such kernels provide an interesting alternative to extend these models to nonlinear contexts. In addition, some works on functional regression and structured-output learning consider operator-valued kernels constructed from the identity operator as in [18]

Algorithm 2 Gauss-Seidel Method

choose an initial vector of functions $\alpha^{(0)}$
repeat
  for $i = 1, 2, \ldots, n$
    $\alpha_i^{(t)} \longleftarrow$ sol. of (13):
      $[K(x_i, x_i) + \lambda I]\alpha_i^{(t)} = s_i$
  end for
until convergence

and [4]. In this work we adopt a functional data analysis point of view and then we are interested in a finite combination of operator-valued kernels constructed from the identity, multiplication and integral operators. A problem encountered when working with operator-valued kernels in infinite-dimensional spaces is that of solving the system of linear operator equations (8). In the following we show how to solve this problem for two cases of operator-valued kernel combinations.

Case 1: multiple scalar-valued kernels and one operator. This is the simpler case where the combination of operator-valued kernels has the following form

$$K(w, z) = \sum_{k=1}^{M} d_k G_k(w, z) T, \tag{11}$$

where G_k is a scalar-valued kernel, T is an operator in $\mathcal{L}(\mathcal{G}_y)$, and d_k are the combination coefficients. In this setting, the block operator kernel matrix $\mathbf{K}$ can be expressed as a Kronecker product between the multiple scalar-valued kernel matrix $\mathbf{G} = \sum_{k=1}^{M} d_k \mathbf{G}_k$, where $\mathbf{G}_k = [G_k(x_i, x_j)]_{i,j=1}^{n}$, and the operator T. Thus we can compute an analytic solution of the system of equations (8) by inverting $\mathbf{K} + \lambda I$ using the eigendecompositions of $\mathbf{G}$ and T as in [13].

Case 2: multiple scalar-valued kernels and multiple operators. This is the general case where multiple operator-valued kernels are combined as follows

$$K(w, z) = \sum_{k=1}^{M} d_k G_k(w, z) T_k, \tag{12}$$

where G_k is a scalar-valued kernel, T_k is an operator in $\mathcal{L}(\mathcal{G}_y)$, and d_k are the combination coefficients. Inverting the associated block operator kernel matrix $\mathbf{K}$ is not feasible in this case, that is why we propose a Gauss-Seidel iterative procedure (see Algorithm 2) to solve the system of linear operator equations (8). Starting from an initial vector of functions $\alpha^{(0)}$, the idea is to iteratively compute, until a convergence condition is satisfied, the functions α_i according to the following expression

$$[K(x_i, x_i) + \lambda I]\alpha_i^{(t)} = 2y_i - \sum_{j=1}^{i-1} K(x_i, x_j)\alpha_j^{(t)} - \sum_{j=i+1}^{n} K(x_i, x_j)\alpha_j^{(t-1)}, \tag{13}$$

where t is the iteration index. This problem is still challenging because the kernel $K(\cdot, \cdot)$ still involves a positive combination of operator-valued kernels. Our algorithm is based on the idea that instead of inverting the finite combination of

operator-valued kernels $[K(x_i, x_i)+\lambda I]$, we can consider the following variational formulation of this system

$$\min_{\alpha_i^{(t)}} \frac{1}{2} \langle \sum_{k=1}^{M+1} K_k(x_i, x_i)\alpha_i^{(t)}, \alpha_i^{(t)} \rangle_{\mathcal{G}_y} - \langle s_i, \alpha_i^{(t)} \rangle_{\mathcal{G}_y},$$

where $s_i = 2y_i - \sum_{j=1}^{i-1} K(x_i, x_j)\alpha_j^{(t)} - \sum_{j=i+1}^{n} K(x_i, x_j)\alpha_j^{(t-1)}$, $K_k = d_k G_k T_k$, $\forall k \in \{1, \ldots, M\}$, and $K_{M+1} = \lambda I$.

Now, by means of a variable-splitting approach, we are able to decouple the role of the different kernels. Indeed, the above problem is equivalent to the following one :

$$\begin{aligned} \min_{\boldsymbol{\alpha}_i^{(t)}} \frac{1}{2} \langle \hat{\mathbf{K}}(x_i, x_i)\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\alpha}_i^{(t)} \rangle_{\mathcal{G}_y^M} - \langle \mathbf{s_i}, \boldsymbol{\alpha}_i^{(t)} \rangle_{\mathcal{G}_y^M} \\ \text{with} \quad \alpha_{i,1}^{(t)} = \alpha_{i,k}^{(t)} \quad \text{for} \quad k = 2, \ldots, M+1, \end{aligned}$$

where $\hat{\mathbf{K}}(x_i, x_i)$ is the $(M+1) \times (M+1)$ diagonal matrix $[K_k(x_i, x_i)]_{k=1}^{M+1}$. $\boldsymbol{\alpha}_i^{(t)}$ is the vector $(\alpha_{i,1}^{(t)}, \ldots, \alpha_{i,M+1}^{(t)})$ and the $(M+1)$-dimensional vector $\mathbf{s_i} = (s_i, 0, \ldots, 0)$. We now have to deal with a quadratic optimization problem with equality constraints. Writing down the Lagrangian of this optimization problem and then deriving its first-order optimality conditions leads us to the following set of linear equations

$$\left\{ \begin{array}{lcl} K_1(x_i, x_i)\alpha_{i,1} - s_i + \sum_{k=1}^{M} \gamma_k & = & 0 \\ K_k(x_i, x_i)\alpha_{i,k} - \gamma_k & = & 0 \\ \alpha_{i,1} - \alpha_{i,k} & = & 0 \end{array} \right. \tag{14}$$

where $k = 2, \ldots, M+1$ and $\{\gamma_k\}$ are the Lagrange multipliers related to the M equality constraints. Finally, in this set of equations, the operator-valued kernels have been decoupled and thus, if their inversion can be easily computed (which is the case in our experiments), one can solve the problem (14) with respect to $\{\alpha_{i,k}\}$ and γ_k by means of another Gauss-Seidel algorithm after simple reorganization of the linear system.

4 Experiments

In order to highlight the benefit of our multiple operator-valued kernel learning approach, we have considered a series of experiments on a real dataset, involving functional output prediction in a brain-computer interface framework. The problem we addressed is a sub-problem related to finger movement decoding from Electrocorticographic (ECoG) signals. We focus on the problem of estimating if a finger is moving or not and also on the direct estimation of the finger movement amplitude from the ECoG signals. The development of the full BCI application is beyond the scope of this paper and our objective here is to prove that this problem of predicting finger movement can benefit from multiple kernel learning.

To this aim, the fourth dataset from the BCI Competition IV [21] was used. The subjects were 3 epileptic patients who had platinium electrode grids placed

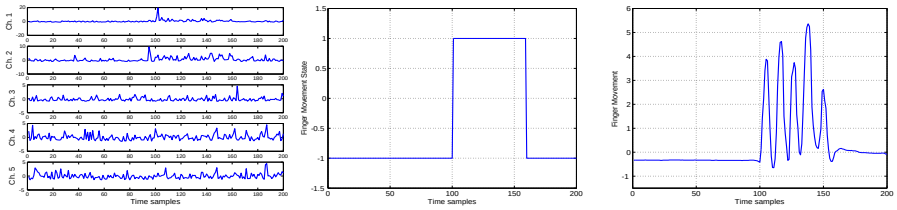


Figure 1: Example of a couple of input-output signals in our BCI task. (left) Amplitude modulation features extracted from ECoG signals over 5 pre-defined channels. (middle) Signal of labels denoting whether the finger is moving or not. (right) Real amplitude movement of the finger.

on the surface of their brains. The number of electrodes varies between 48 to 64 depending on the subject, and their position on the cortex was unknown. ECoG signals of the subject were recorded at a 1KHz sampling using BCI2000 [28]. A band-pass filter from 0.15 to 200Hz was applied to the ECoG signals. The finger flexion of the subject was recorded at 25Hz and up-sampled to 1KHz by means of a data glove which measures the finger movement amplitude. Due to the acquisition process, a delay appears between the finger movement and the measured ECoG signal [21]. One of our hopes is that this time-lag can be properly learnt by means of multiple operator-valued kernels. Features from the ECoG signals are built by computing some band-specific amplitude modulation features, which is defined as the sum of the square of the band-specific filtered ECoG signal samples during a fixed time window.

For our finger movement prediction task, we have kept 5 channels that have been manually selected and split ECoG signals in portions of 200 samples. For each of these time segments, we have the label of whether at each time sample, the finger is moving or not as well as the real movement amplitudes. The dataset is composed of 487 couples of input-output signals, the output signals being either the binary movement labels or the real amplitude movement. An example of input-output signals are depicted in Figure 1. In a nutshell, the problem boils down to be a functional regression task with functional responses.

To evaluate the performance of the multiple operator-valued kernel learning approach, we use both: **(1)** the percentage of labels correctly recognized (LCR) defined by $(W_r/T_n) \times 100\%$, where W_r is the number of well-recognized labels and T_n the total number of labels to be recognized; **(2)** the residual sum of squares error (RSSE) as evaluation criterion for curve prediction

$$RSSE = \int \sum_i \{y_i(t) - \widehat{y}_i(t)\}^2 dt, \tag{15}$$

where $\widehat{y}_i(t)$ is the prediction of the function $y_i(t)$ corresponding to real finger movement or the finger movement state.

For the multiple operator-valued kernels having the form (12), we have used a Gaussian kernel with 5 different bandwidths and a polynomial kernel of degree 1 to 3 combined with three operators T: identity $Ty(t) = y(t)$, multiplication operator associated with the function e^{-t^2} defined by $Ty(t) = e^{-t^2}y(t)$, and the integral Hilbert-Schmidt operator with the kernel $e^{-|t-s|}$ proposed in [13], $Ty(t) = \int e^{-|t-s|}y(s)ds$. The inverses of these operators can be computed ana-

Table 1: Results for the movement state prediction. Residual Sum of Squares Error (RSSE) and the percentage number of Labels Correctly Recognized (LCR) of : (1) baseline KRR with the Gaussian kernel, (2) functional response KRR with the integral operator-valued kernel, (3) MovKL with ℓ_∞, ℓ_1 and ℓ_2-norm constraint.

Algorithm	**RSSE**	**LCR(%)**
KRR - scalar-valued -	68.32	72.91
KRR - functional response -	49.40	80.20
MovKL - ℓ_∞ norm -	45.44	81.34
MovKL - ℓ_1 norm -	48.12	80.66
MovKL - ℓ_2 norm -	**39.36**	**84.72**

lytically. While the inverses of the identity and the multiplication operators are easily and directly computable from the analytic expressions of the operators, the inverse of the integral operator is computed from its spectral decomposition as in [13]. The number of eigenfunctions as well as the regularization parameter λ are fixed using "one-curve-leave-out cross-validation" [26] with the aim of minimizing the residual sum of squares error.

Empirical results on the BCI dataset are summarized in Tables 1 and 2 . The dataset was randomly partitioned into 65% training and 35% test sets. We compare our approach in the case of ℓ_1 and ℓ_2-norm constraint on the combination coefficients with: (1) the baseline scalar-valued kernel ridge regression algorithm by considering each output independently of the others, (2) functional response ridge regression using an integral operator-valued kernel [13], (3) kernel ridge regression with an evenly-weighted sum of operator-valued kernels, which we denote by ℓ_∞-norm MovKL.

As in the scalar case, using multiple operator-valued kernels leads to better results. By directly combining kernels constructed from identity, multiplication and integral operators we could reduce the residual sum of squares error and enhance the label classification accuracy. Best results are obtained using the MovKL algorithm with ℓ_2-norm constraint on the combination coefficients. RSSE and LCR of the baseline kernel ridge regression are significantly outperformed by the operator-valued kernel based functional response regression. These results confirm that by taking into account the relationship between outputs we can improve performance. This is due to the fact that an operator-valued kernel induces a similarity measure between two pairs of input/output.

5 Conclusion

In this paper we have presented a new method for learning simultaneously an operator and a finite linear combination of operator-valued kernels. We have extended the MKL framework to deal with functional response kernel ridge regression and we have proposed a block coordinate descent algorithm to solve the resulting optimization problem. The method is applied on a BCI dataset to predict finger movement in a functional regression setting. Experimental

Table 2: Residual Sum of Squares Error (RSSE) results for finger movement prediction.

Algorithm	**RSSE**
KRR - scalar-valued -	88.21
KRR - functional response -	79.86
MovKL - ℓ_∞ norm -	76.52
MovKL - ℓ_1 norm -	78.24
MovKL - ℓ_2 norm -	**75.15**

results show that our algorithm achieves good performance outperforming existing methods. It would be interesting for future work to thoroughly compare the proposed MKL method for operator estimation with previous related methods for multi-class and multi-label MKL in the contexts of structured-output learning and collaborative filtering.

Appendix

A Existence of Minimizers

We discuss in this section the existence of minimizers of problems (1) and (6). Because in both problems, we deal with infinite dimensional spaces in the optimization problem, we have to consider appropriate tools for doing so.

Existence of f_λ in the problem given in Equation (1) is guaranteed, for $\lambda > 0$ by the generalized Weierstrass Theorem and one of its corollary that we both remind below [16].

Theorem 2 *Let X be a reflexive Banach space and $C \subseteq X$ a weakly closed and bounded set. Suppose $h : C \mapsto \mathbb{R}$ is a proper lower semi-continuous function. Then h is bounded from below and has a minimizer on C.*

Corollary 3 *Let H be a Hilbert space and $h : H \mapsto \mathbb{R}$ is a strongly lower semi-continuous, convex and coercive function. Then h is bounded from below and attains a minimizer.*

We can straighforwadly apply this corollary to Problem (1) by defining

$$h(f) = \sum_{i=1}^{n} \|y_i - f(x_i)\|^2_{\mathcal{G}_y} + \lambda \|f\|^2_{\mathcal{F}}$$

with $f \in \mathcal{F}$ (which is an Hilbert space). It is easy to note that h is continuous and convex. Besides, h is coercive for $\lambda > 0$ since $\|f\|^2_{\mathcal{F}}$ is coercive and the sum involves only positive terms. Hence f_λ exists.

Regarding the MKL problem given in (6), we show existence of a solution in $\mathbf{d}$ and $\{f_k\}$ by defining the function, for fixed $\{d_k\}$

$$h_1(f_1, \cdots, f_k; \{d_k\}_k) = \sum_{i=1}^{n} \|y_i - \sum_k f_k(x_i)\|^2_{\mathcal{G}_y} + \lambda \sum_k \frac{\|f_k\|^2_{\mathcal{F}}}{d_k}$$

and by rewriting problem (6) as

$$\min_{\mathbf{d}\in\mathcal{D}} J(\mathbf{d}) \quad \text{with} \quad J(\mathbf{d}) = \min_{\{f_k\}} h_1(f_1, \cdot, , f_k; \{d_k\}_k)$$

Using similar arguments as above, it can be shown that h_1 is proper, strictly convex and coercive for fixed non-negative $\{d_k\}$ and $\lambda > 0$ (remind the convention that $\frac{x}{0} = 0$ if $x = 0$ and ∞ otherwise). Hence, minimizers of h_1 w.r.t. $\{f_1, \cdots, f_k\}$ exists and are unique. Since the function $J(\mathbf{d})$ which is equal to $h_1(f_1^\star, \cdots, f_k^\star; \{d_k\}_k)$ is continuous over the compact subset of $\mathbb{R}^M$ defined by the constraints on $\mathbf{d}$, it also attains its minimum. This conclude the proof that a solution of problem (6) exists.

B Dual Formulation of Functional Ridge Regression

Essential computational details regarding the dual formulation of functional ridge regression presented in Section 2 are discussed here.

The functional response ridge regression optimization problem has the following form:

$$\begin{aligned} &\min_{f\in\mathcal{F}} \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{2n\lambda}\sum_{i=1}^{n}\|\xi_i\|_{\mathcal{G}_y}^2 \\ &\text{with } \xi_i = y_i - f(x_i). \end{aligned} \tag{16}$$

where $(x_i, y_i)_{i=1,\ldots,n} \in (\mathcal{G}_x, \mathcal{G}_y)$. $\mathcal{G}_x$ and $\mathcal{G}_y$ are the Hilbert space $L^2(\Omega)$ which consists of all equivalence classes of square integrable functions on a finite set Ω, and $\mathcal{F}$ is a RKHS whose elements are continuous linear operators on $\mathcal{G}_x$ with values in $\mathcal{G}_y$. K is the $\mathcal{L}(\mathcal{G}_y)$-valued reproducing kernel of $\mathcal{F}$.

Since $\mathcal{G}_x$ and $\mathcal{G}_y$ are functional spaces, to derive a "dual version" of problem (16) we use the method of Lagrange multipliers on Banach spaces which is suitable to solve certain infinite-dimensional constrained optimization problems. The method is a generalization of the classical method of Lagrange multipliers. The existence of Lagrangian multipliers for the problem (16) which involves an equality constraint is guaranteed by Theorem 4.1 [1] in [15]. As consequence, the Lagrangian function associated to (16) is well defined and Fréchet-differentiable. Let $\alpha = (\alpha_i)_{i=1,\ldots,n} \in \mathcal{G}_y^n$ the vector of functions containing the Lagrange multipliers, the Lagrangian function is given by

$$L(f, \alpha, \xi) = \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{2n\lambda}\|\xi\|_{\mathcal{G}_y^n}^2 + \langle \alpha, y - f(x) - \xi\rangle_{\mathcal{G}_y^n}, \tag{17}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathcal{G}_y^n$, $y = (y_1, \ldots, y_n) \in \mathcal{G}_y^n$, $\xi = (\xi_1, \ldots, \xi_n) \in \mathcal{G}_y^n$, $f(x) = (f(x_1), \ldots, f(x_n)) \in \mathcal{G}_y^n$, and $\forall a, b \in \mathcal{G}_y^n$, $\langle a, b\rangle_{\mathcal{G}_y^n} = \sum_{i=1}^{n}\langle a_i, b_i\rangle_{\mathcal{G}_y}$.

Now, we compute $L'(f)$ the derivative of $L(f, \alpha, \xi)$ with respect to f using the Gâteaux derivative (generalization of the directional derivative) which can be defined for the direction $h \in \mathcal{F}$ by:

$$D_h L(f) = \lim_{\tau \longrightarrow 0} \frac{L(f + \tau h) - L(f)}{\tau}$$

[1] This theorem considers only equality constraint, but it is a particular case of Theorem 3.1 in [15] which deals with more general context.

Using the fact that $D_h L(f) = \langle L'(f), h\rangle_{\mathcal{F}}$, we obtain

i. $G(f) = \|f\|^2_{\mathcal{F}}$

$$\lim_{\tau \longrightarrow 0} \frac{\|f+\tau h\|^2_{\mathcal{F}} - \|f\|^2_{\mathcal{F}}}{\tau} = 2\langle f, h\rangle \implies G^{'}(f) = 2f$$

ii. $H(f) = \langle \alpha, y - f(x) - \xi\rangle_{\mathcal{G}_y^n}$

$$\lim_{\tau \longrightarrow 0} \frac{\langle \alpha, y - f(x) - \tau h(x) - \xi\rangle_{\mathcal{G}_y^n} - \langle \alpha, y - f(x) - \xi\rangle_{\mathcal{G}_y^n}}{\tau} = -\langle \alpha, h(x)\rangle_{\mathcal{G}_y^n} =$$

$$-\sum_i \langle \alpha_i, h(x_i)\rangle_{\mathcal{G}_y} = -\langle \sum_i K_{\mathcal{F}}(x_i, \cdot)\alpha_i, h\rangle_{\mathcal{F}} \quad \text{(using the reproducing property)}$$

$$\implies H^{'}(f) = -\sum_i K_{\mathcal{F}}(x_i, \cdot)\alpha_i$$

(i), (ii), and $L^{'}(f) = 0$, we obtain the (representer theorem) solution:

$$f(\cdot) = \sum_{i=1}^{n} K_{\mathcal{F}}(x_i, \cdot)\alpha_i \tag{18}$$

Substituting this into (17), the problem (16) becomes

$$\min_{\xi} \max_{\alpha} -\frac{1}{2}\langle \mathbf{K}\alpha, \alpha\rangle_{\mathcal{G}_y^n} + \frac{1}{2n\lambda}\|\xi\|^2_{\mathcal{G}_y^n} + \langle \alpha, y - \xi\rangle_{\mathcal{G}_y^n} \tag{19}$$

where $\mathbf{K} = [K(x_i, x_j)]_{i,j=1}^n$ is the block operator kernel matrix.

Differentiating (19) with respect to ξ using the same procedure as described above, we obtain $\xi = n\lambda\alpha$ and then the dual of the functional response ridge regression problem is given by

$$\max_{\alpha} -\frac{n\lambda}{2}\|\alpha\|^2_{\mathcal{G}_y^n} - \frac{1}{2}\langle \mathbf{K}\alpha, \alpha\rangle_{\mathcal{G}_y^n} + \langle \alpha, y\rangle_{\mathcal{G}_y^n} \tag{20}$$

C Convergence of Algorithm 1

In this section, we present a proof of convergence of Algorithm 1. The proof is an extension of results obtained by [2] and [24] to infinite dimensional Hilbert spaces with operator-valued reproducing kernels. Let $R(f, d)$ be the objective function of the MovKL problem defined by (6):

$$R(f, d) = L + \sum_{k=1}^{M} \frac{\|f_k\|^2_{\mathcal{F}_k}}{d_k}$$

where $L = \frac{1}{\lambda}\sum_i \|y_i - \sum_{k=1}^M f_k(x_i)\|^2_{\mathcal{G}_y}$. Substituting Equation (10) in R we obtain the objective function:

$$S(f) := R(f, d(f)) = L + \left(\sum_{k=1}^{M} \|f_k\|_{\mathcal{F}_k}^{\frac{2r}{r+1}}\right)^{\frac{r+1}{r}}$$

The function S is **strictly convex** since L is convex and the function defined by $f \longmapsto \left(\sum_{k=1}^M \|f_k\|^{\frac{2r}{r+1}}\right)^{\frac{r+1}{r}}$ is strictly convex (this follows directly from strict

convexity of the function $x \longmapsto x^p$ when $x \geq 0$ and $p > 1$). Thus, $S(f)$ admits a **unique minimizer**.

Now let us define the function g by:

$$g(f) = \min_u \{R(u, d(f))\}.$$

The function g is **continuous**. This comes from the fact that the function:

$$G(d) = \min_u \{R(u, d)\}$$

is continuous. Indeed, G is the minimal of value of a functional response kernel ridge regression problem in a function-valued RKHS associated to an operator-valued kernel K. So, $G(d) = R(d, u^*)$ with $u^* = \big(\mathbf{K}(d) + \lambda I\big)^{-1} y$ (see Equation (8)). u^* is continuous, and hence $G(d)$ is also continuous.

By definition we have $S(f) = R(f, d(f))$, and since $d(f)$ minimizes $R(f, \cdot)$, we obtain that:

$$S(f^{(n+1)}) \leq g(f^{(n)}) \leq S(f^{(n)})$$

where n is the number of iteration. So, the sequence $\{S(f^{(n)}),\ n \in \mathbb{N}\}$ is nonincreasing and then it is bounded since L is bounded from below. Thus, as $n \longrightarrow \infty$, $S(f^{(n)})$ converges to a number which we denote by S^*. $\{S(f^{(n)})\}$ is convergent and S is a coercive function, then the sequence $\{\|f^{(n)}\|,\ n \in \mathbb{N}\}$ is bounded. Consequently, the sequence $\{f^{(n)},\ n \in \mathbb{N}\}$ is **bounded**.

Next we show the following subsequence convergence property which underlies the convergence of Algorithm 1.

Proposition 4 *If $\mathcal{F}$ is a RKHS associated to a compact-operator-valued kernel, the sequence $\{f^{(n)} \in \mathcal{F},\ n \in \mathbb{N}\}$, since it is bounded, has a convergent subsequence.*

Proof. The analogue of the Bolzano-Weierstrass theorem[2] in Hilbert spaces states that there exists a **weakly convergent** subsequence $\{f^{(n_l)},\ l \in \mathbb{N}\}$ of the bounded sequence $\{f^{(n)}\}$ which (weakly) converges to $f \in \mathcal{F}$. By definition of weakly convergence, we have $\forall g \in \mathcal{F}$ ($\mathcal{F}$ is a RKHS with the operator-valued kernel K):

$$\lim_{n_l \to \infty} \langle f^{(n_l)}(\cdot), g\rangle_{\mathcal{F}} = \langle f(\cdot), g\rangle_{\mathcal{F}} \tag{21}$$

Let $g = K(x, \cdot)\beta$. Using the reproducing property, we obtain

$$\lim_{n_l \to \infty} \langle f^{(n_l)}(\cdot), g\rangle_{\mathcal{F}} = \lim_{n_l \to \infty} \langle f^{(n_l)}(x), \beta\rangle_{\mathcal{G}_y} \quad \text{and} \quad \langle f(\cdot), g\rangle_{\mathcal{F}} = \langle f(x), \beta\rangle_{\mathcal{G}_y}$$
$$\Rightarrow \lim_{n_l \to \infty} \langle f^{(n_l)}(x), \beta\rangle_{\mathcal{G}_y} = \langle f(x), \beta\rangle_{\mathcal{G}_y} \quad \text{using (21)}$$

Thus the subsequence $\{f^{(n_l)}(\cdot)\}$ is (weakly) **pointwise convergent**.

Now we show that:

$$\lim_{n_l \to \infty} \|f^{(n_l)}(\cdot)\|_{\mathcal{F}} = \|f(\cdot)\|_{\mathcal{F}} \tag{$*$}$$

[2]The Bolzano-Weierstrass theorem states that each bounded sequence in $\mathbb{R}^n$ has a convergent subsequence. For infinite-dimensional spaces, strong convergence of the subsequence is not reached and only weak convergence is obtained. Proposition 4 shows that strong convergence can be reached for the sequence $\{f^{(n)},\ n \in \mathbb{N}\}$ solution of our MovKL optimization problem in Hilbert spaces with reproducing compact operator-valued kernels.

Since $f^{(n_l)} \in \mathcal{F}$ is solution of the minimization of the optimization problem (6) with the kernel combination coefficients d_k fixed, it can be written as $\sum_i K(x_i, \cdot)\alpha_i^{(n_l)}$ (representer theorem). We now that $f^{(n_l)}(x)$ converges weakly to $f(x)$, so $\alpha^{(n_l)} = \left(\alpha_i^{(n_l)}\right)_{i\geq 1}$ converges weakly to $\alpha \in \mathcal{G}_y^n$. Indeed, $\forall \beta \in \mathcal{G}_y$ we have:

$$\begin{aligned}
&\lim_{n_l\to\infty} \langle f^{(n_l)}(x), \beta\rangle_{\mathcal{G}_y} = \langle f(x), \beta\rangle_{\mathcal{G}_y} \\
&\Rightarrow \lim_{n_l\to\infty} \langle \sum_i K(x_i, x)\alpha_i^{(n_l)}, \beta\rangle_{\mathcal{G}_y} = \langle \sum_i K(x_i, x)\alpha_i, \beta\rangle_{\mathcal{G}_y} \\
&\Rightarrow \lim_{n_l\to\infty} \langle \mathbf{K}_{\mathbf{x}}\alpha^{(n_l)}, \beta\rangle_{\mathcal{G}_y} = \langle \mathbf{K}_{\mathbf{x}}\alpha, \beta\rangle_{\mathcal{G}_y} \quad \text{where } \mathbf{K}_{\mathbf{x}} \text{ is the row vector } (K(x_i, x))_{i\geq 1} \\
&\Rightarrow \lim_{n_l\to\infty} \langle \alpha^{(n_l)}, \mathbf{K}_{\mathbf{x}}^*\beta\rangle_{\mathcal{G}_y^n} = \langle \alpha, \mathbf{K}_{\mathbf{x}}^*\beta\rangle_{\mathcal{G}_y^n} \\
&\Rightarrow \lim_{n_l\to\infty} \langle \alpha^{(n_l)}, z\rangle_{\mathcal{G}_y^n} = \langle \alpha, z\rangle_{\mathcal{G}_y^n} \quad \forall z \in \mathcal{G}_y^n \\
&\Rightarrow \alpha^{(n_l)} \text{ converges weakly to } \alpha
\end{aligned}$$

Moreover

$$\begin{aligned}
\lim_{n_l\to\infty} \|f^{(n_l)}(\cdot)\|_{\mathcal{F}}^2 &= \lim_{n_l\to\infty} \langle \sum_i K(x_i, \cdot)\alpha_i^{(n_l)}, \sum_j K(x_j, \cdot)\alpha_j^{(n_l)}\rangle_{\mathcal{F}} \\
&= \lim_{n_l\to\infty} \sum_{i,j} \langle K(x_i, x_j)\alpha_i^{(n_l)}, \alpha_j^{(n_l)}\rangle_{\mathcal{G}_y} \text{ (using the reproducing property)} \\
&= \lim_{n_l\to\infty} \langle \mathbf{K}\alpha^{(n_l)}, \alpha^{(n_l)}\rangle_{\mathcal{G}_y^n} \\
&= \langle \mathbf{K}\alpha, \alpha\rangle_{\mathcal{G}_y^n} \quad \text{(because of the compactness}^3 \text{ of } \mathbf{K}) \\
&= \sum_{i,j} \langle K(x_i, x_j)\alpha_i, \alpha_j\rangle_{\mathcal{G}_y} \\
&= \sum_{i,j} \langle K(x_i, \cdot)\alpha_i, K(x_j, \cdot)\alpha_j\rangle_{\mathcal{F}} = \|f\|_{\mathcal{F}}^2
\end{aligned}$$

Using $(*)$ and weak convergence, we obtain the **strong convergence** of the subsequence $\{f^{(n_l)}\}$.

$$\begin{aligned}
\lim_{n_l\to\infty} \|f^{(n_l)} - f\|_{\mathcal{F}}^2 &= \lim_{n_l\to\infty} \langle f^{(n_l)} - f, f^{(n_l)}\rangle_{\mathcal{F}} - \langle f^{(n_l)} - f, f\rangle_{\mathcal{F}} \\
&= \lim_{n_l\to\infty} \|f^{(n_l)}\|_{\mathcal{F}}^2 - \lim_{n_l\to\infty} 2\langle f, f^{(n_l)}\rangle_{\mathcal{F}} + \|f\|_{\mathcal{F}}^2 \\
&= \lim_{n_l\to\infty} \|f^{(n_l)}\|_{\mathcal{F}}^2 - \|f\|_{\mathcal{F}}^2 \text{ (using weak convergence)} \\
&= 0 \quad \text{(using}(*)\text{)} \\
&\Rightarrow f^{(n_l)} \text{ converges strongly to } f \qquad \square
\end{aligned}$$

By Proposition 4, there exists a convergent subsequence $\{f^{(n_l)},\ l \in \mathbb{N}\}$ of the bounded sequence $\{f^{(n)},\ n \in \mathbb{N}\}$, whose limit we denote by f^*. Since $S(f^{(n+1)}) \leq g(f^{(n)}) \leq S(f^{(n)})$, $g(f^{(n)})$ converges to S^*. Thus, by the continuity

[3] Compact operator maps weakly convergent sequences into strongly convergent sequences.

of g and S, $g(f^*) = S(f^*)$. This implies that f^* is a minimizer of $R(\cdot, d(f^*))$, because $R(f^*, d(f^*)) = S(f^*)$. Moreover, $d(f^*)$ is the minimizer of $R(\cdot, f^*)$ subject to the constraints on d. Thus, since the objective function R is smooth, the pair $(f^*, d(f^*))$ is the minimizer of R.

At this stage, we have shown that any convergent subsequence of $\{f^{(n)},\ n \in \mathbb{N}\}$ converges to the minimizer of R. Since the sequence $\{f^{(n)},\ n \in \mathbb{N}\}$ is bounded, it follows that the whole sequence **converges** to minimizer of R.

References

[1] J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. Saketha Nath, and S. Raman. Variable sparsity kernel learning. *JMLR*, 12:565–592, 2011.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] F. Bach. Consistency of the group Lasso and multiple kernel learning. *JMLR*, 9:1179–1225, 2008.

[4] C. Brouard, F. d'Alché-Buc, and M. Szafranski. Semi-supervised penalized output kernel regression for link prediction. In *Proc. ICML*, 2011.

[5] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *JMLR*, 68:1615–1646, 2008.

[6] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

[7] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

[8] C. Cortes, M. Mohri, and A. Rostamizadeh. L_2 regularization for learning kernels. In *Proc. UAI*, 2009.

[9] C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proc. ICML*, 2010.

[10] F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. In *Proc. ICML*, 2011.

[11] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *JMLR*, 6:615–637, 2005.

[12] H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional RKHS approach. In *Proc. AISTATS*, pages 111–125, 2010.

[13] H. Kadri, A. Rabaoui, P. Preux, E. Duflos, and A. Rakotomamonjy. Functional regularized least squares classification with operator-valued kernels. In *Proc. ICML*, 2011.

[14] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning. *JMLR*, 12:953–997, 2011.

[15] S. Kurcyusz. On the existence and nonexistence of lagrange multipliers in Banach spaces. *Journal of Optimization Theory and Applications*, 20:81–110, 1976.

[16] A. Kurdila and M. Zabarankin. *Convex Functional Analysis*. Birkhauser Verlag, 2005.

[17] G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *JMLR*, 5:27–72, 2004.

[18] H. Lian. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *The Canadian Journal of Statistics*, 35:597–606, 2007.

[19] C. Micchelli and M. Pontil. Learning the kernel function via regularization. *JMLR*, 6:1099–1125, 2005.

[20] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

[21] K. J. Miller and G. Schalk. Prediction of finger flexion: 4th brain-computer interface data competition. BCI Competition IV, 2008.

[22] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Prediction of arm movement trajectories from ecog-recordings in humans. *Journal of Neuroscience Methods*, 167(1):105–114, 2008.

[23] A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

[24] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu. l(p)-l(q) penalty for sparse linear and sparse multiple kernel multitask learning. *IEEE Trans. Neural Netw.*, 22(8):1307–20, 2011.

[25] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, 2nd ed.* Springer Verlag, New York, 2005.

[26] John A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B*, 53(1):233–243, 1991.

[27] G. Schalk, J. Kubanek, K. J. Miller, N. R. Anderson, E. C. Leuthardt, J. G. Ojemann, D. Limbrick, D. Moran, L. A. Gerhardt, and J. R. Wolpaw. Decoding two-dimensional movement trajectories using electrocorticographic signals in humans. *Journal of Neural Engineering*, 4(3):264–275, 2007.

[28] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. BCI2000: a general-purpose brain-computer interface system. *Biomedical Engineering, IEEE Trans. on*, 51:1034–1043, 2004.

[29] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

[30] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[31] P. Tseng. Convergence of block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109:475–494, 2001.

Centre de recherche INRIA Lille – Nord Europe
Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399