Research Note RN/05/11

Department of Computer Science, University College London

Feature space perspectives for learning the kernel[1]

Charles A. Micchelli

Department of Mathematics and Statistics
State University of New York
The University at Albany
1400 Washington Avenue, Albany, NY, 12222, USA
E-mail: *cam@math.albany.edu*

and

Massimiliano Pontil

Department of Computer Science
University College London
Gower Street, London WC1E, England, UK
E-mail: *m.pontil@cs.ucl.ac.uk*

18 June, 2005

Abstract

In this paper, we continue our study of learning the kernel. We present a reformulation of this problem within a feature space environment. This leads us to study regularization in the dual space of all continuous functions on a compact domain with values in a Hilbert space with a mix norm. We also relate this problem in a special case to L^p regularization.

[1]This work was supported by NSF Grant ITR-0312113 and EPSRC Grant GR/T18707/01.

1 Introduction

A central theme of this paper is the problem of learning a kernel in a prescribed convex set of kernels $\mathcal{K}$. Our previous work on this problem which was motivated by its potential application in machine learning focused on finding a suitable optimal kernel. Here, we study an equivalent feature space formulation of this problem. This leads us to explore the relationship between the problem of finding an optimal kernel and regularization in the dual space of the space of continuous functions on a compact domain with values in a Hilbert space. We also describe related regularization techniques in L^p spaces which naturally arise in our investigation.

In [1, 17] we proposed to find a good kernel K by solving the variational problem

$$\min\left\{\sum_{j\in\mathbb{N}_m} A(y_j, f(x_j)) + \mu\|f\|_K^2 \;:\; f\in\mathcal{H}_K,\; K\in\mathcal{K}\right\} \tag{1.1}$$

where $A : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ is a prescribed loss function, $\|\cdot\|_K$ the norm in a reproducing kernel Hilbert space with kernel K, μ a positive parameter and $\mathcal{K}$ is a prescribed set of kernels. This problem has been studies from different perspectives in a substantial number of papers. Specifically, in statistics, it has been motivated in [12, 13] as a generalization of the *lasso* introduced in [21], a technique which also relates to basis pursuit denoising [6] and to a linear programming approach for feature section [8]; in machine learning, problem (1.1) has been studied in the context of support vector machines as a mean to optimize the margin or soft-margin error used therein [3, 11]; in approximation theory, it has been investigated with the intention of improving the approximation error [23, 18]. For additional interesting observations related to the theme of this paper, see [5, 7, 9, 10, 19].

In Section 2, we describe the main result of the paper which relates problem (1.1) to our feature space extremal problem. Indeed, the problem described above concerns the choice of an optimal kernel for kernel based learning algorithms while the second problem we study is the reformulation of it within a feature space environment. We demonstrate in great generality that these problems are equivalent and characterize the form of the solutions for both problems. We also provide a description for an optimal feature map solution analogous to the one we derived in our earlier work on learning the kernel, [1, 17]. A detailed description of this result appear in Section 2. However, the proof is postponed until Section 6. In Section 3, we present specific motivating examples when $\mathcal{K}$ is the convex hull of a *finite* set of prescribed kernels. Moreover, for these examples we provide an alternate derivation of the main result in Section 2 by using a decomposition theorem from [2]. In Section 4, we discuss the connection between learning the kernel and L^1 regularization. Section 5 contains related results for L^p regularization and provide a representer theorem in the spirit of our paper, [16]. We end the paper with a discussion of future research directions and commentaries on our results presented here.

We remark that an interesting aspect of the feature space regularization we present here is not only does it involve linear functionals, but also that it is a Banach space regularization method. Indeed, as we shall show, the appropriate norm for the functionals is induced by a mix norm on a space of functions with values in the Hilbert space associated with the feature map. Finally, we also explore similar issues for an L^p analog of the convex hull of a fix set of kernels.

2 Main result

In this section we present our main result. First, we recall the notion of reproducing kernel Hilbert spaces and continuously parameterized convex set of kernels.

2.1 Integrals of kernels

Let $\mathcal{X}$ be an *input set*. By a *kernel* we mean a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for every finite set of inputs $\mathbf{x} = \{x_j : j \in \mathbb{N}_m\} \subseteq \mathcal{X}$ and every $m \in \mathbb{N}$, the $m \times m$ matrix $K_{\mathbf{x}} := (K(x_i, x_j) : i, j \in \mathbb{N}_m)$ is *positive semi-definite*. According to Aronszajn and Moore, every kernel has associated to it an (essentially) *unique* Hilbert space $\mathcal{H}_K$ with inner product $\langle \cdot, \cdot \rangle_K$ such that K is its reproducing kernel, [2]. This means, for every $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$, that $\langle f, K(x, \cdot) \rangle_K = f(x)$.

We let $\mathcal{L}(\mathbb{R}^m)$ be the set of $m \times m$ positive semi-definite matrices and $\mathcal{L}_+(\mathbb{R}^m)$ the subset of *positive definite* ones. We use the notation $\mathcal{A}(\mathcal{X})$ for the set of all kernels on the set $\mathcal{X}$ and $\mathcal{A}_+(\mathcal{X})$ for the subset of kernels K such that, for *each* input $\mathbf{x}$, $K_{\mathbf{x}} \in \mathcal{L}_+(\mathbb{R}^m)$.

Let Θ be a compact Hausdorff space, $C(\Theta)$ the space of continuous real–valued functions on Θ and $\mathcal{M}(\Theta)$ the set of all *probability measures* on Θ. Let $G : \Theta \to \mathcal{A}_+(\mathcal{X})$ be a continuous map. By this we mean that, for each $x, t \in \mathcal{X}$, the function of $\theta \mapsto G(\theta)(x, t)$ is continuous on Θ. The set of kernels $\mathcal{G} := \{G(\theta) : \theta \in \Theta)\}$ induces the convex set of kernels

$$\mathcal{K}(\mathcal{G}) := \left\{ \int_\Theta G(\theta) dp(\theta) : p \in \mathcal{M}(\Theta) \right\} \tag{2.1}$$

which we shall consider below.

2.2 Regularization error functional

Let $D := \{(x_j, y_j) : j \in \mathbb{N}_m\} \subset \mathcal{X} \times \mathbb{R}$ be prescribed data and y the vector $(y_j : j \in \mathbb{N}_m)$. Each kernel $K \in \mathcal{K}(\mathcal{G})$ gives rise to a RKHS $\mathcal{H}_K$. For each $f \in \mathcal{H}_K$, we introduce the *information operator* $I_{\mathbf{x}}(f) := (f(x_j) : j \in \mathbb{N}_m) \in \mathbb{R}^m$ of values of f on the set of inputs in $\mathbf{x}$. We consider $I_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ a linear map and on $\mathbb{R}^m$ we put the usual inner product. Thus, for any two vectors $c = (c_j : j \in \mathbb{N}_m)$ and $d = (d_j : j \in \mathbb{N}_m)$ we write $(c, d) := \sum_{j \in \mathbb{N}_m} c_j d_j$. A straightforward computation identifies the adjoint $I_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}_K$, for every $c = (c_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$, as

$$I_{\mathbf{x}}^* c = \sum_{j \in \mathbb{N}_m} c_j K_j.$$

A *regularization error function* is any function $q : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$. We write any vector $v \in \mathbb{R}^m \times \mathbb{R}_+$ in the form (c, t) for some $c \in \mathbb{R}^m$ and $t \in \mathbb{R}_+$. In other words, the vector v is the concatenation of the vector c and the scalar t. There should be no confusion with this "double duty" notation since in this case one argument is a vector and the other is a scalar. Likewise, we shall denote $q(v)$ as $q(c, t)$ and $q_{\inf} := \inf\{q(v) : v \in \mathbb{R}^m \times \mathbb{R}_+\}$. The regularization error function will allow us to balance the data $I_{\mathbf{x}} f$ with the norm $\|f\|_K := \sqrt{\langle f, f \rangle_K}$ and leads us to the following definition.

Definition 2.1. *We say that $q : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$ is an acceptable regularization error function provided that*

1. *q is lower semi-continuous, that is, for each $\lambda \in \mathbb{R}$ the set $\mathcal{U}_\lambda := \{v : v \in \mathbb{R}^m \times \mathbb{R}_+,\ q(v) \leq \lambda\}$ is a closed subset of $\mathbb{R}^m$;*

2. *given any $\lambda > 0$ there exists $\rho > 0$ such that whenever $(c, t) \in \mathcal{U}_\lambda$ then $t \in [0, \rho]$;*

3. *for each $c \in \mathbb{R}^m$ the function $q(c, \cdot) : \mathbb{R}_+ \to \mathbb{R}$ is non-decreasing on $\mathbb{R}_+$.*

An important example for an acceptable regularization error function in machine learning has the form

$$q(c, t) = \sum_{j \in \mathbb{N}_m} A(y_j, c_j) + \mu B(t) \tag{2.2}$$

where $A : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a continuous loss function (typically convex), μ a positive constant and $B : \mathbb{R}_+ \to \mathbb{R}_+$ is a *strictly increasing* function. The standard choice for the function B is $B(t) = t$, $t \in \mathbb{R}_+$. This leads to the usual kernel–based regularization algorithm studied extensively in the literature. However, as we shall see later, from the feature space point of view, the choice $B(t) = \sqrt{t}$, $t \in \mathbb{R}_+$, is widely studied. Whenever we consider the special case (2.2) we always assume that the above properties of A and B are satisfied.

An acceptable regularization error function gives rise to a functional $Q(\cdot, K) : \mathcal{H}_K \to \mathbb{R}$ defined, for all $f \in \mathcal{H}_K$, as

$$Q(f, K) = q(I_{\mathbf{x}} f, \|f\|_K^2).$$

Properties 1–3 above guarantee, for every $K \in \mathcal{A}(\mathcal{X})$, that $Q(\cdot, K)$ has a minimum f_K over $f \in \mathcal{H}_K$. Since we do not assume here that q is strictly convex this minimum may not be unique, hence our notation f_K is to be interpreted to mean that f_K is *any* minimum for $Q(\cdot, K)$. We let

$$E(K) := \min\{Q(f, K) : f \in \mathcal{H}_K\} = Q(f_K, K).$$

In our previous work [1] we studied the problem of choosing a kernel $K \in \mathcal{K}(\mathcal{G})$ which solves the variational problem

$$E_G := \inf\{E(K) : K \in \mathcal{K}(\mathcal{G})\}. \tag{2.3}$$

Any kernel $\hat{K} \in \mathcal{K}(\mathcal{G})$ for which $E_G = E(\hat{K})$ is called an optimal kernel and $f_{\hat{K}}$ is called an optimal function. A main goal of this paper is to give the variational problem (2.3) a feature space interpretation. To this end, we express $G : \Theta \to \mathcal{A}_+(\mathcal{X})$ in terms of a *feature map* $\Phi : \Theta \times \mathcal{X} \to \mathcal{W}$, where $\mathcal{W}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$ and corresponding norm $\| \cdot \|_{\mathcal{W}}$. That is, for each $x, y \in \mathcal{X}$ and $\theta \in \Theta$, we have that $G(\theta)(x, y) = \langle \Phi(\theta, x), \Phi(\theta, y) \rangle_{\mathcal{W}}$. Note that since G is continuous it follows, for every $x \in \mathcal{X}$, that the function $\Phi(\cdot, x)$ is in $C(\Theta, \mathcal{W})$, the set of *all* continuous functions on Θ with values in $\mathcal{W}$, where we give any $\Psi \in C(\Theta, \mathcal{W})$ the norm

$$\|\Psi\|_{\infty, \mathcal{W}} := \max\{\|\Psi(\theta)\|_{\mathcal{W}} : \theta \in \Theta\}.$$

We introduce *feature functions* $\Phi_j : \Theta \to \mathcal{W}$ induced by the inputs $\mathbf{x}$ and defined, for every $\theta \in \Theta$, as $\Phi_j(\theta) = \Phi(\theta, x_j)$ and assume that they are linearly independent on Θ. This is equivalent to the assumption that G maps into $\mathcal{A}_+(\mathcal{X})$. Corresponding to any feature map and any regularization error function as described above we introduce a feature space regularization functional $V : C^*(\Theta, \mathcal{W}) \to \mathbb{R}$ given, for any $L \in C^*(\Theta, \mathcal{W})$, by

$$V(L) = q(D_{\mathbf{x}}(L), \|L\|^2) \tag{2.4}$$

where $D_{\mathbf{x}} : C^*(\Theta, \mathcal{W}) \to \mathbb{R}^m$ is the linear operator defined as $D_{\mathbf{x}}(L) = (L(\Phi_j) : j \in \mathbb{N}_m)$. Recall that the norm of the linear functional L is defined as

$$\|L\| := \sup \{L(\Psi) : \|\Psi\|_{\infty,\mathcal{W}} \leq 1\}.$$

We introduce the variational problem

$$V_\Phi := \inf\{V(L) : L \in C^*(\Theta, \mathcal{W})\}$$

which we will henceforth refer to as the feature space variational problem. Any linear functional $\hat{L} \in C^*(\Theta, \mathcal{W})$ for which $V_\Phi = V(\hat{L})$ is called an optimal linear functional. Our main result is the following fact.

Theorem 2.1. *Under the above hypotheses, we have that $E_G = V_\Phi$. Moreover, there exist a finitely supported measure $\hat{p} \in \mathcal{M}(\Theta)$ with at most $m+1$ atoms and a vector $\hat{c} \in \mathbb{R}^m$ such that the kernel*

$$\hat{K} = \int_\Theta G(\theta) d\hat{p}(\theta)$$

is an optimal kernel, the function

$$f_{\hat{K}} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \hat{K}_j$$

is an optimal function and the linear functional $\hat{L} \in C^*(\Theta, \mathcal{W})$ defined, for any $\Psi \in C(\Theta, \mathcal{W})$, as

$$\hat{L}(\Psi) = \int_\Theta \langle \hat{\Gamma}(\theta), \Psi(\theta) \rangle_{\mathcal{W}} d\hat{p}(\theta)$$

is an optimal linear functional, where the function $\hat{\Gamma} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \Phi_j$ has the property, for every atom θ of $\hat{p}$, that $\|\hat{\Gamma}(\theta)\|_{\mathcal{W}} = \|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \|\hat{L}\|$.

We wish point out another connection between the optimal function, optimal linear functional and the feature map, namely, for every $x \in \mathcal{X}$, we have that

$$f_{\hat{K}}(x) = \hat{L}(\Phi(\cdot, x)).$$

We mention here an immediate corollary of the above theorem.

Corollary 2.1. *Under the hypothesis above, we have that $E_G = V_\Phi$ where*

$$E_G = \inf \left\{ \sum_{j \in \mathbb{N}_m} A(y_j, f(x_j)) + \mu B(\|f\|_K^2) : f \in \mathcal{H}_K, K \in \mathcal{K}(\mathcal{G}) \right\}$$

and

$$V_\Phi = \min \left\{ \sum_{j \in \mathbb{N}_m} A\left(y_j, L(\Phi_j)\right) + \mu B(\|L\|^2) : L \in C^*(\Theta, \mathcal{W}) \right\}.$$

Moreover, there exist a finitely supported measure $\hat{p} \in \mathcal{M}(\Theta)$ with at most $m+1$ atoms and a vector $\hat{c} \in \mathbb{R}^m$ such that the kernel

$$\hat{K} = \int_\Theta G(\theta) d\hat{p}(\theta)$$

is an optimal kernel, the function

$$f_{\hat{K}} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \hat{K}_j$$

is an optimal function and the linear functional $\hat{L} \in C^*(\Theta, \mathcal{W})$ *defined, for any* $\Psi \in C(\Theta, \mathcal{W})$*, as*

$$\hat{L}(\Psi) = \int_\Theta \langle \hat{\Gamma}(\theta), \Psi(\theta) \rangle_{\mathcal{W}} d\hat{p}(\theta)$$

is an optimal linear functional, where the function $\hat{\Gamma} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \Phi_j$ *has the property, for every atom* θ *of* $\hat{p}$*, that* $\|\hat{\Gamma}(\theta)\|_{\mathcal{W}} = \|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \|\hat{L}\|$.

Before we prove Theorem 2.1 we describe several examples of some potential practical importance in the next sections. We postpone the proof of Theorem 2.1 to Section 6.

3 A practical example: finitely many kernels

In this section we specialize our analysis to the practically important case that $\Theta = \mathbb{N}_n$. Thus, we write $\mathcal{G} = \{G_r : r \in \mathbb{N}_n\}$, $\mathcal{M}(\Theta) = \mathbb{S}^n$, the $n-$dimensional simplex given by $\mathbb{S}^n = \{(\lambda_r : r \in \mathbb{N}_n) \in \mathbb{R}^n : \lambda_r \geq 0,\ \sum_{s \in \mathbb{N}_n} \lambda_s = 1\}$, and

$$\mathcal{K}(\mathcal{G}) = \left\{ \sum_{r \in \mathbb{N}_n} \lambda_r G_r \ : \ (\lambda_r : r \in \mathbb{N}_n) \in \mathbb{S}^n \right\}.$$

We let $\mathcal{W}^n$ be the $n-$fold cross product of $\mathcal{W}$, that is,

$$\mathcal{W}^n = \left\{ w = (w_r : r \in \mathbb{N}_n) : w_r \in \mathcal{W},\ r \in \mathbb{N}_n \right\}$$

equipped it with the ℓ_n^p norm, where $1 \leq p \leq \infty$, given, for any $w \in \mathcal{W}^n$, as

$$\|w\|_{p,\mathcal{W}} := \left(\sum_{r \in \mathbb{N}_n} \|w_r\|_{\mathcal{W}}^p \right)^{\frac{1}{p}}$$

and denote the resulting Banach space by $\mathcal{W}^{p,n}$. In the case that $p = 2$, this is a Hilbert space with inner product defined, for every $u, w \in \mathcal{W}^n$ as $\langle w, u \rangle_{\mathcal{W}^n} := \sum_{r \in \mathbb{N}_n} \langle w_r, u_r \rangle_{\mathcal{W}}$. Clearly, the space $C(\Theta, \mathcal{W})$ is identified with $\mathcal{W}^{\infty,n}$. and its dual space, $C^*(\Theta, \mathcal{W})$, with $\mathcal{W}^{1,n}$ Thus, a linear functional $L \in C^*(\Theta, \mathcal{W})$ corresponds uniquely to a vector $w \in \mathcal{W}^n$ by means of the equation

$$L(u) = \langle w, u \rangle_{\mathcal{W}^n}, \quad u \in \mathcal{W}^n$$

and its norm is given by

$$\|L\| = \|w\|_{1,\mathcal{W}}.$$

We shall now specialize Corollary 2.1 to this case.

Corollary 3.1. *Under the hypotheses above, we have that*

$$E_G = V_\Phi \tag{3.1}$$

where

$$E_G = \inf\left\{\sum_{j\in\mathbb{N}_m} A\left(y_j, f(x_j)\right) + \mu B(\|f\|_K^2) : f \in \mathcal{H}_K, K \in \mathcal{K}(\mathcal{G})\right\}$$

and

$$V_\Phi = \min\left\{\sum_{j\in\mathbb{N}_m} A\left(y_j, \langle w, \Phi(x_j)\rangle_{\mathcal{W}^n}\right) + \mu B(\|w\|_{1,\mathcal{W}}^2) : w \in \mathcal{W}^n\right\}. \tag{3.2}$$

Moreover, there exist $\hat{\lambda} \in \mathbb{S}^n$ *and a vector* $\hat{c} \in \mathbb{R}^m$ *such that the set* $J := \{r : \hat{\lambda}_r > 0\}$ *has cardinality at most* $\min(m+1, n)$*, the kernel*

$$\hat{K} = \sum_{r\in J} \hat{\lambda}_r G_r \tag{3.3}$$

is an optimal kernel, the function

$$f_{\hat{K}} = \sum_{j\in\mathbb{N}_m} \hat{c}_j \hat{K}_j$$

is an optimal function and the vector

$$\hat{w} = \left(\hat{\lambda}_r \hat{\Gamma}_r : r \in \mathbb{N}_n\right)$$

is an optimal solution to problem (3.2), where the $\hat{\Gamma} = \sum_{j\in\mathbb{N}_m} \hat{c}_j \Phi(x_j) \in \mathcal{W}^n$ *has the property, for every* $r \in J$*, that* $\|\hat{\Gamma}_r\|_{\mathcal{W}} = \|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \|\hat{w}\|_{1,\mathcal{W}}$.

Note that the last statement in the corollary concerning the optimal vector $\hat{w} = (\hat{w}_r : r \in \mathbb{N}_n)$ says, for every $r \in \mathbb{N}_n$, that

$$\|\hat{w}_r\|_{\mathcal{W}} = \hat{\lambda}_r \|\hat{\Gamma}\|_{\infty,\mathcal{W}}.$$

Thus, summing both sides of this equation over $r \in \mathbb{N}_n$, we conclude that

$$\|\hat{w}\|_{1,\mathcal{W}} = \|\hat{\Gamma}\|_{\infty,\mathcal{W}}$$

and, so

$$\hat{\lambda}_r = \frac{\|\hat{w}_r\|_{\mathcal{W}}}{\|\hat{w}\|_{1,\mathcal{W}}}. \tag{3.4}$$

This formula demonstrate that a solution to the feature space variational problem provides a choice for the optimal kernel in equation (3.3).

When $\mathcal{W} = \mathbb{R}$, problem (3.2) becomes

$$\min\left\{\sum_{j\in\mathbb{N}_m} A\left(y_j, \langle w, \Phi(x_j)\rangle_{\mathbb{R}^n}\right) + \mu B\left(\left(\sum_{r\in\mathbb{N}_n} |w_r|\right)^2\right) \ : \ w \in \mathbb{R}^n\right\}. \tag{3.5}$$

This problem is closely related to some well-known function estimation techniques. In particular, when the loss function A is the square loss and B is chosen to be $B(t) = \sqrt{t},\ t \in \mathbb{R}_+$, the

variational problem (3.5) has been studied in statistics under the name of *lasso* [21], in signal processing as *basis pursuit denoising* [6], and in linear programming [8] as a *feature selection* method. The common theme of these methods is that the solution $\hat{w}$ of (3.5) is sparse, that is, most of its components are zero. The nonzero components of $\hat{w}$ identify informative features for representing the given data. Indeed, Corollary 3.1 establishes that there exists an optimal vector $\hat{w}$ with at most $\min(n, m+1)$ non zero components.

We note that the equivalence between the functionals E_G and V_Φ described in Corollary 3.1 when the loss function A is the hinge loss used in support vector machines is also described in [3]. Moreover, a method similar to problem (3.2) has been recently proposed in statistics under the name of *cosso*, where it has been proposed as a generalization of the *lasso*, see [13] and also [12] for related results. A specific instance of the above environment is provided by ANOVA kernels, see [22, 13] for a detailed discussion.

We now present an alternate derivation of equation (3.1). To this end, we require the following result of Aronszajn concerning the norm induced by a sum of reproducing kernels, [2, §7].

Theorem 3.1. *If $\{G_r, r \in \mathbb{N}_n\} \subseteq \mathcal{A}(\mathcal{X})$ and $K = \sum_{r\in\mathbb{N}_n} G_r$ then, for every $f \in \mathcal{H}_K$, we have that*

$$\|f\|_K^2 = \min\left\{ \sum_{r\in\mathbb{N}_n} \|f_r\|_{G_r}^2 \ : \ f = \sum_{r\in\mathbb{N}_n} f_r,\ f_s \in \mathcal{H}_{G_s},\ s \in \mathbb{N}_n \right\}.$$

Without elaborating on the technical details, we note, by Theorem 3.1 that

$$\begin{aligned}
E_G &= \min\left\{ \sum_{j\in\mathbb{N}_m} A\left(y_j, \sum_{r\in\mathbb{N}_n} f_r(x_j)\right) + \mu B\left(\sum_{r\in\mathbb{N}_n} \frac{\|f_r\|_{G_r}^2}{\lambda_r}\right) : f_r \in \mathcal{H}_{G_r}, \lambda \in \mathbb{S}^n \right\} \\
&= \min\left\{ \sum_{j\in\mathbb{N}_m} A\left(y_j, \sum_{r\in\mathbb{N}_n} f_r(x_j)\right) + \mu B\left(\Big(\sum_{r\in\mathbb{N}_n} \|f_r\|_{G_r}\Big)^2\right) : f_r \in \mathcal{H}_{G_r}, r \in \mathbb{N}_n \right\} \\
&= \min\left\{ \sum_{j\in\mathbb{N}_m} A\left(y_j, \langle w, \Phi(x_j)\rangle_{\mathcal{W}}\right) + \mu B\left(\|w\|_{1,\mathcal{W}}^2\right) : w \in \mathcal{W}^n \right\} \ = \ V_\Phi
\end{aligned}$$

where the first equality follows by Theorem 3.1, the second follows by taking the minimum over $\lambda \in \mathbb{S}^n$ and the third equality uses the feature map representation of the function f_r, that is, $f_r = \langle w_r, \Phi_r\rangle_{\mathcal{W}}$, $r \in \mathbb{N}_n$, where $\Phi = (\Phi_r : r \in \mathbb{N}_n)$, and $\|f_r\|_{G_r} = \|w_r\|_{\mathcal{W}}$. Moreover, the optimal value of $\lambda = (\lambda_r : r \in \mathbb{N}_n)$ is given by

$$\lambda_r = \frac{\|w_r\|_{\mathcal{W}}}{\|w\|_{1,\mathcal{W}}}. \tag{3.6}$$

Note that this alternate derivation only reveals equation (3.1) and does not provide information about the structure for the extremal solutions for the associated variational problems. Nevertheless, it suggests the following extension of this equation. To this end, we choose $h \in \mathbb{R}_+$, set $\mathbb{S}^{n,h} := \{(\lambda_r : r \in \mathbb{N}_n) : \lambda_r \geq 0, \sum_{r\in\mathbb{N}_n} \lambda_r^h = 1\}$,

$$\mathcal{K}_h(\mathcal{G}) = \left\{ \sum_{r\in\mathbb{N}_n} \lambda_r G_r : \lambda \in \mathbb{S}^{n,h} \right\},$$

$$E_{G,h} = \inf \left\{ \sum_{j \in \mathbb{N}_m} A(y_j, f(x_j)) + \mu B\left(\|f\|_K^2\right) : f \in \mathcal{H}_K, K \in \mathcal{K}_h(\mathcal{G}) \right\}$$

and

$$E_{\Phi,p} = \min \left\{ \sum_{j \in \mathbb{N}_m} A(y_j, \langle w, \Phi(x_j) \rangle_{\mathcal{W}}) + \mu B\left(\|w\|_{p,\mathcal{W}}^2\right) : w \in \mathcal{W}^n \right\}. \tag{3.7}$$

Proposition 3.1. *If $h \in \mathbb{R}_+$ and $p = \frac{2h}{h+1}$ then $E_{G,h} = E_{\Phi,p}$.*

The proof of this proposition follows from Theorem 3.1, the following lemma, whose proof can be found in the appendix of [17] and the same technique used above to give an alternate proof of equation (3.1).

Lemma 3.1. *If $h \geq 0$, $p := \frac{2h}{h+1}$, and $a = (a_r : r \in \mathbb{N}_n) \in \mathbb{R}^n$ then*

$$\min \left\{ \left(\sum_{r \in \mathbb{N}_n} \frac{|a_r|^2}{\lambda_r} \right)^{\frac{1}{2}} \; : \; \lambda = (\lambda_r : r \in \mathbb{N}_n) \in \mathbb{S}^{n,h} \right\} = \|a\|_p^2$$

and the equality occurs for

$$\lambda_r := \left(\frac{|a_r|}{\|a\|_p} \right)^{2-p}, \quad r \in \mathbb{N}_n. \tag{3.8}$$

We hope on a future occasion, to use this alternate approach to discover the structure of the optimal solutions for $E_{G,h}$ and $V_{\Phi,p}$.

4 Single feature kernels and L^1 regularization

In this section, we consider another case of our main result in Section 2 corresponding to the choice $\mathcal{W} = \mathbb{R}$. Equivalently, the kernels in $\mathcal{G}$ are all expressed as a *single feature*. We have already observed in the previous section that in this case, under the additional assumption that Θ is a finite set, problem (3.2) reduces to problem (3.5) which is a type of L^1 regularization problem. An analogous observation is summarized in the corollary below in the general case that Θ is any compact set. We note that in this case $C(\Theta, \mathbb{R}) = C(\Theta)$.

Corollary 4.1. *Under the hypothesis above, we have that $E_G = V_\Phi$ where*

$$E_G = \inf \left\{ \sum_{j \in \mathbb{N}_m} A(y_j, f(x_j)) + \mu B\left(\|f\|_K^2\right) : f \in \mathcal{H}_K, K \in \mathcal{K}(\mathcal{G}) \right\}$$

and

$$V_\Phi = \min \left\{ \sum_{j \in \mathbb{N}_m} A(y_j, L(\Phi_j)) + \mu B\left(\|L\|^2\right) : L \in C^*(\Theta) \right\}.$$

Moreover, there exist a finitely supported measure $\hat{p} \in \mathcal{M}(\Theta)$ with at most $m+1$ atoms and a vector $\hat{c} \in \mathbb{R}^m$ such that the kernel

$$\hat{K} = \int_\Theta G(\theta) d\hat{p}(\theta)$$

is an optimal kernel, the function

$$f_{\hat{K}} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \hat{K}_j$$

is an optimal function and the linear functional $\hat{L} \in C^*(\Theta)$ defined, for any $\Psi \in C(\Theta)$, as

$$\hat{L}(\Psi) = \int_\Theta \hat{\Gamma}(\theta)\Psi(\theta) d\hat{p}(\theta)$$

is an optimal linear functional, where the function $\hat{\Gamma} = \sum_{j \in \mathbb{N}_m} \hat{c}_j \Phi_j$ has the property, for every atom θ of $\hat{p}$, that $|\hat{\Gamma}(\theta)| = \|\hat{\Gamma}\|_\infty = \|\hat{L}\|$.

Note that we can rewrite the linear functional $\hat{L}$ as a finitely supported signed measure, namely,

$$\hat{L} = \sum_{j \in \mathbb{N}_q} \hat{\gamma}_j \delta(\theta_j - \cdot)$$

where $\{\theta_i : i \in \mathbb{N}_q\} \subseteq \Theta$ are the atoms of $\hat{p}$, $q \leq m+1$, $\hat{\gamma}_i = \hat{\lambda}_i \|\hat{\Gamma}\|_\infty \mathrm{sgn}\hat{\Gamma}(\theta_i)$ and for each $\theta \in \Theta$, we interpret $\delta(\theta - \cdot)$ as the delta function concentrated at θ. Moreover, we can alternatively express the optimal function $f_{\hat{K}}$ as a linear combination of the features evaluated at the atoms of $\hat{p}$, that is

$$f_{\hat{K}} = \sum_{j \in \mathbb{N}_q} \hat{\gamma}_j \Phi(\theta_j \cdot).$$

We view the feature space variational problem appearing in the above corollary as the analog of L^1 regularization. We shall now change our perspective and explain in detail what we have in mind. Our intention is also to have this discussion encompass an L^p extension of the variational problem above for $p \in (1, \infty)$. To this end, we describe the necessary terminology and notation to cover this case too.

The appropriate context for this discussion is a measurable space Θ with finite measure ν and $L^p(\Theta, \nu)$ the space of functions $\omega : \Theta \to \mathbb{R}$ with norm $\|\omega\|_p$ defined, for $p \in [1, \infty)$, as

$$\|\omega\|_p := \left(\int_\Theta |\omega(\theta)|^p d\nu(\theta) \right)^{\frac{1}{p}}.$$

We wish to learn a function in $L^p(\Theta, \nu)$ based on a finite set of linear functionals of it, that is, we have a set of examples of the form $\{(M_j, y_j) : j \in \mathbb{N}_m\}$ where M_j are *linear* functionals in $L^q(\Theta, \nu)$ and $y_j \in \mathbb{R}$ is a noisy measurement of $M_j(\omega)$ from the unknown target function ω. Furthermore, we assume that the linear functionals $\{M_j : j \in \mathbb{N}_m\}$ are *linearly independent*.

To estimate ω we consider the problem of minimizing the functional $E_p : L^p(\Theta, \nu) \to \mathbb{R}$ defined, for $\omega \in L^p(\Theta, \nu)$, as

$$E_p(\omega) := q\left(M(\omega), \|\omega\|_p\right) \tag{4.1}$$

over its domain, where q is an admissible regularization function and $M : L^p(\Theta, \nu) \to \mathbb{R}^m$ is the linear operator defined, for $\omega \in L^p(\Theta, \nu)$, as $M(\omega) = (M_j(\omega) : j \in \mathbb{N}_m)$. Recall, for $p \in [1, \infty)$, that the linear functionals M_j can be expressed as

$$M_j(\omega) = \int_\Theta \varphi_j(\theta)\omega(\theta)d\nu(\theta)$$

where the function $\varphi_j \in L^q(\Theta, \nu)$ and $\frac{1}{p} + \frac{1}{q} = 1$, see, for example, [20]. A special case of this setup is covered by the regularization error functional

$$E_p(\omega) := \sum_{j\in\mathbb{N}_m} A(y_j, M_j(\omega)) + \mu\|\omega\|_p^p, \quad \omega \in L^p(\Theta, \nu) \tag{4.2}$$

where $A : \mathbb{R} \to \mathbb{R}_+$ is some prescribed loss function and μ is a positive parameter.

As an example of the above we let $N : \Theta \times \Theta \to \mathbb{R}$ be a prescribed continuous function and $\mathcal{N} : L^p(\Theta, \nu) \to C(\Theta)$ the associated integral operator, that is, for $\omega \in L^p(\Theta, \nu)$, we define

$$\mathcal{N}\omega(\cdot) = \int_\Theta N(\cdot, \theta)\omega(\theta)d\nu(\theta).$$

We introduce a linear space of functions $T := \text{range}\mathcal{N}$. We assume $\mathcal{N}$ is one-to-one and observe that the norm of $h \in T$ defined as $\|h\| := \|\omega\|_p$ where $h = \mathcal{N}\omega$ makes T a Banach space.

We choose $\varphi_j := N_j, \quad j \in \mathbb{N}_m$, where $N_j(\cdot) := N(\theta_j, \cdot)$, and express the regularization function (4.1) in the form

$$E_p(\omega) = q(I_{\boldsymbol{\theta}}(h), \|h\|) \tag{4.3}$$

where $h = \mathcal{N}\omega$ and $\boldsymbol{\theta} = \{\theta_j : j \in \mathbb{N}_m\}$ is a prescribed set of inputs. Clearly, minimizing the left hand side of equation above over $\omega \in L^p(\Theta, \nu)$ is equivalent to minimize the right hand side of this equation over $h \in T$.

In general, T is not a Hilbert space. However, for the special case that $p = 2$, T is a reproducing kernel Hilbert space with inner product defined as $\langle h, h'\rangle = \int_\Theta \omega(\theta)\omega'(\theta)d\nu(\theta)$, where $h = \mathcal{N}\omega$, $h' = \mathcal{N}\omega'$ and the reproducing kernel K is given by

$$K(\theta, \theta') = \int_\Theta N(\theta, s)N(\theta', s)d\nu(s), \quad \theta, \theta' \in \Theta.$$

Indeed, to see that K is the reproducing kernel of T we observe that the above formula means, for $\theta \in \Theta$, that $K(\theta, \cdot) = \mathcal{N}(N(\theta, \cdot))$ and, so, if $h = \mathcal{N}\omega$, by definition we have that

$$\langle K(\theta, \cdot), h\rangle = \int_\Theta N(\theta, s)\omega(s)d\nu(s) = h(\theta).$$

In this special case, it is well-known that the unique minimizer $\hat{h}$ of the functional in the right hand side of equation (4.3) has the form

$$\hat{h} = \sum_{j\in\mathbb{N}_m} c_j K_j \tag{4.4}$$

where $K_j = K(\theta_j, \cdot)$ and the vector of coefficients $c = (c_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$ is obtained by substituting formula (4.4) for $\hat{h}$ into the right hand side of equation (4.3) and minimizing over c. Note that, if $\hat{h} = \mathcal{N}\hat{\omega}$ then $\hat{\omega}$ is given by

$$\hat{\omega} = \sum_{j\in\mathbb{N}_m} c_j N_j$$

or, equivalently, recalling the definition of φ_j in this example, we have that $\hat{\omega} = \sum_{j \in \mathbb{N}_m} c_j \varphi_j$.

Let us turn our attention to the case that $p = 1$. We begin our discussion by observing that, in general, E_1 does not have a minimum on $L^1(\Theta, \nu)$. We illustrate this point with the following example.

Example 1. *Let $d\theta$ be the Lebesgue measure on $[0, 1]$ and consider the problem of minimizing the functional*

$$W(\omega; \alpha) = \alpha \left| 1 - \int_0^1 \omega(\theta) \theta d\theta \right| + \int_0^1 |\omega(\theta)| d\theta$$

over $\omega \in L^1([0, 1], d\theta)$, where α is a nonnegative number. Call the value of this minimum $W_\infty(\alpha) := \inf\{W(\omega; \alpha) : \omega \in L^1([0, 1], d\theta)\}$. With minimal effort it follows that

$$W_\infty(\alpha) = \begin{cases} 1, & \alpha > 1 \\ \alpha, & \alpha \in (0, 1]. \end{cases}$$

Moreover, the minimum does not exist in $L([0, 1], d\theta)$ but does exist as a distribution $\hat{\omega}$ given, for $\theta \in [0, 1]$ by

$$\hat{\omega}(\theta) = \begin{cases} \delta(\theta - 1), & \alpha > 1 \\ 0, \ a.e., & \alpha \in (0, 1]. \end{cases}$$

If we embed $L^1([0, 1], d\theta)$ into $C^*([0, 1])$ and minimize $W(\cdot; \alpha)$ over $C^*([0, 1])$ the minimum exists and is given as above. Likewise, in general, the minimum of functional E_1 defined in (4.2) exists in $C^*(\Theta)$. Indeed, this corresponds to the feature space variational problem treated in Corollary 4.1. Keep in mind that our description of the L_p regularization above takes the point of view of learning ω from the data $M(\omega)$. However, in the case $p = 1$, from our remarks above we *also* view it from the feature space perspective presented in Corollary 4.1.

We complete this digression by describing a representer theorem for the $L^p(\Theta, \nu)$ regularization when $1 < p < \infty$. Before doing so, we think it is advantageous to present another for proof the L^1 regularization case independent of the general theorem presented in Section 2.

Proposition 4.1. *There exist an integer $q \leq m + 1$, a set $\{\theta_j : j \in \mathbb{N}_q\} \subseteq \Theta$ and $\hat{\lambda} = (\hat{\lambda}_j : j \in \mathbb{N}_q) \in \mathbb{S}^q$ such that*

$$\hat{L} = \sum_{j \in \mathbb{N}_q} \hat{\lambda}_j \delta(\theta_j - \cdot)$$

is a minimizer of the regularization functional E_1 above. Moreover, there is a vector $\hat{c} = (\hat{c}_j : j \in \mathbb{N}_m)$ satisfying the constraint that $(\hat{c}, \hat{y}) = 1$, where $\hat{y} := (\hat{L}(\varphi_j) : j \in \mathbb{N}_m)$, such that the function $\hat{\Gamma} := \sum_{j \in \mathbb{N}_m} \hat{c}_j \varphi_j$ has the property that, for every $j \in \mathbb{N}_q$,

$$|\hat{\Gamma}(\theta_j)| = \|\hat{\Gamma}\|_\infty = \min \left\{ \left\| \sum_{j \in \mathbb{N}_m} d_j \varphi_j \right\|_\infty \ : \ d \in \mathbb{R}^m, (d, \hat{y}) = 1 \right\}.$$

Proof. The existence of a minimum of E_1 over $C^*(\Theta)$ follows from weak-*compactness in the unit ball in $C^*(\Theta)$. If $\hat{L}$ is a minimizer of E_1, we set $\hat{y}_j = \hat{L}(\varphi_j), j \in \mathbb{N}_m$, choose any $d \in \mathbb{R}^m$ and note that, for every $L \in C^*(\Theta)$ such that $L(\varphi_j) = \hat{y}_j, j \in \mathbb{N}_m$, that

$$(d, \hat{y}) = \sum_{j \in \mathbb{N}_m} d_j L(\varphi_j) = L\left(\sum_{j \in \mathbb{N}_m} d_j \varphi_j) \right) \leq \|L\| \left\| \sum_{j \in \mathbb{N}_m} d_j \varphi_j \right\|_\infty .$$

Consequently, we have that

$$\|L\| \geq \frac{\sum_{j\in\mathbb{N}_m} d_j\hat{y}_j}{\left\|\sum_{j\in\mathbb{N}_m} d_j\varphi_j\right\|_\infty} \geq \sigma^{-1}$$

where we have defined

$$\sigma := \min\left\{\left\|\sum_{j\in\mathbb{N}_m} d_j\varphi_j\right\|_\infty : \sum_{j\in\mathbb{N}_m} d_j\hat{y}_j = 1\right\}. \tag{4.5}$$

We observe that the variational problem (4.5) has a minimum because the function $U : \mathbb{R}^m \to \mathbb{R}_+$ defined for each $d = (d_j : j \in \mathbb{N}_m)$ by $U(d) := \|\sum_{j\in\mathbb{N}_m} d_j\varphi_j\|_\infty$ is continuous, homogeneous and nonzero for $d \neq 0$. Hence, it tends to infinity as $d \to \infty$.

Let $\hat{c} \in \mathbb{R}^m$ be a minimizer of U. Therefore, this vector is characterized by the fact that the right directional derivative of U at $\hat{c}$ in all directions $a \in \mathbb{R}^m$ such that $(a, \hat{y}) = 1$ is nonnegative. We denote this derivative by $U'_+(\hat{c}; a)$ which is given by

$$U'_+(\hat{c}; a) = \max\left\{\left(\sum_{j\in\mathbb{N}_m} a_j\varphi_j(\theta)\right)\operatorname{sgn}\left(\sum_{j\in\mathbb{N}_m} \hat{c}_j\varphi_j(\theta)\right) : \theta \in \Theta(\hat{c})\right\}$$

where the set $\Theta(\hat{c})$ is defined as

$$\Theta(\hat{c}) := \left\{\theta : \theta \in \Theta,\ \left|\sum_{j\in\mathbb{N}_m} \hat{c}_j\varphi_j(\theta)\right| = \left\|\sum_{j\in\mathbb{N}_m} \hat{c}_j\varphi_j\right\|_\infty\right\}.$$

For each $\theta \in \Theta$, we define the vector $z(\theta) := (\varphi_j(\theta)\operatorname{sgn}(\sum_i \hat{c}_i\varphi_i(\theta)) : j \in \mathbb{N}_m) \in \mathbb{R}^m$ and the set of vectors $Z(\hat{c}) := \{z(\theta) : \theta \in \Theta(\hat{c})\} \subseteq \mathbb{R}^m$.

The condition that $\hat{c}$ is a solution to problem (4.5) means, for all $a \in \mathbb{R}^m$ satisfying $(a, \hat{y}) = 0$, that

$$\max\{(z, a) : z \in Z(\hat{c})\} \geq 0.$$

Clearly, $Z(\hat{c})$ is a bounded subset of $\mathbb{R}^m$. Therefore, its closed convex hull $A := \overline{co(Z(\hat{c}))}$ is compact. We claim that A intersects the line spanned by the vector $\hat{y}$. Indeed, if this is not true then there exists a hyperplane $H := \{d : d \in \mathbb{R}^m, (\beta, d) + \alpha = 0\}$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^m$, which strictly separates A from the set $\{\rho\hat{y} : \rho \in \mathbb{R}\}$, see, for example, [20]. In other words, we must have that

$$(\beta, \rho\hat{y}) + \alpha > 0, \quad \rho \in \mathbb{R}$$

and

$$(\beta, z) + \alpha \leq 0, \quad z \in Z(\hat{c}).$$

The first condition implies that $\alpha > 0$ and, so we conclude that

$$\max\{(\beta, z) : z \in Z(\hat{c})\} < 0$$

which contradicts our hypothesis that $\hat{c}$ is a minimum of U. Hence, we have that $\rho_0\hat{y} \in A$ for some $\rho_0 \in \mathbb{R}$. By the Caratheodory theorem, see, for example, [4, Ch. 2], every vector in A can be

expressed as a convex combination of at most $m+1$ vectors in $Z(\hat{c})$. In particular, there exists ρ_0 such that

$$\rho_0 \hat{y} = \int_\Theta z(\theta) d\hat{p}(\theta) \tag{4.6}$$

where $\hat{p}$ is a probability measure with q atoms, that is,

$$\hat{p} = \sum_{j \in \mathbb{N}_q} \gamma_j \delta(\cdot - \theta_j)$$

q is at most $m+1$, $\{\theta_j : j \in \mathbb{N}_q\} \subseteq \Theta(\hat{c})$, $\gamma_j \geq 0$, $j \in \mathbb{N}_m$ and $\sum_{j \in \mathbb{N}_m} \gamma_j = 1$. Taking the inner product of both sides of equation (4.6) with c we conclude that

$$\rho_0 = \left\| \sum_{j \in \mathbb{N}_m} \hat{c}_j \varphi_j \right\|_\infty = \|g\|_\infty > 0$$

where $g = \sum_{j \in \mathbb{N}_q} \hat{c}_j \varphi_j$. Next, we introduce the linear functional $\hat{L} : C(\Theta) \to \mathbb{R}$ defined for $\omega \in C(\Theta)$ as $\hat{L}(\omega) = \int_\Theta \hat{\Gamma}(\theta)\omega(\theta) d\hat{p}(\theta)$ where

$$\hat{\Gamma} = \frac{1}{\rho_0} \mathrm{sgn}(g) \hat{p} = \sum_{j \in \mathbb{N}_q} \hat{\lambda}_j \delta(\theta_j - \cdot).$$

where we have defined $\hat{\lambda}_j = \frac{\gamma_j}{\rho_0}$. The result follows by noting that $\hat{L}(\varphi_j) = \hat{y}_j$ and $\|\hat{L}\| = \|\hat{\Gamma}\|_\infty^{-1}$. ■

5 Regularization in L^p spaces

In this section provide a representation result for the minimizer of functional (4.1) when $p \in (1, \infty)$.

Proposition 5.1. *If the function $q : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}_+$ is admissible and $p \in (1, \infty)$ then there exists a minimizer $\hat{\omega} \in L^p(\Theta, \nu)$ of functional (4.1), given by the form*

$$\hat{\omega} = \left(\left| \sum_{j \in \mathbb{N}_m} \hat{c}_j \varphi_j \right| \right)^{q-1} \mathrm{sign} \left(\sum_{j \in \mathbb{N}_m} \hat{c}_j \varphi_j \right) \tag{5.1}$$

where $\hat{c} = (\hat{c}_j : j \in \mathbb{N}_m) \in \mathbb{R}^m$.

Proof. E_p has a minimum in $L^p(\Theta, \nu)$ since q is an admissible regularization function and the unit ball in $L^p(\Theta, \nu)$ is weakly compact. Let $\hat{\omega}$ a minimizer of E_p, define data $\hat{y}_j = M_j(\hat{\omega}), j \in \mathbb{N}_m$ and choose $\tilde{\omega}$ as the solution to the minimal norm interpolation problem

$$\min \left\{ \|\omega\|_p : \omega \in L^p(\Theta, \nu),\ M_j(\omega) = \hat{y}_j, j \in \mathbb{N}_m \right\}. \tag{5.2}$$

Clearly $\tilde{\omega}$ is also a minimizer of E_p and, for any $d \in \mathbb{R}^m$, we have that

$$(d, \hat{y}) = \sum_{j \in \mathbb{N}_m} d_j M_j(\tilde{\omega}) = \int_\Theta \tilde{\omega}(\theta) \sum_{j \in \mathbb{N}_m} d_j \varphi(\theta) d\nu(\theta) \leq \|\tilde{\omega}\|_p \left\| \sum_{j \in \mathbb{N}_m} d_j \varphi_j \right\|_q$$

where the last inequality follows by Hölder's inequality. Consequently, we have that

$$\|\tilde{\omega}\|_p \geq \frac{\sum_{j\in\mathbb{N}_m} d_j\hat{y}_j}{\left\|\sum_{j\in\mathbb{N}_m} d_j\varphi_j\right\|_q} \geq \sigma^{-1}$$

where we have defined

$$\sigma = \min\left\{\left\|\sum_{j\in\mathbb{N}_m} d_j\varphi_j\right\|_q : \sum_{j\in\mathbb{N}_m} d_j\hat{y}_j = 1\right\}. \tag{5.3}$$

Let $\hat{c}\in\mathbb{R}^m$ be a minimizer of (5.3), define $\varphi := \sum_{j\in\mathbb{N}_m}\hat{c}_j\varphi_j$, set $\tilde{\omega} = \|\varphi\|^{-q}|\varphi|^{q-1}\mathrm{sgn}\,(\varphi)$ and note that

$$\|\tilde{\omega}\|_p = \|\varphi\|^{-q}\left(\int_\Theta |\varphi(\theta)|^{p(q-1)}\,d\nu(\theta)\right)^{\frac{1}{p}} = \|\varphi\|^{-q}\|\varphi\|^{\frac{q}{p}} = \|\varphi\|_q^{-1}.$$

This proves the claimed result. ■

In order to compute the coefficient vector $\hat{c} = (\hat{c}_j : j\in\mathbb{N}_m)$ in equation (5.1) we substitute this equation for $\hat{\omega}$ in the right hand side of equation (4.1) obtaining the function

$$\hat{E}_p(\hat{c}) := E_p\left(\left|\sum_{j\in\mathbb{N}_m}\hat{c}_j\varphi_j\right|^{q-1}\mathrm{sgn}\left(\sum_{j\in\mathbb{N}_m}\hat{c}_j\varphi_j\right)\right),\ \ \hat{c}\in\mathbb{R}^m$$

and then optimize $\hat{E}_p$ over the vector $\hat{c}\in\mathbb{R}^m$. Unfortunately, the function $\hat{E}_p$ is not, in general, convex. In fact, for the square loss $A(y,t) = \frac{1}{2}(y-t)^2$, $t\in\mathbb{R}$ with $\Theta = [0,1]$, $m=1$, $y_1 = 1$ and $\varphi = 1$ we have, for every $p\neq 2$, that

$$\hat{E}_p(\hat{c}) = (1-|\hat{c}|^{q-1}\mathrm{sgn}(\hat{c}))^2 + \gamma|\hat{c}|^{q-1}.$$

This function is not convex when $\gamma < 1$.

6 Proof of the main result

In this section we present the proof of our main result in Theorem 2.1. We divide the proof in two parts. In the first part we establish that

$$E_G \geq V_\Phi. \tag{6.1}$$

Recall that, for every $K\in\mathcal{K}(\mathcal{G})$ we defined f_K to be any function in $\mathcal{H}_K$ such that

$$E(K) := \min\{Q(f,K) : f\in\mathcal{H}_K\} = Q(f_K,K).$$

For every $y\in\mathbb{R}^m$ we use the notation $\rho(K,y)$ to denote the minimum norm squared of all functions $f\in\mathcal{H}_K$ which interpolate y at $\mathbf{x}$, that is, $I_{\mathbf{x}}f = y$ and set

$$\rho(K,y) := \min\left\{\|f\|_K^2 : f\in\mathcal{H}_K,\ I_{\mathbf{x}}f = y\right\}. \tag{6.2}$$

As we remarked earlier f_K may not be unique and, hence, we are not certain of its structure. To overcome this difficulty we introduce the vector $y_K := I_{\mathbf{x}} f_K$ and let $g_K \in \mathcal{H}_K$ solve the minimal norm interpolation problem

$$\|g_K\|_K^2 = \rho(K, y).$$

Hence, $g_K = I_{\mathbf{x}}^* c_K$ for a unique $c_K \in \mathbb{R}^m$ identified by the linear equation $I_{\mathbf{x}} g_K = y_K$. This follows from the so-called representer theorem. Consequently, we have that

$$E(K) = q(I_{\mathbf{x}} f_K, \|f_K\|_K^2) = q(y_K, \|f_K\|_K^2) \geq q(y_K, \|g_K\|_K^2).$$

Note that in the last step we used property 3 of an acceptable regularization error function.

Since $K \in \mathcal{K}(\mathcal{G})$, there exists $p \in \mathcal{M}(\Theta)$ such that

$$K = \int_\Theta G(\theta) dp(\theta).$$

We observe, for any $y \in \mathcal{X}$, that

$$g_K(y) = \int_\Theta \langle \Gamma_K(\theta), \Phi(\theta, y) \rangle dp(\theta)$$

where $\Gamma_K := \sum_{j \in \mathbb{N}_m} c_{K,j} \Phi_j$ and $c_K := (c_{K,j} : j \in \mathbb{N}_m)$. This computation suggests that we introduce the linear functional L_K defined, for every $\Psi \in C(\Theta, \mathcal{W})$, as

$$L_K(f\Psi) = \int_\Theta \langle \Gamma_K(\theta), \Psi(\theta) \rangle dp(\theta). \tag{6.3}$$

Therefore, with these observations we obtain that

$$y_K = I_{\mathbf{x}} g_K = D_{\mathbf{x}}(L_K) \tag{6.4}$$

where the linear operator $D : C^*(\Theta, \mathcal{W}) \to \mathbb{R}^m$ was defined earlier, below equation (2.4). Also, observe, for any $\Psi \in C(\Theta, \mathcal{W})$ by the Cauchy-Scwarz inequality used twice (once in $\mathcal{W}$ and then another time in $L^2(\Theta, dp)$), that

$$|L_K(\Psi)| \leq \int_\Theta \|\Gamma_K(\theta)\|_{\mathcal{W}} \|\Psi(\theta)\|_{\mathcal{W}} dp(\theta) \leq \left(\int_\Theta \|\Gamma_K(\theta)\|_{\mathcal{W}}^2 dp(\theta) \right)^{\frac{1}{2}} \left(\int_\Theta \|\Psi(\theta)^2\|_{\mathcal{W}} dp(\theta) \right)^{\frac{1}{2}}$$

and, so, we get the following inequality for the norm of L_K,

$$\|L_K\|^2 \leq \int_\Theta \|\Gamma_K(\theta)\|_{\mathcal{W}}^2 dp(\theta). \tag{6.5}$$

A straightforward computation shows that

$$\int_\Theta \|\Gamma_K(\theta)\|_{\mathcal{W}}^2 dp(\theta) = \|g_K\|_K^2. \tag{6.6}$$

Consequently, combining equations (6.5) and (6.6) we have demonstrated that

$$\|L_K\| \leq \|g_K\|_K. \tag{6.7}$$

Now, we observe from equations (6.4) and (6.7), and the definition of the functional V_Φ, for *any* $K \in \mathcal{K}(\mathcal{G})$, that

$$E(K) = E(y_K, f_K) \geq q(y_K, \|g_K\|_K^2) = q(D_{\mathbf{x}}(L_K), \|g_K\|_K^2) \geq q(D_{\mathbf{x}}(L_K), \|L_K\|^2) \geq V_\Phi.$$

Since this lower bound for $E(K)$ holds for any $K \in \mathcal{K}(\mathcal{G})$, we have proved that

$$E_G \geq V_\Phi.$$

To show the reverse inequality we use a result from [17]. For the convenience of the reader we describe it in detail. To this end, we recall some notation used there. Before, we used $\rho(K, y)$ for the minimum norm squared of all functions $f \in \mathcal{H}_K$ which interpolate the data at y, see equation (6.2). The infimum of this quantity over all $K \in \mathcal{K}(\mathcal{G})$ will be denoted by $\rho(\mathcal{K}(\mathcal{G}), y)$. Moreover, since $\mathcal{K}(\mathcal{G}) \subseteq \mathcal{A}_+$, for each $y \in \mathbb{R}^m$ there is a unique $c_K \in \mathbb{R}^m$ such that

$$(c_K, I_{\mathbf{x}} I_{\mathbf{x}}^* c_K) = \|I^* c_K\|_K^2 = \rho(K, y)$$

and $I_{\mathbf{x}} I_{\mathbf{x}}^* c_K = y$. Since $I_{\mathbf{x}} I_{\mathbf{x}}^* = K_{\mathbf{x}}$ we also have that $c_K = K_{\mathbf{x}}^{-1} y$. For the statement of the theorem below we introduce the vector $\hat{c}_K := \rho^{-1}(K; y) c_K$ which has the property that $(\hat{c}_K, y) = 1$.

Theorem 6.1. *If Θ is a compact Hausdorff topological space and $G : \Theta \to \mathcal{A}_+(\mathcal{X})$ is continuous then there exists a kernel $\hat{K} = \int_\Theta G(\theta) d\hat{p}(\theta) \in \mathcal{K}(\mathcal{G})$ such that $\hat{p}$ is a discrete probability measure in $\mathcal{M}(\Theta)$ with at most $m+1$ atoms. Moreover, for $\hat{c} := \hat{c}_{\hat{K}}$ and any atom $\theta \in \Theta$ of $\hat{p}$, we have that*

$$(\hat{c}, G_{\mathbf{x}}(\theta)\hat{c}) = \max\{(\hat{c}, G_{\mathbf{x}}(\theta)\hat{c}) : \theta \in \Theta\} \tag{6.8}$$

$$\rho(\mathcal{K}(\mathcal{G}), y) = \rho(\hat{K}, y) = (y, \hat{K}_{\mathbf{x}}^{-1} y) \tag{6.9}$$

and for every $c \in \mathbb{R}^m$ with $(c, y) = 1$ and every $K \in \mathcal{K}(\mathcal{G})$ there holds

$$(\hat{c}, K_{\mathbf{x}}\hat{c}) \leq (\hat{c}, \hat{K}_{\mathbf{x}}\hat{c}) \leq (c, \hat{K}_{\mathbf{x}} c). \tag{6.10}$$

Returning to the proof of Theorem 2.1, we observe by weak*–compactness in $C^*(\Theta, \mathcal{W})$ that there is a solution to the problem of minimizing $V_\Phi : C^*(\Theta, \mathcal{W}) \to \mathbb{R}_+$ over its domain, see, for example, [20]. We call the minimum $\hat{L}$ and set $\hat{y} := D_{\mathbf{x}}(\hat{L})$. Hence, by definition we have that

$$V_\Phi = q(\hat{y}, \|\hat{L}\|^2).$$

To estimate this quantity from below, we consider the problem

$$\gamma := \min \left\{ \left\| \sum_{j \in \mathbb{N}_m} c_j \Phi_j \right\|_{\infty, \mathcal{W}}^2 : c \in \mathbb{R}^m,\ (c, \hat{y}) = 1 \right\}.$$

Vector–valued problems of this type were considered in [14]. Note that the minimum exists because the functions $\Phi_j,\ j \in \mathbb{N}_m$ were assumed to be linearly independent over Θ. For the problem

at hand we note that

$$\begin{aligned}\gamma &= \min\left\{\max\left\{\Big\|\sum_{j\in\mathbb{N}_m} c_j\Phi_j(\theta)\Big\|^2 \;:\; \theta\in\Theta\right\} : c\in\mathbb{R}^m, (c,\hat{y})=1\right\}\\ &= \min\left\{\max\left\{\int_\Theta \Big\|\sum_{j\in\mathbb{N}_m} c_j\Phi_j(\theta)\Big\|^2 dp(\theta) \;:\; p\in\mathcal{M}(\Theta)\right\} : c\in\mathbb{R}^m, (c,\hat{y})=1\right\}\\ &= \min\Big\{\max\Big\{(c,K_{\mathbf{x}}c) \;:\; K\in\mathcal{K}(\mathcal{G})\Big\} : c\in\mathbb{R}^m, (c,\hat{y})=1\Big\} = \rho^{-1}(\mathcal{K}(\mathcal{G});y)\end{aligned}$$

where we recall that $K_{\mathbf{x}} = (K(x_i,x_j) : i,j\in\mathbb{N}_m)$. Thus, by Theorem 6.1 there is a discrete probability measure $\hat{p}\in\mathcal{M}(\Theta)$ of support $\le m+1$ and a vector $\hat{c}\in\mathbb{R}^m$ such that for all $c\in\mathbb{R}^m$, $K\in\mathcal{K}(\mathcal{G})$, there holds the inequality

$$(\hat{c},K_{\mathbf{x}},\hat{c}) \le (\hat{c},\hat{K}_{\mathbf{x}}\hat{c}) \le (c,\hat{K}_{\mathbf{x}}c)$$

where

$$\hat{K} = \int_\Theta G(\theta)d\hat{p}(\theta).$$

Moreover, for each atom θ of $\hat{p}$ we have that

$$\rho^{-1}(\mathcal{K}(\mathcal{G});y) = (\hat{c},G_{\mathbf{x}}(\theta)\hat{c}) = \max\{(\hat{c},G_{\mathbf{x}}(\theta)\hat{c}) : \theta\in\Theta\} \tag{6.11}$$

and the kernel $\hat{K}$, has the property that $\hat{K}_{\mathbf{x}}\hat{c} = \gamma\hat{y}$. As before, we let $\hat{\Gamma} := \sum_{j\in\mathbb{N}_m}\hat{c}_j\Phi_j$ and observe that $\|\hat{\Gamma}(\theta)\|_{\mathcal{W}}^2 = (\hat{c},G_{\mathbf{x}}(\theta)\hat{c})$. Consequently, for each atom θ of $\hat{p}$, we have that $\|\hat{\Gamma}(\theta)\|_{\mathcal{W}} = \|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \sqrt{\gamma}$.

Now, let us consider the linear functional $F\in C^*(\Theta,\mathcal{W})$ defined for each $\Psi\in C^*(\Theta,\mathcal{W})$ as

$$F(\Psi) = \gamma^{-1}\int_\Theta \langle\hat{\Gamma}(\theta),\Psi(\theta)\rangle_{\mathcal{W}} d\hat{p}(\theta).$$

As before in the proof of (6.1), we conclude that $\|F\|\le\gamma^{-1}\|\hat{\Gamma}\|_{\infty,\mathcal{W}}$. However, in the present circumstance, since $F(\hat{\Gamma}) = \gamma^{-1}\|\hat{\Gamma}\|_{\infty,\mathcal{W}}^2$ we additionally obtain, by the definition of the norm of F, that $\gamma^{-1}\|\hat{\Gamma}\|_{\infty,\mathcal{W}}^2\le\|F\|\|\hat{\Gamma}\|_{\infty,\mathcal{W}}$. In other words, we obtain that $\gamma^{\frac{1}{2}} = \gamma^{-1}\|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \|F\|$.

To proceed further, we shall demonstrate that F is the minimal norm interpolant to the data $\hat{y}$. First, we observe that the functional F interpolate the data at $\hat{y}$. Indeed, since $K_{\mathbf{x}}\hat{c} = \gamma y$ we have that

$$D_{\mathbf{x}}(F) = \gamma^{-1}\hat{K}_{\mathbf{x}}^{-1}\hat{c} = \hat{y}.$$

Now, let $L\in C^*(\Theta,\mathcal{W})$ be *any* linear functional which interpolates the data, that is, $D_{\mathbf{x}}(L) = \hat{y}$. Therefore, we get that

$$1 = (\hat{c},\hat{y}) = L\Big(\sum_{j\in\mathbb{N}_m}\hat{c}_j\Phi_j\Big) = L(\hat{\Gamma}) \le \|L\|\|\hat{\Gamma}\|_{\infty,\mathcal{W}} = \sqrt{\gamma}\|L\|$$

from which we obtain that $\|F\|\le\|L\|$. In other words, we have established, as anticipated above, that

$$\|F\| = \min\{\|L\| : L\in C^*(\Theta,\mathcal{W}),\ D_{\mathbf{x}}(L) = \hat{y}_j, j\in\mathbb{N}_m\}.$$

Next, we introduce the function $\hat{g} := \gamma^{-1} I_{\mathbf{x}}^* \hat{c}$ and observe as before that $\|\hat{g}\|_{\hat{K}} = \|F\|$. Finally, since $I_{\mathbf{x}} \hat{g} = \hat{y}$, we conclude that

$$V_\Phi = q(\hat{y}, \|\hat{L}\|^2) \geq q(\hat{y}, \|F\|^2) = q(I_{\mathbf{x}} \hat{g}, \|\hat{g}\|_{\hat{K}}^2) \geq E_G.$$

Thus, we have that $V_\Phi = E_G$, g_K is a minimizer for $E(\hat{K})$, $\hat{K}$ is optimal for E and F is optimal for V_Φ.

7 Summary

We have presented an equivalence between the problem of learning a kernel within a prescribed set of continuously parameterized kernels studied in [1, 17] and a feature space reformulation of it. This leads us to study a variational regularization problem in the dual space of all continuous functions with values in the Hilbert space associated with the features map. This equivalence requires only weak conditions on the form of the regularization error function. Not only does it establish that these variational problems are the same but, also, it provides a choice of the optimal solutions to both problems. In particular, it generalizes some results from [1] which required the loss function to be differentiable.

Furthermore, we demonstrate that the problem of learning the kernel, which has been investigated extensively in the literature, see [1, 3, 5, 7, 11, 12, 13, 19, 23], in special cases reduces to L^p regularization, [16]. This connection highlights the importance of studying regularization in a non-Hilbert space framework in machine learning. Indeed, special cases of the feature space problem have been widely studied under the names *lasso* [21], basis pursuit denoising [6] and, recently, as the *cosso* method, [13].

There are a number of issues related to the work presented in this paper which would be valuable to explore. For example, how does the form of the optimal solutions to the variational problems evolve with μ? Our results in Section 2 and the subsequent comments in Section 3 say that there always exists a solution which uses at most $m + 1$ nonzero kernels or features. Do the number of non-zero components diminish with μ, as was seen in [15] for special cases? Finally, the study of generalization error bounds for the methods presented in this paper would definitely be of interest, see [18] for recent progress on this issue. Of course, the central challenge not addressed here is the practical implementation and numerical validation of the methods presented here.

References

[1] A. Argyriou, C.A. Micchelli and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. *Proc. 18-th Annual Conference on Learning Theory (COLT'05)*, Bertinoro, Italy, June, 2005.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **686**: 337–404, 1950.

[3] F.R. Bach, G.R.G Lanckriet and M.I. Jordan. Multiple kernels learning, conic duality, and the SMO algorithm. Proc. of the Int. Conf. on Machine Learning (ICML'04) 2004.

[4] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization. Theory and Examples*. CMS (Canadian Mathematical Society) Springer-Verlag, New York, 2000.

[5] O. Bousquet and D.J.L. Herrmann. On the complexity of learning the kernel matrix. *Advances in Neural Information Processing Systems*, **15**, 2003.

[6] S.S. Chen, D.L. Donoho, M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1): 33–61, 1998.

[7] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J.S. Kandola. On kernel-target alignment *Advances in Neural Information Processing Systems*, **14**, T. G. Dietterich, S. Becker, Z. Ghahramani (eds.), 2002.

[8] G.M. Fung, O.L. Mangasarian. A feature selection Newton method for support vector machine classification. *Comput. Optim. Appl.*, **28**(2): 185–202, 2004.

[9] S.R. Gunn and J.S. Kandola. Structural modelling with sparse kernels *Machine Learning*, **48**(1): 137–163, 2002.

[10] M. Herbster. Relative loss bounds and polynomial-time predictions for the K-LMS-NET algorithm. *Proc. of the 15-th Int. Conference on Algorithmic Learning Theory*, October 2004.

[11] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M.I. Jordan. Learning the kernel matrix with semi-definite programming. *J. of Machine Learning Research*, **5**: 27–72, 2004.

[12] Y. Lee, Y. Kim, S. Lee and J.-Y. Koo. Structured multicategory support vector machine with ANOVA decomposition. Technical Report No. 743, Department of Statistics, The Ohio State University, October 2004.

[13] Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models – cosso. Institute of Statistics Mimeo Series 2556, NCSU, January 2003.

[14] C.A. Micchelli. Curves from variational principles. *Mathematical Modeling and Numerical Analysis*, **26**: 77–93, 1992.

[15] C.A. Micchelli and A. Pinkus. Variational problems arising from balancing different error criteria. *Rendiconti di Matematica, Serie VII*, **14**: 37–86, 1994.

[16] C.A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. *Proc. of the 17–th Annual Conference on Learning Theory (COLT'04)*, Banff, Alberta, June 2004.

[17] C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. *J. of Machine Learning Research* (to appear), 2005.

[18] C.A. Micchelli, M. Pontil, Q. Wu, and D.X. Zhou. Error bounds for learning the kernel. Research Note 05/09, Dept of Computer Science, University College London, June, 2005.

[19] C.S. Ong, A.J. Smola, and R.C. Williamson. Hyperkernels. *Advances in Neural Information Processing Systems*, **15**, S. Becker, S. Thrun, K. Obermayer (Eds.), MIT Press, Cambridge, MA, 2003.

[20] H.L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, 3rd edition, 1988.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B*, **58**: 267–288, 1996.

[22] G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[23] Q. Wu, Y. Ying and D.X. Zhou. Multi-kernel regularization classifiers. *Preprint*, 2004.